



UNIVERSITÀ
DI PAVIA

SCUOLA DI ALTA FORMAZIONE DOTTORALE
MACRO-AREA SCIENZE DELLA VITA

PhD in Psychology, Neuroscience and Data Science
Department of Brain and Behavioral Sciences

Computational Language Analysis for Psychotherapy: Investigating Semantic Change and Clinical Applications

Academic year 2024-2025
Cycle XXVIII

Coordinatore
Prof.ssa Elena Cavallini

Doctoral candidate
Dr.ssa Luisa Siciliano

Tutor
Prof. Davide Liccione

Co-tutor
Prof.ssa Gabriella Bottini

INDEX

Introduction.....	p.4
Section 1 – Computational Language Analysis for Psychotherapy.....	p.9
1. Psychotherapy Process-Outcome Research.....	p.9
1.1. Two Paradigms in Psychotherapy Research: Quantitative and Qualitative.....	p.23
2. Artificial Intelligence and Natural Language Processing.....	p.30
2.1. Text Mining and Topic Modelling.....	p.54
2.2. Generative AI and Prompt Engineering.....	p.63
3. Psychotherapy, Computational Language Analysis, and Semantic Shift.....	p.71
3.1. Text Analysis Tools for Psychotherapy Research.....	p.74
3.2. Tracking Semantic Change in Psychotherapy.....	p.78
3.2.1. Toward a definition of <i>Semantic Shift</i>	p.83
3.3. Research Frontiers and Open Challenges.....	p.85
3.3.1. Bias in Psychotherapy and Bias in Algorithms.....	p.88
4. From Theory to Research: Methods and Tools to Investigate Changes in Psychotherapy.....	p.95
4.1. The Usefulness of Combining Topic Modelling and Statistical Analysis to Investigate the Therapeutic Process: A Single Case Study.....	p.97
4.1.1. Abstract.....	p.97
4.1.2. Methods.....	p.98
4.1.3. Results.....	p.103
4.2. Reconsidering Topic Dynamics in Psychotherapy: A Critical Analysis with LDA and BERTopic.....	p.111
4.2.1. Abstract.....	p.111
4.2.2. Methods.....	p.112
4.2.3. Results.....	p.118
4.3. Sentiment Analysis of a Psychotherapy Case Using RoBERTa-GoEmotions....	p.131
4.3.1. Abstract.....	p.131
4.3.2. Methods.....	p.132
4.3.3. Results.....	p.135

4.4.	Are Generative Models Ready for Psychotherapy Research? Evidence from Prompt-Based Topic Modelling Comparison.....	p.139
4.4.1.	Abstract.....	p.140
4.4.2.	Methods.....	p.141
4.4.3.	Results.....	p.144
4.5.	Discussion.....	p.149
4.5.1.	Topic Dynamics in Psychotherapy.....	p.150
4.5.2.	Tracing Emotional Change across Psychotherapy sessions.....	p.153
4.5.3.	From Topic Modelling to Generative AI.....	p.157
Section 2 – AI-based clinical application: PSICOGENIUS ^{AI}		p.161
1.	From research to the development of PSICOGENIUS ^{AI}	p.162
2.	From session to insight: how and why is it useful?.....	p.165
2.1.	Prompt Engineering and PSICOGENIUS ^{AI}	p.169
2.2.	Individual session outputs.....	p.170
2.3.	Overall output analysis.....	p.176
3.	AI as an ally, not a replacement.....	p.181
4.	Future directions for PSICOGENIUS ^{AI}	p.184
Section 3 – Computational Psychotherapy and Suicidal Risk Analysis.....		p.186
1.	Generative Topic Modelling in Psychotherapy: Revisiting and Comparing with Established Methods.....	p.186
1.1.	Database description.....	p.187
1.2.	Significance and Methods.....	p.188
2.	Extracting Personality Traits from Psychotherapy Transcripts with LLMs.....	p.192
2.1.	Database description.....	p.192
2.2.	Significance and Methods.....	p.193
3.	Human-Machine Collaboration in Predicting Suicide Risk Severity from Text.....	p.194
3.1.	Database description.....	p.195
3.2.	Significance.....	p.196
References.....		p.198
Appendices.....		p.224

Introduction

My doctoral research is situated at the intersection of psychotherapy process–outcome research and artificial intelligence, to explore how computational methods can support the scientific study and clinical monitoring of therapeutic change. When I began this doctoral trajectory, these domains were not as tightly connected as they have become in recent years. In my own training and professional development, however, they naturally converged. Alongside my clinical training in psychology and psychotherapy, I developed a parallel interest in programming and AI, initially sparked by personal study during the COVID-19 pandemic. This thesis proposes a structured attempt to integrate these two epistemic and methodological traditions (Figure 1).

Psychotherapy and Artificial Intelligence

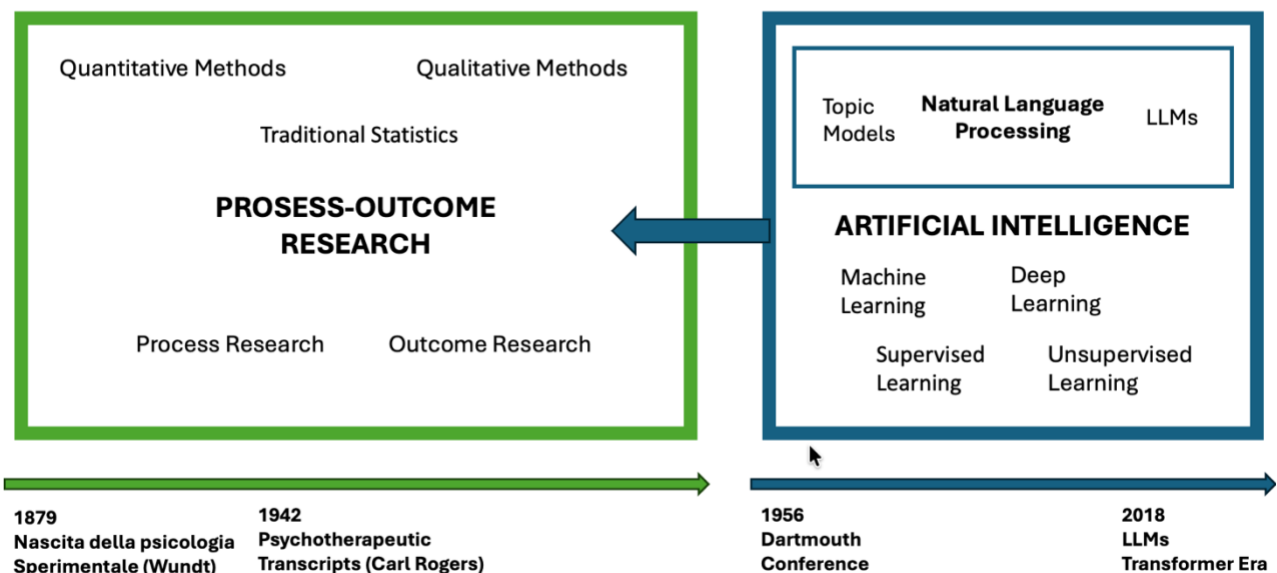


Figure 1. Psychotherapy Process–Outcome Research and Artificial Intelligence. Schematic representation of the conceptual and historical convergence between psychotherapy research traditions and AI-based language technologies, illustrating the framework of this thesis. The figure links established process–outcome methodologies, grounded in quantitative and qualitative analysis of therapeutic discourse, with computational approaches to text, including topic modelling, sentiment analysis, and generative systems. The timeline highlights significant milestones in psychotherapy research and AI development, showing how recent advances in NLP have enabled AI tools to function as decision-support resources for studying narrative and semantic change in psychotherapy. [The image is created by me].

Process-outcome research in psychotherapy has a long and well-established history (see Chapter 1, Section 1). Most evidence on “what works” in psychotherapy has been produced through qualitative methods and quantitative measurement, with statistics and standardized questionnaires as the dominant tools for assessing therapeutic mechanisms and outcomes. Over the last decade, however, machine learning and natural language processing (NLP) have increasingly entered the broader health

domain, first in medical applications and diagnostic support, and more recently in mental health research. A growing body of publications now documents their emerging role in psychotherapy research. Against this background, the present dissertation aims to contribute to the field by testing how AI-based approaches can be meaningfully combined with established process–outcome paradigms, not to replace clinical expertise, but to augment it.

A guiding assumption of this work is that AI systems should be conceptualized as decision-support technologies rather than substitutes for psychotherapy or clinical judgment. The use of machine learning for diagnostic or clinical decision-making, for example, is only defensible when embedded within specialist knowledge, substantial training, and rigorous methodological control. The same principle applies to contemporary generative systems. Large language models can assist with complex linguistic tasks, but only when the user has a solid understanding of the domain and can critically evaluate the outputs. These systems are inherently probabilistic, generating text by selecting the most likely continuations based on patterns learned from large corpora. As a result, the quality, relevance, and even the apparent plausibility of responses depend heavily on how the user formulates instructions, a practice often called prompt engineering. From a research perspective, this probabilistic architecture raises key issues regarding replicability and standardization, because identical inputs do not necessarily yield identical outputs (see Chapter 2, Section 1). Methodological protocols must therefore explicitly address sources of variability and define procedures that ensure transparency and reproducibility.

A further critical aspect concerns bias, both algorithmic and human. In psychotherapy, bias may operate at multiple levels, including the therapist’s interpretative frame, the patient’s narrative positioning, and the implicit norms embedded within the therapeutic setting. Similarly, AI tools reflect biases present in training data, model design, and evaluation procedures. For these reasons, the responsible use of AI in psychotherapy research requires systematic attention to the decision-making processes of both clinicians and computational systems, so that AI-generated results can be interpreted under controlled conditions rather than uncritically trusted. Such caution is significant because generative systems can appear highly reliable, often producing fluent and persuasive responses that foster misplaced confidence. Additionally, the rapid expansion of generative AI raises broader sustainability concerns, given the substantial energy demands of training and deploying large-scale models, as well as the infrastructure investments currently being made to support them.

The scientific motivation for integrating psychotherapy research and AI in this thesis is grounded in a shared focal point: language as a privileged access route to change. In psychotherapy, clinicians work primarily through language to engage the patient’s suffering, patterns of meaning-making, and

habitual ways of experiencing self and others. This work concerns both verbal and nonverbal communication, and it targets disruptions in identity coherence, intelligibility, and experiential integration. The hermeneutic tradition, understood as the study of texts and meaning, therefore provides an essential theoretical background for the present project. Within this perspective, I introduce and operationalize the construct of semantic shift as a marker of therapeutic change. Semantic shift refers to changes in the patient's narrative that reflect transformations in the content and organization of experience over the course of therapy. Because narration shapes how experience is emotionally lived and interpreted, changes in linguistic meaning can be understood as indicators of changes in experiential functioning, including movement toward greater well-being rather than persistent suffering.

At the same time, semantic change cannot be fully captured by AI tools that operate exclusively on textual data. Many clinically essential nuances are conveyed through nonverbal channels such as prosody, facial expressions, pauses, and irony, which remain difficult to model with the AI systems used in this dissertation. For this reason, computational outputs must be regarded as partial representations that require integration with qualitative interpretation and clinical expertise. The therapist's knowledge of the patient and the therapeutic context remain essential for assigning meaning to detected patterns and distinguishing clinically relevant change from artefactual variation. Nevertheless, a necessary contribution of computational methods lies in their ability to systematically extract patterns that may not be immediately visible to human observers, such as the evolution of topic structure, the reconfiguration of thematic associations, and the distribution of affective tone across sessions.

In practical terms, this thesis examines how AI can make semantic shifts observable. For example, topic modelling methods such as Latent Dirichlet Allocation (LDA) cluster words and discourse segments into latent themes, enabling the researcher to examine how the lexical composition of a topic changes across phases of treatment. In this way, topic models can provide a quantitative and interpretable proxy for meaning reorganization. Building on the literature, I was particularly influenced by the work of Imel and colleagues (2015), who applied topic modelling to large psychotherapy corpora to compare therapeutic orientations in terms of their thematic structure and associations with outcomes. Inspired by this approach, I adopted LDA to examine thematic dynamics across the different phases of a single psychotherapy case comprising 26 sessions with a positive outcome, to extract clinically informative indicators from a single completed treatment rather than only from aggregated multi-therapy datasets. A natural extension of this work, and an important direction for future research, would be to replicate the same analytic pipeline on a comparable case

with a negative outcome, for example, a dropout trajectory, using standardized outcome measures such as those employed in the present studies.

These considerations also motivated a broader applied objective: a long-term vision to develop a software tool to support psychotherapists by processing session transcripts and returning structured process indicators. Over the three years of the doctoral program, generative AI models have advanced rapidly, making it increasingly plausible to treat automatic transcription and translation, under specific conditions, as comparable to human performance. In this dissertation, however, the central focus is not on comparing manual and automatic transcription or translation. Rather, the key methodological question is whether human topic coding and algorithmic topic assignment, including both classical topic models and prompt-based generative systems, are comparable in terms of reliability and interpretability for psychotherapy process research.

To sum up, this thesis positions AI as a methodological extension of process–outcome research, leveraging computational tools to study how language, themes, and affect evolve over the course of psychotherapy. By combining topic modelling, sentiment modelling, and generative approaches, the work aims to contribute to an empirically grounded and clinically responsible integration of AI into psychotherapy research and, prospectively, into supportive clinical applications.

Over the course of my doctoral trajectory, the two domains that constitute the backbone of this dissertation, psychotherapy process–outcome research and language-based artificial intelligence, progressively intertwined until they became a single, integrated research program. The first article was published during the PhD, while the remaining manuscripts are currently being finalized and systematically consolidate the data and analyses produced across these years. Taken together, these results provided the methodological foundation for the conception and development of the AI tools implemented in the software described in detail in the second section of the thesis.

The dissertation is organized into two main parts. The first section introduces the theoretical and methodological framework, the tradition of process outcome research in psychotherapy, the role of language as a privileged access route to therapeutic change, and the principal computational approaches used in the literature to analyse clinical texts, with particular emphasis on topic modelling, measures of conversational dynamics, and more recent applications of generative AI. This section also addresses the methodological and ethical implications of using probabilistic tools in clinical contexts, including issues of replicability, bias, and interpretative responsibility.

The second part presents the empirical work conducted during the doctoral project and the applied rationale that follows from it. Chapter 4 brings together the primary studies developed from a complete psychotherapy case, initially transcribed manually and translated into English using reverse translation. In parallel, the rapid evolution of generative systems over the last few years has considerably accelerated transcription and translation workflows. However, the central objective of the doctoral work was not to compare the quality of automatic versus manual transcription, but rather to develop and test AI tools useful for both psychotherapy research and clinical practice, with a specific focus on extracting and monitoring process indicators.

Overall, the dissertation advances an integrated framework in which AI is conceptualized as support for clinical work and research, rather than as a substitute for psychotherapeutic interpretation. The empirical findings and the final discussions converge on a hybrid approach that combines robust statistical methods, NLP models, and generative systems into replicable, clinically meaningful analytical pipelines. On this basis, the second section of the thesis also presents the applied architecture that translates these methodological insights into a software project that produces structured, interpretable indicators from therapy transcripts, while keeping professional expertise and clinical context at the centre of interpretation.

In line with this translational aim, Section 2 is dedicated to PSICOGENIUS^{AI}, the software I developed in parallel with the research program. This section describes how the empirical results informed the design of an AI-based workflow that moves from session to insight, clarifying how prompt engineering, session-level outputs, and longitudinal summaries can support clinicians and researchers, consistent with the principle of AI as an ally rather than a replacement.

To conclude, Section 3 situates the dissertation within a broader international research trajectory. It presents the projects I contributed to during my visit to the University of Utah in Professor Zac Imel's lab, as well as the work I am continuing to develop from that collaboration. These studies expand the thesis's core methodological concerns, reliability, interpretability, and human-machine collaboration to higher-impact domains of computational psychotherapy, including generative topic modelling benchmarks, the extraction of personality-related constructs from psychotherapy transcripts using LLMs, and collaborative approaches to predicting suicide risk severity from text.

Section 1 – Computational Language Analysis for Psychotherapy

1. Psychotherapy Process-Outcome Research

This chapter offers the conceptual and historical background for the current research. It traces the evolution of psychotherapy as both a clinical practice and a scientific discipline, emphasizing its growth across different theoretical approaches and research methods. Understanding this history is essential because it provides the context for how process-outcome research has developed and continues to evolve. Specifically, the chapter covers:

- *The historical evolution of psychotherapy and the growth of various approaches throughout the twentieth century.*
- *The rise of process-outcome research as a key framework for connecting therapeutic processes with outcomes.*
- *The dialogue between quantitative and qualitative traditions culminates in methodological pluralism.*
- *Contemporary advances, including computational and ecological methods, create new opportunities for researching therapeutic change.*

Taken together, these sections provide the essential context and epistemological foundation for the empirical analyses in the following chapters, demonstrating how the current study builds on and advances the tradition of psychotherapy process-outcome research.

Historical Development of Psychotherapy Research

Psychotherapy is traditionally described as a treatment that addresses emotional, cognitive, and behavioural challenges using psychological techniques (American Psychological Association, 2023). As Cautin and Lilienfeld (2015) note, it involves “the treatment of emotional or physical ills through psychological means.” At its core, psychotherapy requires at least two participants – a therapist and a patient – who engage in a dynamic and dialogical interaction. Beyond this definition, Liccione (2019) emphasizes that the primary goal of psychotherapy is not merely to alter how people recount their stories. Instead, it concentrates on shifting experiential modes rather than changing the Self itself. This change occurs when patients, through self-narration during therapy and therapeutic exercises, become able to directly recognize their own specific modes of being – modes that may not have been integrated into their identity or may never have been verbally expressed (Liccione, 2019).

During the twentieth century, psychotherapy underwent significant growth. What began with a few approaches based on psychoanalysis, behaviourism, and humanistic psychology has expanded into over 400 recognized orientations (Lambert, 2013). Despite their diversity, these methods share a core reliance on dialogue, relational exchange, and meaning-making as key mechanisms of therapeutic change (Grencavage & Norcross, 1990; Laska, Gurman, & Wampold, 2014). Beyond the proliferation of orientations, the field has also been shaped by ongoing debates about what truly drives therapeutic change. Wampold and Imel (2015), in "The Great Psychotherapy Debate," argue that variance in outcomes is explained less by specific techniques than by common factors – such as the therapeutic alliance, client expectations, and therapist effects – that cut across different modalities. This perspective challenges the notion that effectiveness is primarily rooted in method-specific interventions, instead emphasizing relational and contextual factors as central to understanding the success of psychotherapy (Wampold & Imel, 2015; Laska, Gurman, & Wampold, 2014).

The field of psychotherapy has a dual nature: it is both a clinical practice and an empirical field of research. Theoretical frameworks give the conceptual foundation for treatment, while empirical studies aim to verify, compare, and improve these approaches through systematic investigation (Borckardt et al., 2008). This ongoing conversation between theory and practice is a key feature of psychotherapy, showcasing its dual identity as both a science and a profession (Epstein, 1994). In this context, the scientific study of psychotherapy has gone through three main phases that overlap but are also distinct conceptually. These phases show the gradual refinement of goals, methods, and epistemological assumptions as efforts to understand how psychotherapy works and prompts change (Russell & Orlinsky, 1996; Shedler, 2010). From early twentieth-century case studies to today's computational methods, the field has maintained a unique link between clinical practice and scientific investigation (Borckardt et al., 2008; Epstein, 1994).

The earliest phase of psychotherapy research was dominated by the attempt to demonstrate whether psychotherapy was effective at all. Early clinical contributions – such as Freud's detailed case studies (Anna O., the "Rat Man" in 1917) – illustrated attempts to link treatment and outcome, yet they lacked systematic controls, generalizability, and reproducibility. A crucial methodological advance occurred in the 1940s, when Carl Rogers began recording and publishing verbatim transcripts of therapeutic encounters, enabling researchers to analyse therapeutic dialogue in a structured, comparative manner. The scientific debate soon crystallized around the efficacy of psychotherapy. Rosenzweig (1936) formulated the "Dodo Bird verdict," claiming that all psychotherapies yield equivalent benefits because their effects primarily derive from common, nonspecific factors such as empathy, expectations of help, or the therapeutic alliance. By contrast, Eysenck (1952) famously argued – based on clinical surveys – that psychotherapy produced little

benefit beyond what could be expected from spontaneous remission, leading him to question the very legitimacy of psychotherapy as a treatment modality. The methodological challenges of the time exacerbated the controversy: unlike pharmacological trials, psychotherapy studies could not implement double-blind designs. At the same time, the diversity of orientations and client populations complicated cross-study comparison.

Nevertheless, subsequent quantitative syntheses transformed the landscape. Smith and Glass (1977) introduced meta-analysis to psychotherapy research, followed by Smith, Glass, and Miller (1980), conclusively showing that psychotherapy was effective and superior to no treatment. This methodological breakthrough not only refuted Eysenck's scepticism but also legitimised psychotherapy as a scientific and clinical enterprise. By the 1960s and 1970s, researchers increasingly recognized that demonstrating the overall effectiveness of psychotherapy was not enough. The key question shifted from "Does it work?" to "How does it work?" and this led to a systematic study of process variables – the observable and subjective events that happen during therapy sessions and influence change (Strupp, 1986; Strupp & Howard, 1992). Howard and Strupp pioneered process research by examining therapist techniques, client responses, and interaction patterns, showing that what occurs in sessions can predict outcomes. However, definitions of process varied widely across different schools: psychodynamic approaches focused on transference, resistance, and countertransference, while cognitive-behavioural therapies emphasized collaborative empiricism, restructuring dysfunctional thoughts, and behaviour change strategies (Beck, 1979; McWilliams, 2011). This variety of theories enriched the field but also made direct comparisons more challenging.

The solution emerged with the creation of psychotherapy treatment manuals. By clearly defining therapeutic procedures, manuals enabled detailed comparisons across different methods and laid the groundwork for the growth of Randomized Controlled Trials (RCTs) (Chambless & Hollon, 1998; Nathan & Gorman, 2015). RCTs quickly became the gold standard in research, enabling causal conclusions and helping identify treatments supported by evidence (ESTs). Professional organizations increasingly endorsed lists of ESTs, thereby promoting accountability and helping to spread evidence-based practices (American Psychological Association, 2006). Although some critics argued that manualisation might oversimplify the complexity of therapy (Truijens et al., 2019), this period established a solid methodological framework. It shifted research focus toward measurable mechanisms of change.

From the late 1970s onward, the field adopted a process-focused paradigm, recognizing that psychotherapy's efficacy cannot be explained solely by outcomes but must be understood through the interplay among therapist, client, and context. Research attention turned to in-session phenomena

such as alliance formation, empathy, rupture–repair sequences, and the sequential organization of talk (Flückiger et al., 2018; Horvath et al., 2011). Longitudinal studies revealed that the quality of the therapeutic alliance is one of the most robust predictors of treatment success, often outweighing allegiance to a specific modality (Flückiger et al., 2018).

Meta-analyses of common factors – such as therapeutic alliance, goal consensus, and collaboration – reinforced the idea that therapeutic efficacy emerges from relational and contextual dynamics rather than technique-specific mechanisms (Flückiger et al., 2018). These findings gradually reshaped clinical training, shifting emphasis toward interpersonal skills and responsiveness rather than adherence to rigid protocols. Methodological innovation supported this shift. Microanalytic sequential methods allowed researchers to study therapy at the level of conversational turns and interactional moves, while significant-events analysis identified client-perceived moments of breakthrough (Llewelyn et al., 1988; Elliott & Shapiro, 1992; Oliveira, Sousa & Pires, 2012).

Elliott and colleagues mapped these diverse approaches into four complementary research traditions: quantitative process-outcome, qualitative inquiry, microanalysis, and significant-events methodology (Elliott, 2002). This pluralism bridged the gap between nomothetic generalization and idiographic understanding, striking a balance between empirical rigor and hermeneutic depth. In parallel, hermeneutic and phenomenological perspectives argued that psychotherapy research must account not only for observable behaviours but also for meaning-making processes, thereby integrating empirical and interpretive epistemologies (Hein & Austin, 2001; Lopez & Willis, 2004). This broadened the scope of research to include qualitative analyses of lived experience, narrative transformation, and the dialogical constitution of self (Lopez & Willis, 2004).

In recent decades, psychotherapy research has entered what some scholars describe as a fourth phase, characterized by the convergence of digitalization, computational methods, and ecological approaches to clinical practice. This phase expands the scope of inquiry by leveraging technological innovations and large-scale clinical data (Emmelkamp 2005; Schillinger et al., 2021). The widespread availability of electronic health records, digital archives of therapy transcripts, and session recordings has enabled researchers to assemble extensive corpora that far exceed the scale of traditional case studies or small-sample trials (Adjerid & Kelley, 2018). These big data offer unprecedented opportunities to model therapeutic dynamics with a certain ecological validity, capturing the complexity of everyday clinical practice and not relying solely on controlled experimental settings (Schillinger et al., 2021). So, advances in artificial intelligence and natural language processing (NLP) have become central to this development. Techniques such as topic modelling (e.g., Latent Dirichlet Allocation), sentiment analysis, and distributional semantics enable the scalable examination of

thematic patterns, affective trajectories, and stylistic markers across thousands of therapeutic interactions (Bedi et al., 2015; Cioffi et al., 2022). More recent applications of deep learning, including neural embeddings and transformer-based language models, further enhance the capacity to detect subtle relational and emotional processes unfolding in session dialogue (Cioffi et al., 2022). This computational turn has been described as a form of computational hermeneutics, insofar as it extends the interpretive sensitivity of clinical and qualitative traditions with data-driven, replicable methodologies. Rather than replacing the hermeneutic perspective, it integrates it, enabling the identification of process variables (e.g., alliance ruptures, repair sequences, shifts in self-narrative) on a scale and precision previously unattainable (Schillinger et al., 2021; Oudin et al., 2023).

Psychotherapy research has progressed through three overlapping but distinct phases. The first (1950s-1970s) established psychotherapy's general efficacy, overcoming scepticism through meta-analysis. The second (1960s-1980s) integrated process and outcome, institutionalising manuals and RCTs as methodological benchmarks. The third (1970s-present) shift focused on in-session processes, relational dynamics, and methodological pluralism, complemented in recent decades by computational innovations. By combining clinical sensitivity, hermeneutic insight, and computational scalability, research in psychotherapy provides nuanced accounts of how psychotherapy works, yielding knowledge of scientific and practical relevance (Cooper & McLeod, 2011; McLeod, 2016).

Process-Outcome Research: Definition and Significance

Process-outcome research examines the relationship between events that occur during psychotherapy sessions and clinical outcomes at the conclusion of treatment. It explores “both the events in psychotherapy sessions, or the constructs thought to change during or in between therapy sessions” and their association with subsequent change in “problems, symptoms and functioning” (Crits-Christoph, Gibbons, & Mukherjee, 2013). This research approach encompasses a broad range of process elements, including therapist techniques, client behaviours, interpersonal dynamics such as empathy and alliance, and broader service contexts (Llewelyn, Macdonald, & Aafjes-van Doorn, 2016). The defining feature of process-outcome research is its focus not only on whether psychotherapy works but on uncovering how change unfolds within therapy. The importance of this paradigm comes from its ability to address a key question in clinical psychology: why does psychotherapy work? While outcome studies determine whether psychotherapy is effective, process–outcome research offers insights into how change occurs. As Orlinsky, Rønnestad, and Willutzki argue, process variables are essential for understanding the conditions that lead to successful or unsuccessful therapy, helping refine interventions and improve clinician training. Empirical evidence

highlights the therapeutic bond, client engagement, and therapist empathy as predictors of positive outcomes across various treatment methods (Orlinsky, Rønnestad & Willutzki, 2004; Norcross & Lambert, 2011). Recent studies have further emphasized the core of the therapeutic alliance, showing that its quality is linked to session outcomes and treatment success (Flückiger et al., 2018).

A key strength of process-outcome research lies in its ecological validity. Rather than relying exclusively on highly standardized trials with rigid protocols, this approach investigates psychotherapy within its naturalistic context, where therapist and patient co-construct meanings over time (Llewelyn, Macdonald, & Aafjes-van Doorn, 2016). Consequently, process–outcome studies bridge the gap between empirical evidence and clinical practice, offering knowledge directly relevant to therapists in their everyday work (Crits-Christoph, Gibbons, & Mukherjee, 2013).

Methodological Approaches to Psychotherapy Process Research

The study of psychotherapy processes has necessitated the development of increasingly sophisticated methodologies designed to capture the complexity of therapeutic interactions. These methods combine qualitative and quantitative approaches, designed to link in-session phenomena with treatment outcomes (Llewelyn, Macdonald, & Aafjes-van Doorn, 2016).

Observational and Coding Systems. One of the earliest strategies for studying the therapeutic process was direct observation and systematic coding of therapist-patient interactions. Several influential instruments were developed during the 1960s and 1970s. For instance, the Psychotherapy Process Q-Set (PQS) (Jones, 1985) provides a structured method for describing therapeutic sessions through a comprehensive set of items rated by independent observers (Jones, Cumming, & Horowitz, 1988). Similarly, the Verbal Response Modes (VRM) model (Stiles, 1992) categorizes the therapist's interventions based on their communicative function (Stiles, 1992). These systems contributed to the standardisation of process research, enabling comparisons of sessions across different orientations and contexts. However, they also revealed certain limitations: the reliance on trained raters introduced subjectivity, and the time-consuming nature of coding restricted large-scale application.

Rating Scales and Self-Report Measures. In parallel with observational systems, researchers developed rating scales and self-report instruments to assess patients' and therapists' experiences of the therapeutic process. Among the most prominent tools is the Working Alliance Inventory (WAI) (Horvath & Greenberg, 1989), which operationalised Bordin's (1979) tripartite model of alliance – agreement on goals, collaboration on tasks, and the affective bond (Horvath & Greenberg, 1989). Such instruments enabled the empirical study of constructs such as therapeutic alliance, empathy, and rupture-repair sequences, which have been consistently linked to outcomes (Safran & Muran, 2000).

Self-report measures also enhanced the feasibility of large-scale studies by allowing patients and therapists to contribute data directly, without the need for external coders.

Micro-Analytic and Sequential Approaches. Another line of development has been micro-analysis of interaction. Inspired by conversation analysis and linguistic pragmatics, these approaches examine therapy as a sequence of turns at talk, focusing on how meaning is co-constructed moment by moment (Peräkylä et al., 2008). Sequential analysis allows the study of contingencies, such as how a therapist's empathic response increases the likelihood of patient disclosure or how specific interventions may lead to shifts in affective tone. Micro-analytic methods emphasize the performative dimension of language in psychotherapy: words not only represent experiences but also influence the therapeutic relationship by constraining or transforming dialogue (Buchholz & Kächele, 2017).

Text Mining and Computational Approaches. The digitalisation of clinical records and the diffusion of computational methods have transformed psychotherapy process research. Large corpora of therapy transcripts can now be subjected to natural language processing (NLP) techniques that detect thematic structures, semantic networks, and linguistic markers of change. For example, topic modelling approaches such as Latent Dirichlet Allocation (Blei, Ng, & Jordan, 2003) permit the extraction of latent themes across sessions, revealing how therapeutic conversations shift over time. Similarly, sentiment analysis and measures of linguistic style (Pennebaker, Boyd, Jordan, & Blackburn, 2015) provide indicators of affect regulation, cognitive restructuring, or narrative integration (Bedi et al., 2015). These computational techniques do not replace traditional qualitative analysis; instead, they extend it, offering novel perspectives on the dynamics of psychotherapy. By quantifying aspects of discourse that were previously accessible only through interpretive judgement, text mining creates new possibilities for linking process data with clinical outcomes (Mergenthaler & Stinson, 1992; Bedi et al., 2015; Corcoran & Cecchi, 2020).

Theoretical and Clinical Implications of Process–Outcome Research

Process-outcome research has reshaped the scientific understanding of psychotherapy, showing that therapeutic effectiveness cannot be explained solely by adopting a particular orientation or technique. Instead, efficacy emerges from a complex interplay of relational, cognitive, emotional, and linguistic processes that unfold over time within the therapeutic setting (Llewelyn, Macdonald, & Aafjes-van Doorn, 2016; Crits-Christoph, Gibbons, & Mukherjee, 2013). This paradigm has challenged reductionist explanations of change. Psychotherapy is increasingly conceptualized as a dynamic system in which multiple variables interact at different levels of analysis: individual characteristics of patient and therapist, dyadic relational patterns, and broader contextual and cultural factors (Hayes & Andrews, 2020; Schiepek et al., 2020). This systemic view reflects the

epistemological shift from linear cause-and-effect models to complex, non-linear perspectives (Castonguay et al., 2015; Hayes & Andrews, 2020; Schiepek et al., 2020).

Clinically, process-outcome research has practical consequences for training and intervention. By identifying the mechanisms that underlie therapeutic change – such as alliance formation, empathy, rupture-and-repair sequences, or discursive transformations – it becomes possible to refine therapist training and improve clinical effectiveness across orientations. Rather than focusing exclusively on technique, emphasis is placed on therapist responsiveness, flexibility, and the capacity to attune to the unfolding process (Castonguay et al., 2015; Safran & Muran, 2000).

In the field of psychotherapy research, a multidisciplinary orientation has been adopted, encompassing diverse methods and methodologies to improve both the understanding and the effectiveness of treatments (Castonguay et al., 2015). Within this broad framework, the present study falls within the branch of psychotherapy research known as process-outcome research (POR). In this area, it is increasingly recognized that quantitative and qualitative methodologies must converge toward deeper integration (Giorgi & Giorgi, 2003; Castonguay et al., 2015). Such integration is justified by the fact that the scientific method provides the tools to analyse large and complex datasets while accounting for multiple variables; however, these data must subsequently be interpreted within a conceptualization of the human being as embodied, relational, and situated in a historical context (Gelo & Salvatore, 2016; Gallagher, 2019). This epistemological stance is essential to avoid reducing research to the identification of psychotherapeutic processes or diagnostic traits, which may overlook the situated richness of lived experience across individual patients (Giorgi, 2009).

Building on this perspective, the current trajectory of psychotherapy research points toward an increasing integration between traditional process-outcome approaches and computational methods. The availability of large corpora of clinical transcripts, combined with advances in artificial intelligence (AI) and natural language processing (NLP), offers unprecedented opportunities to explore the micro-dynamics of therapeutic dialogue on a scale previously unimaginable (Llewelyn, Macdonald, & Aafjes-van Doorn, 2016; Miner et al., 2022).

Future research is therefore likely to emphasize hybrid methodologies. The primary challenge in the coming years will be to develop models that capture the complex, evolving, and situated nature of psychotherapy while maintaining rigorous methodological standards. In this respect, process-outcome research stands as a vital bridge between clinical practice and scientific inquiry, contributing to the ongoing refinement of psychotherapy as both an art and a science (Castonguay et al., 2015; Gelo & Salvatore, 2016).

Theoretical Foundations of Psychotherapy Research

Psychotherapy, as a scientific discipline, has developed at the crossroads of philosophy, psychology, and clinical practice. Its epistemological roots can be traced back to the late nineteenth and early twentieth centuries, when psychoanalysis, behaviourism, and phenomenology laid the foundation for understanding the human mind and therapeutic methods. Freud's psychoanalysis offered the first systematic model of therapy based on the interpretation of unconscious processes (Freud, 1917/1963). Behaviourism, on the other hand, focused on observable behaviour and how to change it through conditioning (Skinner, 1953). Later, phenomenology and existential psychology introduced a hermeneutic approach, emphasizing the importance of subjective experience and its meaning (Heidegger, 1927/1962; Ricoeur, 1976).

These diverse epistemological traditions converged in shaping psychotherapy research as both a clinical and an investigative enterprise. The tension between explanation (*Erklären*) and understanding (*Verstehen*), central to the human sciences, is particularly evident in psychotherapy, where empirical measurement and interpretive analysis coexist in a dynamic and sometimes conflictual relationship (Gadamer, 1960/2004; Polkinghorne, 1988). One of the most enduring debates in psychotherapy concerns the relationship between scientific evidence and clinical practice. On one hand, the evidence-based practice movement has promoted the standardization of interventions and the use of randomized controlled trials as the benchmark of efficacy (Chambless & Hollon, 1998). On the other hand, many scholars argue that psychotherapy, by its very nature, resists complete standardization because each therapeutic encounter is unique and shaped by the individual's context, biography, and interactions (Wampold & Imel, 2015).

This dialectic highlights a broader epistemological challenge: psychotherapy is both a nomothetic science, aiming for generalizable laws, and an idiographic practice, rooted in the uniqueness of individual lives (Polkinghorne, 1988). Hermeneutic approaches emphasize the importance of interpreting personal narratives and situating them within their historical and cultural contexts (Gadamer, 1960/2004). The challenge is to develop a research paradigm that can integrate these two perspectives. The increasing focus on mixed-methods designs and process-outcome studies precisely reflects this effort to balance empirical rigor with the complexity of lived experience (Castonguay et al., 2015; Giorgi & Giorgi, 2003).

Hermeneutics, originally conceived as the art of interpreting texts, has played a decisive role in the development of contemporary psychotherapy. Its fundamental premise is that human experience is always mediated by language and that understanding involves interpretation rather than

mere observation (Ricoeur, 1981). This orientation contrasts with positivist epistemologies, which prioritize measurement and causal explanation.

Martin Heidegger (1927/1962) transformed hermeneutics from a methodological approach to an ontological inquiry, arguing in *Being and Time* that human existence (Dasein) is fundamentally interpretive. To exist is to understand oneself within a horizon of meanings shaped by history and culture – a perspective with profound implications for psychotherapy. Therapeutic work thus transcends mere correction of dysfunctional cognitions or behavioural modification; it involves the reinterpretation of one's being-in-the-world (Heidegger, 1927/1962; Palmer, 1969).

Hans-Georg Gadamer (1960/2004) further advanced hermeneutics by emphasizing dialogue as the medium of understanding. He proposed that interpretation arises from a “fusion of horizons,” in which the perspectives of interlocutors meet and transform one another. Applied to psychotherapy, this framework highlights the inherently dialogical nature of the therapeutic encounter, in which meaning emerges from the co-construction of understanding rather than from the unilateral application of technique (Gadamer, 1960/2004; Malpas, 2010).

Paul Ricoeur (1976, 1986) extended hermeneutics into the realm of narrative and textual analysis. He distinguished between explanation and understanding, depicting interpretation as a dialectical interplay between structural analysis and phenomenological comprehension. From this vantage point, psychotherapy can be viewed as a narrative reconfiguration, where patients re-author their life stories through therapeutic dialogue (Ricoeur, 1976, 1986).

This emphasis on textuality establishes a direct connection to process research that analyzes therapy transcripts. Treating discourse as a “fixed text” enables systematic linguistic study while preserving interpretive meaning, inspiring hybrid methodologies that merge computational analysis with hermeneutic interpretation (Llewelyn, Macdonald, & Aafjes-van Doorn, 2016). The epistemological positioning of psychotherapy has long been a subject of contention. Is it a natural science bound by objectivity and replicability, or a human science where interpretation, context, and meaning predominate? Contemporary consensus increasingly views psychotherapy as a hybrid discipline that integrates scientific rigor and hermeneutic depth (Gallagher, 2004; Castonguay et al., 2015).

Attempts to strictly align psychotherapy with natural science paradigms confront limitations. Randomized controlled trials (RCTs), while indispensable for establishing efficacy, often oversimplify complex therapeutic processes, exclude patients with comorbidities, restrict therapist flexibility, and reduce treatment to manualized procedures, thereby compromising ecological validity (Kazdin, 2008; Wampold & Imel, 2015). A hermeneutic stance insists that psychotherapy cannot be

abstracted from the interpretive processes through which meaning is negotiated. Each therapeutic dialogue is embedded within the histories, values, and lifeworlds of both the therapist and the patient. Understanding, as Gadamer (1960/2004) argued, is always situated within traditions and shaped by pre-understandings (*Vorurteil*) in its original sense (Gadamer, 1960/2004; Palmer, 1969). Therefore, psychotherapy demands examination not merely as a technique set but as a meaning-making process. Process-outcome research exemplifies this synthesis by linking measurable changes to the evolving subjective and relational processes within therapy (Castonguay et al., 2015; Ricoeur, 1986).

Language, Discourse, and the Therapeutic Dialogue

Language constitutes the essential medium of psychotherapy. Regardless of theoretical orientation – psychoanalytic, cognitive-behavioural, systemic, or humanistic – the therapeutic encounter is primarily realized through spoken words, narratives, and discursive exchanges. The patient's experience is brought into the session through language, and the therapist's task is to respond, interpret, and co-construct meaning within the dialogue. Discourse is not merely a representation of inner states, but an active process that shapes and transforms experience. As Ricoeur (1986) observed, speech always goes beyond itself, projecting new possibilities of meaning. In psychotherapy, the act of narrating past events or articulating emotions can itself initiate transformation, as language reconfigures the relation between self and world (Ricoeur, 1986; Ricoeur, 1976). The therapeutic dialogue thus functions as both a medium of reflection and an instrument of change. Through processes such as re-narration, metaphorical elaboration, and interpretive reframing, patients integrate fragmented experiences into more coherent and adaptive life stories (White & Epston, 1990; Angus, Levitt, & Hardtke, 1999).

Contemporary conversation-analytic approaches emphasize that therapeutic dialogue is co-constructed between therapist and patient (Peräkylä et al., 2008). The meaning of an utterance cannot be understood in isolation but depends on its sequential position and the interlocutor's response. For example, a patient's expression of despair may be met by empathic validation, interpretive inquiry, or cognitive reframing, each shaping the trajectory of subsequent discourse in distinct ways. This sequential co-construction underscores the interactive nature of meaning-making: therapy is not merely the application of techniques to a passive recipient, but an ongoing negotiation in which both parties contribute (Peräkylä et al., 2008). Once transcribed, therapeutic dialogue becomes a text that can be subjected to both hermeneutic interpretation and systematic analysis. Transcripts capture not only content but also interactional structures – such as pauses, overlaps, and shifts in tone – that reveal the microdynamics of the encounter (Llewelyn, Macdonald, & Aafjes-van Doorn, 2016). They

provide a unique resource for linking linguistic features to therapeutic outcomes, forming the empirical basis of much process-outcome research.

The textual dimension also permits the integration of computational methods such as topic modelling, sentiment analysis, and semantic network analysis. These tools, while rooted in quantitative traditions, are particularly suited to the hermeneutic orientation, as they illuminate how meaning emerges and transforms within discourse (Blei, Ng, & Jordan, 2003; Pennebaker et al., 2015; Bedi et al., 2015). In text interpretation, particularly within natural language processing (NLP) and linguistic analysis, a text can be analysed at multiple levels (Smith & Johnson, 2022). The main levels comprise eight categories, forming a comprehensive framework for comprehending and interpreting text in computational and human linguistic analysis (Table 1).

Semantic Analysis is essential for uncovering the deeper meanings and themes within patients' narratives in psychotherapy research (Anderson et al., 1999). By analysing the semantic content of therapeutic dialogues, researchers and clinicians can gain valuable insights into patients' emotional states, cognitive patterns, and the meanings they attribute to their experiences (Imel et al., 2015; Gaut et al., 2017; Udvardi, 2019). Therefore, SA is an interpretative method that can greatly benefit from hermeneutics – the science of interpretation – both in its philosophical-ontological characterization (as in the works of Heidegger and Gadamer) and in its more epistemological aspects (such as those explored by Ricoeur). Contemporary hermeneutics is a methodology for understanding how humans develop their understanding of the world and their experiences within it, based on constantly interpreting the information provided by their contexts, senses, thoughts, and feelings (Gadamer, 1975; Gadamer, 2004; Davidson & Vallée, 2021).

For Ricoeur, interpretative acts serve as the bridge between language and lived experience (Ricoeur, 1992). This notion is particularly relevant in the context of analysing psychotherapy session transcripts, as it involves translating lived experiences into verbal expressions (Imel et al., 2015; Gaut et al., 2017). Taking a hermeneutic-phenomenological approach, the therapist's task is to interpret the patient's narrative in a manner that uncovers their ways of being (Arciero et al., 2012; Liccione, 2019; Fuchs, 2001; de Haan, 2022). A circular causal relationship exists between self-narrative and lived experience, with psychotherapeutic interpretation significantly shaping this experience. Interpretive understanding creates the potential to perceive things differently and to orient oneself in alternative ways in the world. This interconnection between knowledge and understanding forms the cornerstone of Ricoeur's theory of interpretation (Ricoeur, 1981; Davidson & Vallée, 2021).

1) Lexical Analysis	Focuses on the words themselves, their structure, and meaning. It involves identifying words, their forms (e.g., root words, prefixes, suffixes), and their role in the text. This is the most basic level, dealing with individual words and their dictionary definitions.
2) Syntactic Analysis	Deals with the structure of sentences. This involves analyzing the grammatical structure, understanding how words combine to form phrases and sentences, and parsing the sentence to understand the relationships between different parts of the sentence.
3) Semantic Analysis	Focuses on the meaning of words and sentences. It goes beyond syntax to understand the actual meaning conveyed by the text, including the roles of words in context, the meaning of phrases, and the overall message or proposition expressed by the text.
4) Pragmatic Analysis	Concerned with understanding how context influences the interpretation of meaning. Pragmatic analysis considers the speaker's intent, the relationship between the speaker and listener, and how the context (such as the situation or cultural background) affects the interpretation of the text.
5) Discourse Analysis	Examines how sentences and larger units of text (like paragraphs or entire documents) connect and interact to create meaning. It involves analyzing the flow of text, coherence, and how meaning is constructed across larger sections of text.
6) Stylistic Analysis	Looks at the style of the text, including the choice of words, tone, and rhetorical devices. It helps in understanding the text's aesthetic qualities and how the author's style contributes to meaning and effect.
7) Phonological Analysis	Though more relevant in spoken language, phonological analysis examines how sounds contribute to meaning. In written text, it might involve analyzing rhyme, rhythm, and sound patterns, especially in poetry.
8) Morphological Analysis	Focuses on the structure of words, examining how words are formed from morphemes (the smallest units of meaning) and how these affect the meaning of the text.

Table 1. Main Levels of Textual Analysis. This table presents eight key levels of textual analysis, providing a framework for interpreting text, particularly in psychotherapy research.

Textuality and the Role of Transcripts in Psychotherapy

Language serves as the primary medium of psychotherapy. Regardless of orientation, psychotherapeutic work is fundamentally centered on spoken and written discourse, where experiences are narrated, interpreted, and redefined. Ricoeur (1986) described language as a hermeneutic act, emphasizing that discourse is not a neutral tool but a transformative process through which new meanings emerge (Ricoeur, 1981, 1986). In psychotherapy, the importance of discourse means that the textual aspect of clinical interactions becomes a focus. Transcripts of therapy sessions are especially valuable for process-outcome research because they record the step-by-step flow of dialogue (Mergenthaler & Stinson, 1992). By examining these texts, researchers can identify thematic patterns, recurring linguistic markers, and interaction styles associated with therapeutic progress (Llewelyn, Macdonald, & Aafjes-van Doorn, 2016).

The methodological potential of transcripts has greatly increased with the rise of computational linguistics and natural language processing (NLP). Automated analyses now enable the detection of hidden themes, semantic areas, and discursive changes that would be hard to identify through manual coding alone (Rumshisky et al., 2016; Miner et al., 2022). Consequently, the study of textuality has become more central in modern psychotherapy research, blending qualitative and quantitative methods in innovative ways.

Integrating Hermeneutics and Empirical Research

The current landscape of psychotherapy research increasingly emphasizes the need to combine hermeneutic and empirical approaches. Each perspective alone is inadequate: hermeneutics without empirical evidence risks remaining purely speculative, while empirical research without interpretive depth risks reducing the therapeutic process to decontextualized variables (Hein & Austin, 2001; Gadamer, 1960/2004). Hermeneutics enhances understanding by placing clinical phenomena within personal, historical, and cultural contexts. Empirical research broadens knowledge by identifying patterns, regularities, and predictive relationships. Combining these paradigms enables a more comprehensive understanding of psychotherapy (Castonguay et al., 2015). Mixed-methods research represents a practical embodiment of this integration. Studies that combine qualitative narrative analysis with quantitative outcome measures allow researchers to trace how changes in discourse correspond with improvements in symptomatology or well-being (Creswell & Plano Clark, 2017). For example, micro-analytic studies of rupture-repair sequences, articulated initially by Safran and Muran (2000), can be paired with longitudinal outcome data to examine their predictive validity (Safran & Muran, 2000).

Recent advances in natural language processing (NLP) have opened possibilities for what might be termed computational hermeneutics. Large-scale analyses of therapy transcripts can reveal latent structures of meaning – topics, semantic fields, affective trajectories – that resonate with hermeneutic concerns about interpretation and narrative (Blei, Ng, & Jordan, 2003; Pennebaker, Boyd, Jordan, & Blackburn, 2015). Unlike traditional coding, computational approaches enable analysis of thousands of sessions, thereby enhancing generalizability while preserving attention to meaning (Llewelyn, Macdonald, & Aafjes-van Doorn, 2016).

The goal is not the dominance of one paradigm over the other but the creation of an integrated epistemology. In this framework, empirical findings are interpreted through hermeneutic lenses, while hermeneutic insights inform the formulation of empirical hypotheses. This bidirectional relationship ensures that psychotherapy research remains both scientifically rigorous and faithful to the lived experience of patients and therapists (Creswell & Plano Clark, 2017; Gadamer, 1960/2004).

1.1. Two Paradigms in Psychotherapy Research: Quantitative and Qualitative

Psychotherapy research has historically evolved within a methodological tension between quantitative and qualitative approaches. This duality reflects broader epistemological debates regarding the status of psychotherapy as a discipline: whether it should be aligned with the natural sciences (*Naturwissenschaften*), where general laws and causal mechanisms are privileged, or with the human sciences (*Geisteswissenschaften*), where interpretation, context, and meaning are central (Guba & Lincoln, 1994; Gelo, Braakmann & Benetka, 2008; Gelo, Salcuni & Colli, 2012).

At its core, the primary aim of psychotherapy research as a scientific discipline is to gather as much knowledge as possible about “*what works for whom, under what conditions, and why*” (Paul, 1967; Roth & Fonagy, 2005). Within this scientific framework, three main methodological traditions have been identified: quantitative research, qualitative research, and theoretical research (Gelo et al., 2020). Each of these traditions is grounded in a specific epistemological stance and methodological principle, and together they constitute the pluralistic landscape of psychotherapy research.

Quantitative Research

Quantitative research, rooted in the natural sciences and specifically in the post-positivist tradition that dominated social science research since the early 20th century (Guba & Lincoln, 1994; Gelo, Salcuni & Colli, 2012), involves the systematic collection and statistical analysis of numerical data. Its methodological principles emphasise objectivity, explanation, deduction, replicability, and

generalisation (Gelo et al., 2020). This tradition emerged in response to Eysenck's (1952) challenge to the efficacy of psychotherapy, representing an effort to demonstrate the feasibility and necessity of applying scientific methods to the study of psychotherapy (Orlinsky, Grawe & Parks, 1994).

Among the most widely used quantitative methods in psychotherapy research are Randomised Controlled Trials (RCTs), longitudinal cohort studies, meta-analyses, and standardised psychometric assessments. RCTs have historically been considered the *gold standard* for demonstrating the efficacy of psychotherapeutic interventions. For instance, early RCTs comparing Cognitive-Behavioural Therapy (CBT) with pharmacological treatments for depression established the empirical foundation for CBT as an effective intervention (Rush et al., 1977; Beck, 1979). RCTs investigating exposure therapy for anxiety disorders have consistently provided strong evidence of efficacy (Foa et al., 2005).

Rooted in the biomedical model, RCTs are structured to maximise internal validity by isolating causal relationships between intervention and outcome (Chambless & Hollon, 1998). They prioritize standardization, replicability, and statistical generalizability, thereby enabling the accumulation of a robust evidence base that has shaped the evidence-based practice movement in contemporary clinical psychology and psychiatry. Beyond RCTs, longitudinal studies have made significant contributions by examining trajectories of change over extended periods. For example, the Penn Psychotherapy Project (Luborsky, Singer & Luborsky, 1975) employed repeated psychometric measures to track symptom improvement and therapeutic alliance across sessions, offering insights into the temporal unfolding of therapeutic processes. Meta-analyses and systematic reviews further aggregate findings across studies, as exemplified by the landmark analysis of Smith, Glass, and Miller (1980), which demonstrated that psychotherapy, across modalities, is more effective than no treatment.

Nevertheless, the limitations of these paradigms are increasingly acknowledged. The strict inclusion criteria of RCTs often exclude patients with comorbidities or complex psychosocial conditions, thereby reducing ecological validity and limiting generalisability to real-world practice. Moreover, reliance on manualised treatments, while ensuring replicability, may constrain therapist responsiveness and obscure relational and contextual dimensions of therapy. Quantitative approaches also privilege symptom reduction as measured by instruments such as the Beck Depression Inventory-II (BDI-II) (Beck, Steer, & Brown, 1996) or the Hamilton Rating Scale for Anxiety (HAM-A) (Hamilton, 1959). While invaluable for outcome assessment, such tools capture only a partial dimension of change, leaving out phenomena such as narrative coherence, reflexivity, or relational atonement (Kazdin, 2008). Research on narrative processes (Angus & McLeod, 2004) illustrates how patients may experience profound change through re-authoring personal stories, even when symptom scores do not reflect dramatic improvements.

Finally, quantitative methods may risk perpetuating a reductionist perspective by operationalising complex psychological phenomena into discrete variables. While useful for statistical modelling, this approach can obscure the emergent, dialogical, and contextual nature of therapeutic change. As critics have argued, psychotherapy should not be reduced to a set of techniques applied to passive patients, but rather understood as a dynamic, co-constructed process in which relational and interpretive factors are central (Norcross & Wampold, 2011). Advances in NLP have attempted to address this gap. For instance, studies by Imel et al. (2015) utilized machine learning to analyse session transcripts, automatically detecting therapist empathy, and demonstrating how quantitative methods can be adapted to capture relational phenomena. These developments suggest the emergence of a more integrative quantitative science, one that retains statistical rigour while engaging with the complexity of therapeutic dialogue.

Qualitative Research

In contrast, qualitative research is rooted in the human sciences and closely associated with constructivist-interpretivist and critical-ideological traditions that emerged in the 1970s (Polkinghorne, 1988; Guba & Lincoln, 1994; Gelo, Salcuni & Colli, 2012). Qualitative inquiry emphasises the subjective, relational, and narrative aspects of psychotherapy, investigating how therapists and patients construct meaning, negotiate identities, and transform experiences within the therapeutic dialogue (Elliott, Fischer, & Rennie, 1999). It collects and analyses non-numerical data – such as interview transcripts, narratives, images, and conversational interactions – using hermeneutic, phenomenological, or narrative frameworks (Rennie, 2012). Its methodological principles include subjectivity, induction, abduction, contextualization, uniqueness, and understanding (Gelo et al., 2020). Rather than seeking statistical generalization, qualitative methods aim to capture the richness and uniqueness of lived experience, to understand phenomena in their factual and relational contexts (Gelo, Braakmann, & Benetka, 2008).

Among the most established qualitative approaches in psychotherapy research are phenomenological and hermeneutic methods, which aim to capture clients' subjective experiences and the meaning structures that emerge in therapy. Phenomenological studies, such as those conducted by Giorgi (1994), emphasise careful description of lived experience, while hermeneutic approaches inspired by Gadamer (1975) and Ricoeur (1981) seek to interpret therapeutic dialogue as a process of meaning-making. These methods have provided critical insights into how clients construct and transform personal narratives, highlighting psychotherapy as a fundamentally interpretive practice.

Another important family of methods includes grounded theory and thematic analysis, which generate theoretical models directly from qualitative data. Grounded theory has been widely used to explore

therapeutic mechanisms, such as how clients negotiate identity transformation during therapy (Charmaz, 2006). For example, Rennie (1994) employed grounded theory interviews with clients to investigate their in-session experiences, revealing the role of tacit negotiation and reflexivity in the therapeutic process. Thematic analysis, on the other hand, offers a flexible tool for identifying patterns in client discourse, making it particularly suitable for exploring common themes across diverse therapeutic contexts (Braun & Clarke, 2006).

Narrative and discourse analytic methods represent another strand of qualitative research, focusing on how stories and language use shape therapeutic change. Angus and McLeod (2004), for instance, demonstrated how narrative re-authoring processes allow clients to reorganise past experiences and develop more coherent life stories. Similarly, conversation analysis has been applied to psychotherapy transcripts to investigate turn-taking, interruptions, and repair sequences, revealing how the microstructure of dialogue supports or hinders alliance formation (Peräkylä et al., 2008). These methods illuminate the interactional dynamics of therapy, capturing nuances that are often invisible to quantitative measurement.

A further development has been the study of rupture and repair sequences in the therapeutic alliance, which relies heavily on qualitative microanalysis of session transcripts. Safran and Muran (2000) demonstrated how qualitative coding of interactions can uncover the processes by which therapist and client collaboratively overcome misunderstandings. These findings underscore the central role of responsiveness and negotiation in sustaining therapeutic progress—dimensions often marginalised in manualised, quantitatively driven approaches.

Despite their richness, qualitative methods are sometimes criticised for limited generalisability and lack of standardisation. However, frameworks such as Elliott, Fischer, and Rennie's (1999) *Guidelines for Qualitative Research in Psychology* have established methodological criteria for ensuring rigour, transparency, and credibility. Moreover, qualitative findings often complement quantitative data, offering explanatory depth that clarifies how and why therapeutic change occurs. For instance, studies that integrate qualitative interviews with quantitative symptom measures have shown that clients may report increased self-understanding and relational satisfaction even when symptom reduction is modest (McLeod, 2011).

Finally, qualitative inquiry has been enriched by the integration of technological and computational tools. With the growing availability of psychotherapy transcripts, methods such as computer-assisted qualitative data analysis (CAQDAS) enable systematic coding and thematic exploration across large datasets. Combined with natural language processing, these approaches open new possibilities for hybrid methodologies that retain interpretive depth while expanding the scale of

analysis (Edwards, 2020). Such innovations point toward a future in which qualitative research not only complements but actively integrates with quantitative frameworks in the service of a pluralistic science of psychotherapy.

Theoretical Research

Alongside empirical approaches, theoretical research plays a crucial role in psychotherapy science. Based on the idea that humans attain knowledge primarily through reasoning (Feyerabend, 1981), theoretical inquiry constructs, extends, and revises the conceptual frameworks underpinning psychotherapy. It provides the models that empirical research tests, and in turn integrates findings into broader theoretical perspectives (Gelo et al., 2020).

Toward Methodological Pluralism

The emphasis on common factors aligns with qualitative and hermeneutic traditions, which foreground the relational and narrative dynamics of therapy. As Wampold and Imel (2015) note, much of the variability in treatment outcomes can be attributed to processes embedded in dialogue, empathy, and meaning-making rather than to the application of discrete techniques. Thus, methodological pluralism, that integrating quantitative rigour with qualitative depth, offers the most promising framework for capturing these multifaceted processes.

The field of process-outcome research represents a key effort to bridge the quantitative-qualitative divide. It investigates how specific processes observed in therapy sessions contribute to therapeutic outcomes, combining quantitative measures of change with qualitative or interpretive analyses of interactional dynamics. According to Elliott (2002), process research outcomes can be examined through four complementary approaches: quantitative process-outcome studies, qualitative exploration of factors, microanalytic sequential methods, and the significant events approach. Quantitative designs typically sample process variables and examine their statistical association with outcomes (Hill, 2011; McLeod, 2013), whereas qualitative approaches unravel clients' lived experiences, often through self-reported perspectives (Levitt, Pomerville & Surace, 2016). Studies of rupture-repair sequences (Safran & Muran, 2000) or linguistic markers of emotional expression (Pennebaker, 2011) exemplify how micro-processes in dialogue can predict long-term outcomes.

Process-outcome research provides a fertile ground for addressing the questions raised in the debate between specific techniques and common factors. Wampold and Imel (2015) argue that factors such as the therapeutic alliance, rupture-repair sequences, and therapist variability account for a greater proportion of variance in outcomes than specific model-based interventions. Empirical process-outcome studies substantiate this claim by demonstrating that the quality of the therapeutic

relationship and the capacity to co-construct meaning are robust predictors of change across treatment modalities. In this respect, process–outcome research not only bridges methodological paradigms but also helps reframe psychotherapy as a fundamentally relational and contextual practice.

Despite their contributions, both quantitative and qualitative methods have limitations in capturing the complex, multifaceted nature of psychotherapy processes. As Atkins et al. (2014) note, even robust designs struggle to accommodate the dense, dynamic data present in therapeutic interactions. The availability of verbatim therapy transcripts has further enriched this field, transforming conversations into analysable texts. These can be approached hermeneutically to explore lived meanings or subjected to systematic examination through computational techniques. Recent technological advances, particularly in Artificial Intelligence (AI), have opened novel pathways: natural language processing (NLP), machine learning, topic modelling, semantic network analysis, and sentiment analysis enable large-scale examinations of discursive structures and allow detection of psychological processes and change markers that would otherwise remain inaccessible (Smink et al., 2019). This convergence exemplifies the growing call for mixed-methods designs, which integrate the statistical robustness of quantitative approaches with the interpretive depth and contextual sensitivity of qualitative inquiry (Stiles, 1980; Creswell & Plano Clark, 2017). Such methodological pluralism reflects the dual nature of psychotherapy: it is simultaneously a measurable intervention and a hermeneutic dialogue. On the one hand, outcome measures such as symptom reduction provide evidence of therapeutic efficacy; on the other hand, qualitative insights into narrative transformation and meaning-making reveal how such change is co-constructed in dialogue. By uniting these perspectives, process–outcome research moves beyond paradigm polarisation, offering a comprehensive and ecologically valid understanding of therapeutic change.

The apparent dichotomy between quantitative and qualitative paradigms has thus increasingly given way to methodological pluralism. Contemporary psychotherapy research acknowledges that no single method can fully capture the multifaceted and multidimensional nature of therapeutic change. Instead, integration is sought through process–outcome studies, mixed-methods designs, and computational linguistics, which unite evidence with meaning, measurement with interpretation (Stiles, 1980; Creswell & Plano Clark, 2017). For example, symptom reduction assessed through standardised instruments can be correlated with discourse-analytic findings from therapy transcripts, allowing researchers to trace how specific linguistic or narrative processes predict clinical improvement. In this way, qualitative insights into meaning-making are not opposed to quantitative outcomes but rather complement and enrich them. Studies have shown that increases in patients' emotional expression, cognitive reorganisation, or narrative complexity are predictive of positive therapeutic outcomes (Pennebaker, 2011; Angus et al., 2012).

The availability of session transcripts has further facilitated the merging of different methods. Once transcribed, therapeutic dialogues become texts that can be analysed both through hermeneutic interpretation and systematic methods. Natural language processing techniques – such as topic modelling, semantic network analysis, or sentiment analysis – make it possible to identify hidden structures of meaning, emotional patterns, and thematic trends across extensive collections of data. These computational tools incorporate elements traditionally seen in hermeneutic work, helping create what could be called computational hermeneutics.

In this context, psychotherapy research emerges not only as a branch of applied science but also as a field where epistemological diversity is essential. Quantitative approaches offer rigor, generalizability, and predictive ability; qualitative methods provide depth, contextual richness, and focus on meaning; and theoretical inquiry establishes the conceptual foundation for integration. Positioned between the poles of quantitative and qualitative methodologies, psychotherapy research is increasingly moving toward synthesis. Process–outcome research and mixed-methods design exemplify this trend, demonstrating that therapeutic change can be studied both as a measurable phenomenon and as a hermeneutic dialogue (Castonguay et al., 2015; Creswell & Plano Clark, 2017; Llewelyn, Macdonald, & Aafjes-van Doorn, 2016). This pluralism reflects the inherent complexity of psychotherapy, providing a more comprehensive framework for advancing both science and clinical practice (Levitt et al., 2017)

2. Artificial Intelligence and Natural Language Processing

This chapter introduces the foundational concepts of artificial intelligence (AI) to contextualise its relevance for the present dissertation. It is organised into four major sections: artificial intelligence, natural language processing, text mining and text analysis, and generative AI.

By the end of the chapter, the reader will have an overview of:

- *What AI is, how it has evolved, its key systems, and their primary domains of application, with a particular focus on the exponential growth of AI in both research and commerce.*
- *The nature of natural language processing (NLP), its key tools, main applications, and the current state of research and technological development.*
- *a focused discussion on text mining and text analysis as two central applications of NLP, highlighting both their methodological distinctions and complementarities.*
- *An outline of generative AI, including the role of large language models (LLMs), their development, domains of application, rapid expansion in research and industry, and the emerging field of prompt engineering.*

The purpose of this chapter is to provide a conceptual and methodological framework for understanding what AI can achieve and the principal models it encompasses. This overview serves as the basis for the following chapter, which explores how computational analysis and psychotherapy may be integrated to advance both research and clinical progress in understanding processes of change in psychotherapy.

Artificial Intelligence (AI)

The concept of artificial intelligence (AI) has evolved significantly since its initial formulation, reflecting both advances in computational science and changing understandings of intelligence itself. At its core, AI is defined as the field of computer science dedicated to developing systems capable of performing tasks that typically require human intelligence, such as reasoning, learning, perception, decision-making, and natural language understanding (Russell & Norvig, 2021). This broad definition encompasses a variety of paradigms, from symbolic rule-based reasoning to statistical machine learning and, more recently, deep neural networks. The term “artificial intelligence” was first coined by John McCarthy at the 1956 Dartmouth Conference, widely regarded as the discipline's foundational moment (McCarthy, Minsky, Rochester, & Shannon, 2006;

Dartmouth College, 2022). Intelligence was initially conceptualized as a computational process: the manipulation of symbols to simulate aspects of human reasoning. Early pioneers, including John McCarthy, Marvin Minsky, Allen Newell, and Herbert Simon, envisioned encoding human thought into formal logical rules that machines could execute (Crevier, 1993).

This early symbolic paradigm, later commonly referred to as “Good Old-Fashioned AI” (GOFAI), focused on logic, explicit knowledge representation, and rule-based systems that are fundamental to knowledge engineering, automated reasoning, and expert systems (Haugeland, 1985; Russell & Norvig, 2021). Although GOFAI established a vital foundation for computer-based intelligence, the following decades saw the incorporation of probabilistic, statistical, and neural network methods as the concept of “intelligence” itself expanded within scientific and philosophical discussions. From a conceptual standpoint, AI is not merely a technological artefact but also a theoretical construct situated at the intersection of philosophy, mathematics, and cognitive science. The attempt to formalise intelligence raises foundational questions about the nature of thought, the possibility of mechanising reasoning, and the relationship between human and artificial cognition. Turing’s seminal paper, “Computing Machinery and Intelligence” (1950), addressed the problem of machine intelligence through the notion of functional equivalence: a machine could be considered intelligent if its performance was indistinguishable from that of a human, as operationalised by the “Imitation Game” or Turing Test (Turing, 1950). While later criticised for its behaviourist and operationalist limits, the Turing Test remains central to ongoing debates on the nature and evaluation of AI.

In contemporary scholarship, AI is increasingly envisioned as a multi-layered system involving perception, learning, and adaptation. Kaplan and Haenlein (2019) define AI as “a system’s ability to correctly interpret external data, to learn from such data, and to use those learnings to achieve specific goals and tasks through flexible adaptation” (Kaplan & Haenlein, 2019, p. 17). This definition foregrounds the adaptive, dynamic, and data-driven aspects of modern AI systems, clearly distinguishing them from earlier rule-based or symbolic approaches.

The conceptual foundations of AI also include the distinction between narrow AI and general AI. Narrow AI (or weak AI) refers to systems designed to perform specific tasks, such as image recognition, speech processing, or chess playing. These systems often outperform humans in narrowly defined domains but lack the capacity for generalisation beyond their training data. In contrast, Artificial General Intelligence (AGI) denotes the hypothetical development of machines with human-level cognitive flexibility, capable of transferring knowledge across tasks and contexts (Goertzel & Pennachin, 2007). While AGI remains largely aspirational, the pursuit of general intelligence continues to inform both philosophical and technical debates, shaping long-term visions of AI

development. Moreover, the theoretical landscape of AI is marked by different paradigmatic approaches. The symbolic paradigm emphasises logical reasoning and explicit knowledge representation, whereas the connectionist paradigm, emerging in the 1980s with artificial neural networks, focuses on distributed representations and learning through weight adjustment (Rumelhart, Hinton, & McClelland, 1986). The subsequent rise of machine learning shifted the emphasis from hand-coded rules to data-driven algorithms that identify patterns and make predictions. In the 2010s, the advent of deep learning revolutionised the field, leveraging multilayered neural architectures to achieve unprecedented performance in tasks such as image classification (Krizhevsky, Sutskever, & Hinton, 2012) and natural language processing (Vaswani et al., 2017). Another conceptual foundation of AI lies in its interdisciplinary nature. From its inception, AI has drawn upon insights from mathematics (formal logic, probability theory, optimisation), cognitive psychology, neuroscience, and linguistics. This interdisciplinarity explains both the richness and the complexity of the field, as well as its rapid expansion into diverse areas of application (Russell & Norvig, 2021; Crevier, 1993).

Ultimately, it is crucial to recognize that definitions of AI are not purely technical but are also socially and historically situated. Floridi and Chiriatti (2020) argue that AI should be understood as an “artefact of intelligence,” reflecting both technological affordances and societal expectations about automation, intelligence, and human–machine interaction. As such, the conceptualisation of AI evolves alongside cultural imaginaries, ethical debates, and political priorities. For example, the exponential growth of AI research in the twenty-first century is not only a function of computational advances but also of the increasing societal demand for automation, personalisation, and predictive analytics in fields such as healthcare, education, and security (Brynjolfsson & McAfee, 2017).

Historical Development and Key Milestones of Artificial Intelligence

The historical trajectory of artificial intelligence (AI) reflects alternating phases of optimism and disillusionment, often described as cycles of boom and “AI winters” (Crevier, 1993). From its inception in the mid-twentieth century to the contemporary era of deep learning, the field has undergone profound conceptual and technological transformations that have shaped both research agendas and societal imaginaries.

The origins of AI can be traced to the 1940s and 1950s, when theoretical advances in mathematics, logic, and computer science first suggested the feasibility of machine-based intelligence. Alan Turing’s seminal essay *Computing Machinery and Intelligence* (1950) introduced the idea that digital computers could simulate any process of formal reasoning, culminating in the proposal of the Turing Test as an operational criterion for machine intelligence. Developments in

cybernetics, led by Norbert Wiener (1948), conceptualised intelligent behaviour in terms of feedback systems and information theory, laying the groundwork for computational models of cognition.

The Dartmouth Conference of 1956 is widely recognised as the official birth of AI as a research discipline. Organised by John McCarthy, Marvin Minsky, Claude Shannon, and Nathaniel Rochester, the conference proposed the ambitious goal of enabling machines to “simulate every aspect of learning or any other feature of intelligence” (McCarthy et al., 2006). This event consolidated the symbolic paradigm, privileging rule-based approaches that sought to encode intelligence through logical formalisms and heuristic search. Early successes, such as Newell and Simon’s Logic Theorist (1956) and General Problem Solver (1957), demonstrated the potential of symbolic reasoning systems to solve complex mathematical and logical problems.

The 1960s and 1970s saw both growth and setbacks. Research in natural language processing led to early systems, such as ELIZA (Weizenbaum, 1966), a program that mimicked Rogerian psychotherapy by matching patterns in user inputs. At the same time, expert systems began to emerge, with applications in medical diagnosis (e.g., MYCIN, Shortliffe, 1976) and chemical analysis (e.g., DENDRAL, Buchanan & Feigenbaum, 1978). These systems demonstrated the utility of symbolic AI in specialized fields, although their reliance on manually created rules limited their potential for expansion. Despite these successes, unrealistic expectations led to disillusionment when progress slowed. The first AI winter, beginning in the mid-1970s, was characterized by decreasing funding and scepticism about AI’s potential.

The 1980s experienced a revival with the rise of expert systems and the resurgence of connectionism. Driven by commercial interests, expert systems gained popularity in industries looking for automated decision-support tools. Concurrently, advances in artificial neural networks revitalized the connectionist paradigm, particularly with the introduction of the backpropagation algorithm (Rumelhart, Hinton, & Williams, 1986), which allowed the efficient training of multilayer networks. However, limitations in computational power and data availability restricted neural network performance, leading to a second period of stagnation – often called the second AI winter in the late 1980s and early 1990s.

The 1990s and early 2000s marked the emergence of statistical machine learning as a dominant paradigm in the field. With the availability of larger datasets and advances in probabilistic modelling, researchers increasingly moved away from rigid symbolic systems toward algorithms capable of learning patterns from data. Hidden Markov Models (Rabiner, 1989), support vector machines (Cortes & Vapnik, 1995), and Bayesian networks (Pearl, 1988) became central tools for applications such as speech recognition, natural language processing, and computer vision. A

symbolic milestone of this era was IBM's Deep Blue, which defeated world chess champion Garry Kasparov in 1997 (Campbell, Hoane, & Hsu, 2002). This event showcased the capacity of AI systems to surpass human expertise in well-defined domains, although critics argued that brute computational power, rather than actual intelligence, underpinned the achievement.

The 2010s ushered in the deep learning revolution, enabled by exponential growth in computational power, the availability of massive datasets, and innovations in neural architectures. Breakthroughs included convolutional neural networks (CNNs) for image recognition, popularised by Krizhevsky, Sutskever, and Hinton's (2012) ImageNet success, and recurrent neural networks (RNNs) for sequence modelling. Perhaps the most transformative development was the introduction of the transformer architecture (Vaswani et al., 2017), which revolutionised natural language processing by enabling scalable attention mechanisms. This innovation underpinned subsequent large language models such as BERT (Devlin et al., 2019) and GPT-3 (Brown et al., 2020), which achieved unprecedented performance across diverse linguistic tasks. Parallel advances in reinforcement learning produced landmark achievements, such as DeepMind's AlphaGo, which defeated the world champion Go player Lee Sedol in 2016 (Silver et al., 2016). Unlike earlier systems, AlphaGo integrated deep learning with reinforcement learning, exemplifying the hybridisation of paradigms and demonstrating AI's potential to tackle problems once deemed intractable for machines.

Today, AI has become a widespread social-technical phenomenon. Its applications extend beyond specialized fields to influence healthcare, finance, education, transportation, mental health, and more (Topol, 2019). The rapid increase in AI-related publications indicates both technological progress and the global focus on AI research (Zhang et al., 2021). Simultaneously, the field is becoming more connected to ethical, legal, and social debates about transparency, bias, accountability, and the future of human-machine interactions (Floridi & Cowls, 2019). The historical development of AI reveals the interplay between technical innovations, conceptual shifts, and societal expectations. Each stage has helped shape a complex discipline, characterized by cycles of enthusiasm and scepticism, yet steadily progressing toward more advanced models of intelligence. The journey from symbolic reasoning to deep learning highlights not only technological development but also the growing possibilities of simulating intelligent behaviour.

Major Paradigms: Symbolic AI, Machine Learning, and Deep Learning

Successive paradigms have shaped the development of artificial intelligence, each reflecting distinctive epistemological assumptions, computational strategies, and technological affordances. Three paradigms have been particularly influential in defining the field: symbolic AI, machine

learning, and deep learning. These approaches, although often presented as sequential, coexist and frequently interact in contemporary research, underscoring the pluralistic nature of AI.

Symbolic AI is the earliest and most dominant paradigm in AI research. Emerging in the 1950s and 1960s, it conceptualises intelligence as the manipulation of symbols according to formal logical rules (Newell & Simon, 1976). The central idea is that cognition can be represented in terms of explicit structures – facts, rules, and heuristics – that machines can store and process. A key feature of symbolic AI is knowledge representation, which involves formalizing real-world entities and their relationships using languages such as predicate logic, semantic networks, and frames (Brachman & Levesque, 2004). This method supports early expert systems, such as DENDRAL for chemical analysis (Buchanan & Feigenbaum, 1978) and MYCIN for medical diagnosis (Shortliffe, 1976). These systems showed that rule-based reasoning could match human expertise in specific areas, inspiring strong optimism about AI’s potential.

However, symbolic AI faced significant limitations. Manually encoding large amounts of knowledge was unsustainable and is often referred to as the “knowledge acquisition bottleneck” (Lenat & Feigenbaum, 1991). Additionally, symbolic systems had difficulty with uncertainty, ambiguity, and the contextual flexibility that characterizes natural human cognition. While effective in structured settings, symbolic AI was less capable in dynamic or noisy environments, such as natural language understanding and real-world perception. Despite these issues, symbolic methods remain relevant, particularly in areas that require explainability, formal reasoning, and logical constraints. Modern AI research increasingly investigates hybrid models that combine symbolic reasoning with statistical learning, aiming to connect structured knowledge with data-driven adaptability (Marcus, 2020).

Machine learning (ML) marks a fundamental shift from rule-based programming to data-driven inference. Instead of explicitly encoding knowledge, ML systems learn patterns and decision rules from empirical data. Mitchell (1997) defined machine learning as the process by which “a computer program is said to learn from experience E with respect to some class of tasks T and performance measure P , if its performance at tasks in T , as measured by P , improves with experience E .” Machine learning encompasses a variety of methods, including: a) supervised learning, where models are trained on labelled datasets to predict outcomes (e.g., linear regression, support vector machines, random forests); b) unsupervised learning, which identifies hidden structures within unlabelled data (e.g., clustering, principal component analysis); c) reinforcement learning, where agents learn through trial and error by interacting with an environment and receiving feedback in the form of rewards or penalties (Sutton & Barto, 2018).

A notable example of ML in practice is the use of support vector machines for text categorisation (Joachims, 1998), which established the utility of statistical methods in natural language processing. Similarly, Bayesian networks (Pearl, 1988) introduced probabilistic reasoning into AI, allowing systems to handle uncertainty in a principled way. The success of ML lies in its flexibility and scalability. By leveraging increasing computational power and data availability, ML algorithms became applicable to diverse fields, including bioinformatics, finance, image recognition, and speech processing. Yet, ML systems also face limitations, particularly in terms of interpretability, susceptibility to overfitting, and dependence on large, high-quality datasets.

Deep learning constitutes a specialised subfield of machine learning that leverages artificial neural networks with multiple hidden layers to capture complex, hierarchical patterns in data. Inspired by the structure of the human brain, deep neural networks (DNNs) perform representation learning, automatically extracting features from raw input without extensive manual engineering (LeCun, Bengio, & Hinton, 2015). Convolutional neural networks (CNNs) revolutionised image classification, achieving state-of-the-art performance on large-scale benchmarks such as ImageNet (Krizhevsky, Sutskever, & Hinton, 2012). Recurrent neural networks (RNNs), particularly long short-term memory (LSTM) models (Hochreiter & Schmidhuber, 1997), have enabled the effective modelling of sequential data, including speech and text. More recently, the introduction of the transformer architecture (Vaswani et al., 2017) has transformed natural language processing by employing self-attention mechanisms that scale efficiently to massive corpora. These innovations culminated in the development of large-scale pretrained models, such as BERT (Devlin et al., 2019) and GPT-3 (Brown et al., 2020), which demonstrate remarkable generalisation across a wide range of linguistic tasks. It has also achieved breakthroughs in reinforcement learning, exemplified by AlphaGo's victory over world champion Go players (Silver et al., 2016).

Although deep learning has had a transformative impact, it also has limitations. These systems are often referred to as "black boxes," which raises concerns about transparency, interpretability, and ethical accountability (Doshi-Velez & Kim, 2017). Additionally, the environmental and social costs of training large-scale models, including energy use and the reinforcement of data biases, have become significant concerns (Bender et al., 2021). Nonetheless, deep learning continues to fuel the rapid growth of AI capabilities. Its strength comes from its ability to scale with data and computational power, expanding what machines can do in perception, reasoning, and interaction.

Applications of AI Across Domains

Artificial intelligence has evolved from a specialised area of computer science into a pervasive socio-technical force with applications across a wide array of domains. These applications illustrate

both the versatility of AI methodologies and their capacity to transform knowledge practices, professional routines, and societal infrastructures. The following overview surveys major areas of application, highlighting representative achievements and critical debates (see Figure 2).

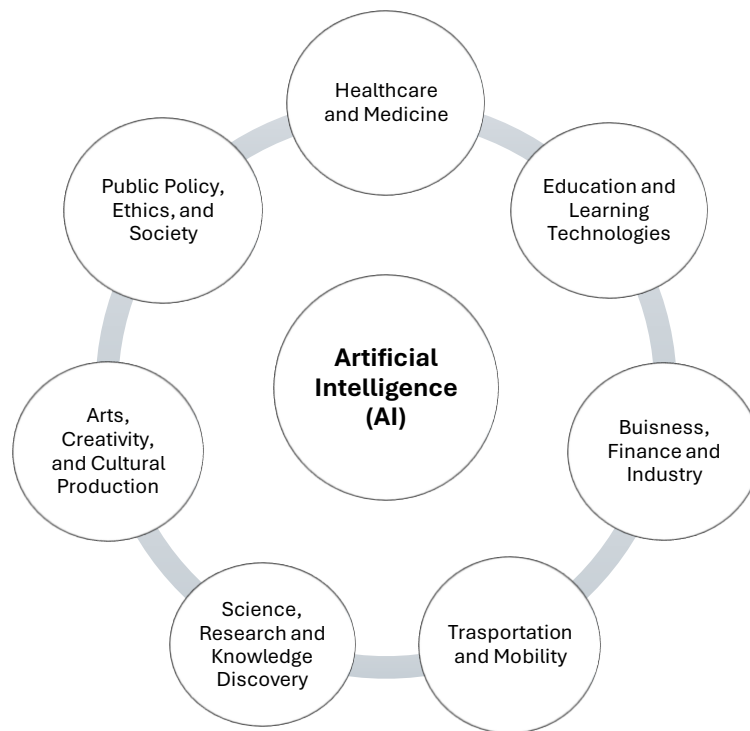


Figure 2. The significant applications of AI across domains. [Image created by me].

Healthcare has become one of the most prominent areas for AI deployment. Machine learning algorithms are widely used in medical imaging, where convolutional neural networks have achieved diagnostic accuracy comparable to that of human experts in detecting tumours, fractures, and retinal diseases (Esteva et al., 2017; Gulshan et al., 2016). In oncology, AI systems aid early cancer detection by analyzing patterns in histopathology slides and radiographic scans, thereby enhancing precision medicine. Beyond imaging, AI is utilized in clinical decision support systems (CDSS), which integrate patient data with predictive models to support physicians in diagnosis and treatment planning (Shortliffe & Sepúlveda, 2018). In psychiatry and psychology, NLP techniques have been used to identify markers of depression, suicidality, and psychosis from speech and text data (Calvo et al., 2017; Salemi, Foroozandeh & Ashkzari, 2025). These systems demonstrate the potential of AI for mental health assessment and intervention, although ethical debates about privacy, consent, and clinical responsibility remain essential. The COVID-19 pandemic accelerated interest in AI-based epidemiological modelling, drug discovery, and telemedicine platforms. However, deploying AI in

healthcare also reveals several limitations, including algorithmic bias, data quality issues, and challenges in integrating AI into clinical workflows (Topol, 2019).

AI has become a catalyst for innovation in education through intelligent tutoring systems, adaptive learning platforms, and automated assessment tools. Intelligent tutoring systems personalize learning experiences by adjusting feedback and instructional strategies to individual students, often using reinforcement learning to optimize teaching paths (Woolf, 2008). Automated essay scoring and natural language understanding support large-scale assessment, while recommender systems offer customized educational resources. More recently, large language models such as GPT-3 have been integrated into educational platforms for language practice, summarization, and dialogue-based learning. While these applications improve accessibility and scalability, critics highlight risks of over-reliance, superficial knowledge, and the loss of pedagogical control (Luckin et al., 2016).

In the corporate sector, AI enables predictive analytics, fraud detection, and process automation. Financial institutions use ML algorithms to analyse transaction patterns, spot anomalies, and manage risk portfolios (Bose & Mahapatra, 2001). Algorithmic trading systems rely on predictive models to optimize investment strategies with millisecond-level accuracy. In the industry, robotics and computer vision support automation in manufacturing and logistics, reducing operational costs and enhancing productivity. Natural language technologies, such as chatbots and virtual assistants, enhance customer service, while recommender systems on e-commerce platforms like Amazon and Netflix predict consumer preferences (Ricci, Rokach, & Shapira, 2015). However, the wide adoption of AI in industry also sparks debates about job displacement, algorithmic transparency, and the concentration of technological power in large corporations.

AI has transformed transportation through the development of autonomous vehicles. Self-driving cars combine computer vision, sensor fusion, and reinforcement learning to navigate complex environments (Katrakazas et al., 2015). Companies like Waymo and Tesla have invested heavily in autonomous driving, making AI a key element in the future of mobility. Beyond self-driving cars, AI also enhances traffic optimization, fleet management, and predictive maintenance in aviation and logistics. These applications aim to improve efficiency and safety, although challenges remain, such as ethical decision-making in accident scenarios, liability issues, and infrastructure adaptation (Bonnefon, Shariff, & Rahwan, 2016).

AI techniques are increasingly integrated into scientific research to speed up discovery. In genomics, ML algorithms analyse DNA sequences to find disease-related variants, supporting the growth of personalized medicine (Libbrecht & Noble, 2015). In physics and chemistry, AI has contributed to designing new materials and the prediction of molecular structures, as demonstrated

by DeepMind's AlphaFold, which achieved remarkable accuracy in protein folding prediction (Jumper et al., 2021). In social sciences, NLP techniques enable large-scale text analysis, allowing researchers to study political discourse, cultural trends, and psychological phenomena across extensive collections of text. These methods expand the scope of empirical research, but also raise concerns about validity, interpretability, and reproducibility (Grimmer, Roberts, & Stewart, 2022).

AI is not limited to scientific and industrial domains, but it is increasingly permeating artistic and cultural practices. Generative algorithms produce visual art, music, and literature, often through deep generative models such as Generative Adversarial Networks (GANs) and transformers (Goodfellow et al., 2014; Radford et al., 2019). AI-generated artworks have been exhibited in galleries and even auctioned at major houses such as Christie's, provoking debates about authorship, authenticity, and the nature of creativity. In the film and media industries, AI tools support scriptwriting, editing, and audience analysis. While these technologies enable new forms of cultural expression, they also disrupt traditional notions of artistic labour and intellectual property.

Finally, AI plays a central role in shaping governance and public policy. Predictive policing systems employ ML algorithms to forecast crime hotspots, though these applications have been criticised for reproducing systemic biases (Lum & Isaac, 2016). Governments deploy AI for administrative automation, welfare allocation, and surveillance, raising profound ethical concerns about civil liberties, transparency, and democratic accountability (Eubanks, 2018). The societal impact of AI is thus ambivalent: it offers unprecedented opportunities for efficiency and innovation, yet it also poses risks of inequality, disempowerment, and exploitation. These tensions underscore the need for responsible AI frameworks that embed ethical reflection, human oversight, and regulatory safeguards into technological development (Floridi & Cowls, 2019).

Exponential Growth of AI Research and Publications

The trajectory of artificial intelligence has been accompanied by exponential growth in scientific publications, reflecting both the field's maturation and its increasing societal relevance (see Figure 3). Bibliometric analyses demonstrate that AI is among the fastest-expanding areas of research, with publication rates outpacing those of many traditional disciplines (Zhang et al., 2021).

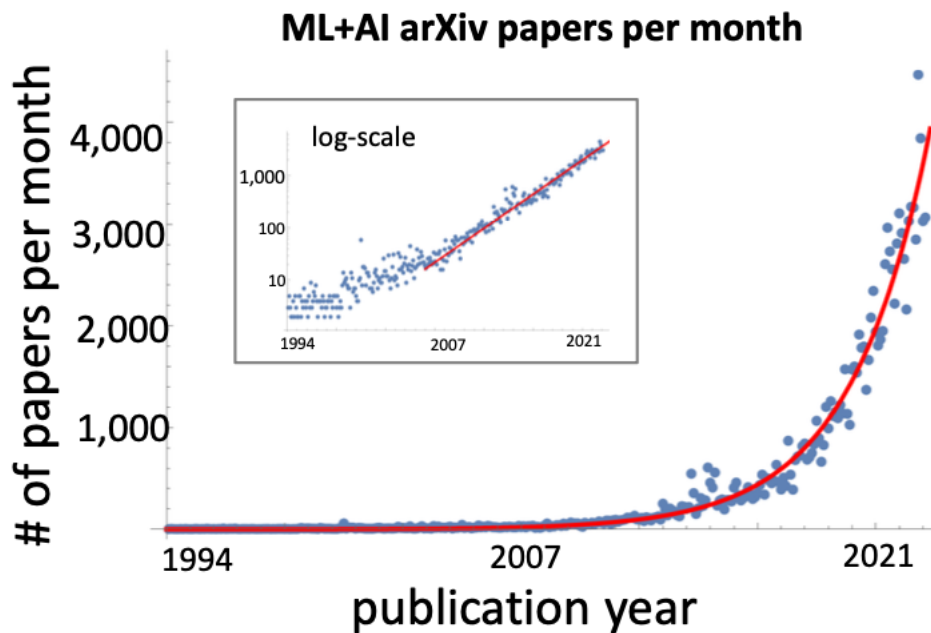


Figure 3. Exponential growth of AI-related publications on arXiv. The blue dots represent the monthly number of papers submitted in categories such as cs.AI, cs. LG, cs.NE, and stat.ML, while the red curve shows the fitted exponential trend. The inset presents the same data on a logarithmic scale, confirming the near-linear trajectory that indicates exponential growth. [Image's source: Krenn et al. (2022)].

The figure illustrates the exponential increase in the number of papers on artificial intelligence and machine learning published on arXiv between 1994 and 2021. According to Krenn et al. (2022), the doubling rate of AI-related publications is approximately every 23 months, reflecting both the vitality of the field and the challenges posed by information overload and concentration of research agendas. Even if this graph shows only publications on arXiv (and is not an example of our field that must be represented on a platform such as PubMed), it reflects an interesting trend in recent years.

In its formative decades, from the 1950s to the 1980s, AI remained a relatively specialized domain, characterized by symbolic reasoning, expert systems, and computational linguistics. Publications were concentrated in niche journals such as *Artificial Intelligence* and conference proceedings from the International Joint Conference on Artificial Intelligence (IJCAI). Growth was modest but steady, with developments often constrained by limited computational resources and data availability (Nilsson, 2010).

The 1990s experienced a significant increase in output, fuelled by the rise of machine learning and probabilistic models. Conferences such as Neural Information Processing Systems (NeurIPS) and the International Conference on Machine Learning (ICML) became important platforms for sharing

research. Greater access to large datasets and advances in computational infrastructure supported the growth of studies on support vector machines, ensemble methods, and probabilistic graphical models (Jordan & Mitchell, 2015). Bibliometric data indicate that between 1990 and 2000, the number of AI-related publications more than doubled, marking a shift from symbolic reasoning to data-driven approaches as the field's primary focus.

The explosive growth of deep learning research has defined the period from 2010 onwards. Landmark breakthroughs, such as the success of convolutional neural networks in ImageNet classification (Krizhevsky, Sutskever, & Hinton, 2012) and the introduction of the Transformer architecture (Vaswani et al., 2017), triggered an unprecedented surge of academic and industrial activity. Between 2010 and 2020, the number of AI publications increased by a factor of six, with deep learning dominating the output in natural language processing, computer vision, and reinforcement learning (Shoham et al., 2018). This exponential growth was mirrored in citation rates, as papers introducing architectures such as BERT (Devlin et al., 2019) and GPT-3 (Brown et al., 2020) rapidly became among the most cited in computer science.

AI research has increasingly penetrated diverse disciplines, including medicine, law, economics, and the humanities. Natural language processing methods are now used in political science to analyse parliamentary debates, in psychology to model therapeutic dialogues, and in literature studies to trace semantic drift across historical corpora (Grimmer, Roberts, & Stewart, 2022). The proliferation of AI across disciplines has further accelerated publication growth by attracting scholars from outside computer science. Interdisciplinary journals such as *Nature Machine Intelligence* and *AI & Society* exemplify the expanding intellectual landscape. Moreover, preprint servers such as arXiv have permitted rapid and open access to research outputs.

The exponential growth of AI publications is not limited to academia; it also extends to industry and policy reports. Companies such as Google, Microsoft, and OpenAI regularly release technical white papers that shape research trajectories and public discourse (Amodei et al., 2016). Similarly, policy organisations, including the OECD and the European Commission, produce reports that frame the ethical, legal, and societal implications of AI (OECD, 2019). These forms of grey literature make a substantial contribution to the overall body of knowledge about AI, influencing regulatory debates and shaping public perceptions.

While exponential growth reflects vitality, it also introduces challenges. The sheer volume of publications raises concerns about information overload, replication crises, and uneven research quality. The dominance of a few highly cited models risks homogenising research agendas, while barriers to entry, as the high computational costs of training large-scale models, concentrate influence

within elite institutions and corporations (Bender et al., 2021). Moreover, the rapid pace of dissemination complicates peer review and amplifies risks of premature claims, necessitating new mechanisms for evaluation and knowledge curation.

Different tools in the Artificial Intelligence field

As summarized in Table 2, there are various AI tools and core concepts.

Symbolic AI: Built on explicit rules, logical inference, and symbolic manipulation (Newell & Simon, 1976). Knowledge frameworks – semantic networks, frames, and ontologies – were created to formalize reasoning. Expert systems, such as MYCIN (Shortliffe, 1976), demonstrated domain-specific success but faced limitations due to ambiguity and scalability issues. Neuro-symbolic methods now aim to combine symbolic reasoning with machine learning.

Machine Learning: Marked a paradigm shift during the 1990s–2000s towards data-driven inference. It includes supervised, unsupervised, and reinforcement learning, all grounded in statistical learning theory (Vapnik, 1998). Standard algorithms include SVMs, decision trees, and ensemble methods, which incorporate probabilistic reasoning to manage uncertainty in real-world data.

Neural Networks: Inspired by biological neural structures, ANNs have evolved from perceptrons to deep networks. CNNs revolutionized computer vision (LeCun et al., 2015), while RNNs and LSTMs advanced sequential modelling (Hochreiter & Schmidhuber, 1997). Transformers, on the other hand, became foundational for NLP (Vaswani et al., 2017). These architectures collectively powered the deep learning revolution.

Embeddings: Offer dense vector representations of words, sentences, and documents. Early models, such as Word2Vec (Mikolov et al., 2013) and GloVe (Pennington et al., 2014), captured distributional semantics. Contextual embeddings, such as BERT (Devlin et al., 2019), introduce sensitivity to polysemy and context, supporting a wide array of downstream NLP tasks.

Parsing and Semantics: Parsing breaks down sentences into syntactic structures. Initially grounded in rule-based grammars (Chomsky, 1956), it has since shifted toward probabilistic and neural dependency approaches (Kübler, McDonald, & Nivre, 2009). Semantic analysis extends to disambiguation, role labelling, and coreference resolution, while pragmatics and discourse-level modelling remain complex challenges.

Reinforcement and Generative Models: Reinforcement Learning enables agents to optimize behaviour through reward-based interaction (Li et al., 2016), exemplified by AlphaGo. Generative methods, like VAEs (Kingma & Welling, 2014) and GANs (Goodfellow et al., 2014), facilitate the

creation of new content, laying the groundwork for modern generative AI, including large language models.

Infrastructures: Open-source libraries (TensorFlow, PyTorch, Hugging Face Transformers) and scalable cloud services (Google Cloud, AWS) have democratized access to advanced techniques. Transfer learning with pre-trained models (Howard & Ruder, 2018) has significantly lowered the entry barrier for research and applications in fields such as psychology and healthcare.

Concept / Approach	Brief Description	Examples
Symbolic AI	Early rule-based reasoning systems relying on logic and structured knowledge representation.	MYCIN, semantic networks, ontologies.
Machine Learning	Statistical models that infer patterns from data, enabling prediction and generalisation.	Support Vector Machines, decision trees, ensembles.
Neural Networks	Multi-layered computational models inspired by the brain, supporting hierarchical feature learning.	CNNs (vision), RNNs/LSTM (sequences), Transformers (NLP).
Embeddings	Vector representations mapping linguistic units into continuous spaces to capture semantic similarity.	Word2Vec, GloVe, BERT embeddings.
Parsing & Semantics	Methods to analyse syntactic structures and semantic roles in natural language.	Context-free grammars, dependency parsing, semantic role labelling.
Reinforcement & Generative Models	Learning through interaction with environments and models for content generation.	AlphaGo, VAEs, GANs, LLMs

Table 2. The Core Concepts in AI and NLP. Overview of the main paradigms in artificial intelligence and natural language processing, with concise definitions and illustrative examples.

Natural Language Processing (NLP)

Natural Language Processing (NLP) is a subfield of artificial intelligence, as illustrated above, that studies computational techniques for analysing, understanding, and generating human language in both spoken and written form. Its scope extends beyond the mechanical manipulation of symbols: it encompasses the design of algorithms and models that, to varying degrees, approximate the

complex cognitive processes underlying linguistic competence. In this sense, NLP can be understood as a bridge between computer science, linguistics, and cognitive science (Jurafsky & Martin, 2023).

Historically, NLP was initially defined within the paradigm of computational linguistics, a discipline rooted in the formal modelling of grammatical structures. Early research in the 1950s and 1960s was characterised by rule-based approaches derived from linguistic theories, such as context-free grammars introduced by Chomsky (1956). These symbolic approaches attempted to capture the logical regularities of syntax and morphology through handcrafted rules and lexicons. Although limited in scalability, they provided the foundation for seminal applications such as early machine translation systems (Hutchins, 2006).

Over time, the conceptual scope of NLP expanded in response to both theoretical and technological developments. The introduction of probabilistic methods in the 1980s and 1990s (Manning & Schütze, 1999) marked a decisive shift towards data-driven approaches. These methods enabled systems to learn linguistic patterns from large corpora rather than relying exclusively on manually encoded rules. Such a transition reflected not only the maturation of computational statistics but also the growing recognition that human language is inherently probabilistic, marked by ambiguity, variation, and contextual dependence.

In its contemporary form, NLP is defined by the integration of machine learning and, more recently, deep learning architectures. Neural models, particularly the transformer architecture (Vaswani et al., 2017), have enabled unprecedented progress in tasks such as machine translation, question answering, and dialogue modelling. Pre-trained language models such as BERT (Devlin et al., 2019), GPT-3 (Brown et al., 2020), and more recent large language models demonstrate the scalability of this paradigm, where vast quantities of textual data and computational resources converge to produce general-purpose linguistic representations.

From a conceptual perspective, NLP can be distinguished into two complementary orientations (Su, Huang & Chen, 2019):

1. Natural language understanding (NLU), which aims to extract meaning from linguistic input through parsing, semantic analysis, and pragmatic interpretation.
2. Natural language generation (NLG), which focuses on producing fluent, contextually appropriate output, whether in the form of translations, summaries, or conversational responses.

This dual orientation reflects the bidirectional nature of human communication, highlighting the complexity of replicating linguistic competence computationally (see Figure 4).

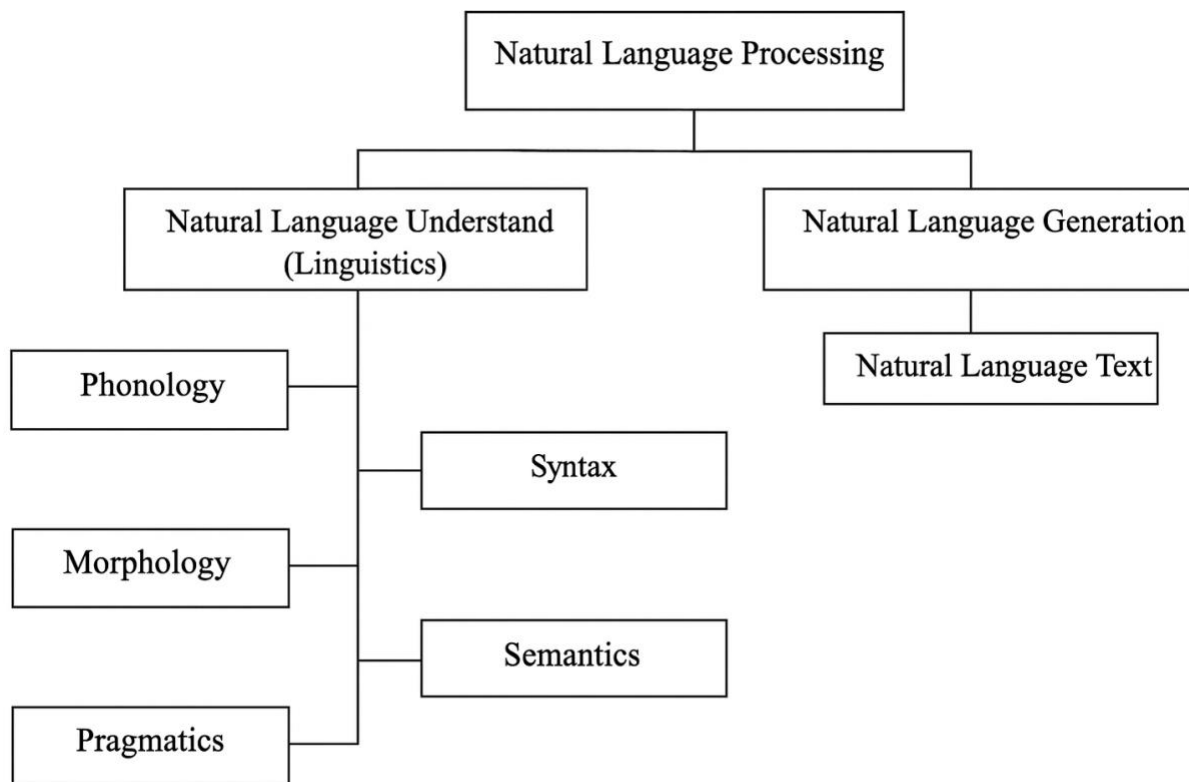


Figure 4. The structure of Natural Language Processing (NLP) is divided into two main areas: Natural Language Understanding (NLU), which encompasses phonology, morphology, syntax, semantics, and pragmatics, and Natural Language Generation (NLG), which focuses on producing coherent natural language text. [Image created by me].

A central premise of Natural Language Processing is that language can be analyzed at multiple, hierarchically organized levels, each corresponding to a dimension of linguistic structure and meaning (see Figure 3). These levels – morphological, syntactic, semantic, and pragmatic – provide a conceptual framework for both linguistic theory and computational implementation (Jurafsky & Martin, 2023). By decomposing language into distinct yet interrelated strata, NLP systems can target specific aspects of human communication while progressively approximating the complexity of natural discourse.

Morphology deals with the internal structure of words, including roots, affixes, and inflectional patterns. Computational morphology aims to identify and model these parts to perform tasks such as lemmatization, stemming, and morphological tagging. Early methods used finite-state transducers (Koskenniemi, 1983), which captured word formation through rule-based automata. These techniques remain influential, especially in highly inflectional languages such as Finnish and Turkish, where morphological variation plays a crucial role in conveying meaning. Modern approaches increasingly

rely on neural sequence-to-sequence models that learn morphological changes directly from annotated corpora. Cotterell et al. (2016) showed that recurrent neural networks can reach near-human accuracy on cross-linguistic morphological tasks. Morphological processing is crucial not only for accurate text analysis but also for downstream tasks, such as information retrieval and machine translation, where errors in morphological recognition can propagate through the system.

Syntax refers to the hierarchical arrangement of words into phrases and sentences, governed by rules of grammatical structure. Computational syntax has historically been dominated by two main paradigms: constituency parsing, based on phrase structure grammars, and dependency parsing, which models syntactic relations between words directly. Probabilistic context-free grammars (PCFGs) provided a statistical foundation for constituency parsing, enabling models to assign probabilities to alternative syntactic structures (Collins, 1999). Dependency parsing, by contrast, has gained prominence for its efficiency and cross-linguistic applicability (Kübler, McDonald, & Nivre, 2009). Neural network-based parsers, particularly those employing transformer architectures, now achieve state-of-the-art performance across benchmarks, demonstrating the efficacy of distributed representations for capturing syntactic patterns (Dozat & Manning, 2017). Syntactic analysis is not merely a structural exercise: it underpins tasks such as question answering, text summarisation, and dialogue systems by providing a scaffold for interpreting sentence meaning.

Semantics addresses the meaning of linguistic units and their relationships within larger contexts. Computational semantics has historically grappled with issues such as polysemy (a word with multiple meanings), synonymy, and entailment. Early approaches, such as semantic networks and frame semantics (Fillmore, 1976), attempted to encode meaning through symbolic representations. The emergence of distributional semantics marked a paradigm shift. Latent Semantic Analysis (LSA), introduced by Landauer and Dumais (1997), demonstrated that meaning could be approximated by patterns of co-occurrence in large corpora. This principle evolved into word embeddings (Mikolov et al., 2013), which represent words as dense vectors in high-dimensional space. More recently, contextual embeddings, introduced by models such as BERT (Devlin et al., 2019), have transformed semantic analysis by capturing word meaning dynamically, depending on the surrounding context. Semantic processing is foundational for applications such as sentiment analysis, information extraction, and knowledge graph construction, where accurate modelling of meaning is critical.

Pragmatics expands on semantic analysis by considering the meaning in context, such as the speaker's intent, discourse structure, and social norms. Unlike morphology, syntax, and semantics, which can often be formalized with some precision, pragmatics presents a greater challenge because it depends on world knowledge and situational factors. Computational pragmatics deals with phenomena like

anaphora resolution (linking pronouns to their antecedents), discourse coherence, and speech act recognition. Early discourse analysis was influenced by theories such as Rhetorical Structure Theory (Mann & Thompson, 1988), which modelled relationships between parts of a text. More recent methods include deep learning to enhance the understanding of coherence and dialogue flow. In clinical and psychological settings, pragmatic analysis is relevant. For instance, tracking shifts in coherence and references within psychotherapy transcripts can offer insights into cognitive organization and therapeutic progress (Angus, Levitt, & Hardtke, 1999).

Although presented as distinct levels, morphological, syntactic, semantic, and pragmatic analyses are deeply interconnected. Errors or simplifications at one level often cascade to others, underscoring the importance of integrative modelling. Current NLP research increasingly embraces multitask and end-to-end approaches, in which a single model jointly learns multiple levels of linguistic structure, reflecting the holistic nature of language itself.

Core Techniques and Models

Successive methodological paradigms have characterised the development of Natural Language Processing, each reflecting advances in computational theory, linguistic science, and statistical modelling. These paradigms – rule-based systems, statistical approaches, machine learning methods, and deep learning architectures – represent not only technological milestones but also shifts in how language itself is conceptualised in computational terms.

Rule-Based and Symbolic Approaches. The earliest systems in NLP were founded on symbolic reasoning and rule-based methods. Rooted in formal linguistics and logic, these approaches sought to capture grammatical and semantic regularities through handcrafted rules, lexicons, and production systems (Winograd, 1972). For example, early machine translation projects such as the Georgetown-IBM experiment in 1954 relied exclusively on manually encoded dictionaries and syntactic rules (Hutchins, 2006). While rule-based methods offered transparency and interpretability, they were also brittle and labour-intensive. The vast variability of natural language made exhaustive rule specification infeasible. Moreover, such systems lacked robustness in handling ambiguity, polysemy, and noisy input and features inherent to real-world communication. Nevertheless, symbolic NLP remains relevant in specific domains. Contemporary hybrid systems often incorporate symbolic reasoning for tasks requiring logical inference, such as legal reasoning or medical knowledge representation (Cambria & White, 2014).

Statistical Methods. The introduction of probabilistic models in the late 1980s and 1990s marked a decisive turning point. Statistical NLP sought to capture linguistic patterns from large

corpora rather than relying exclusively on handcrafted rules (Manning & Schütze, 1999). Key techniques included: a) n-gram models, which estimate the probability of a word given its preceding context, forming the foundation of early speech recognition and machine translation systems (Jelinek, 1990); b) Hidden Markov Models (HMMs), widely used for part-of-speech tagging and speech recognition due to their ability to model sequential dependencies (Rabiner, 1989); Probabilistic Context-Free Grammars (PCFGs), which introduced probabilities into syntactic parsing, enabling models to rank alternative parse trees (Collins, 1999). Statistical methods improved robustness and scalability, allowing systems to adapt to the variability of natural language through corpus-based learning. However, their reliance on local context and independence assumptions limited their ability to capture deeper semantic and pragmatic phenomena.

Machine Learning and Feature-Based Models. Building on statistical foundations, the 2000s witnessed the growing influence of supervised machine learning in NLP. Algorithms such as Support Vector Machines (SVMs), Maximum Entropy classifiers, and Conditional Random Fields (CRFs) became standard tools for tasks including named entity recognition, sentiment analysis, and information extraction (Lafferty, McCallum, & Pereira, 2001). These models required careful feature engineering: linguists and data scientists manually designed input features such as word shape, orthographic cues, or dependency relations. While this approach yielded substantial improvements, its dependence on handcrafted features constrained adaptability and generalisation.

Deep Learning and Neural Architectures. The advent of deep learning radically transformed NLP. Neural networks, particularly recurrent architectures such as LSTMs (Hochreiter & Schmidhuber, 1997), enabled the modelling of long-range dependencies in sequential data. These breakthroughs were accelerated by the availability of large datasets and increased computational power. The paradigm shift was consolidated with the introduction of word embeddings, such as Word2Vec (Mikolov et al., 2013), which provided dense, continuous vector representations of words. This approach operationalised the distributional hypothesis: “you shall know a word by the company it keeps” (Firth, 1957) and facilitated transfer learning across tasks.

The Transformer Revolution. A defining moment in NLP occurred with the introduction of the transformer architecture (Vaswani et al., 2017). Unlike recurrent models, transformers utilize self-attention mechanisms that simultaneously compute contextual relationships between all words in a sequence. This innovation significantly improved efficiency, enabling models to capture hierarchical and global dependencies in language. Transformers paved the way for pre-trained language models, which are trained on massive corpora and subsequently fine-tuned for specific tasks. Landmark examples include: a) BERT (Bidirectional Encoder Representations from Transformers), which

introduced contextual embeddings capturing bidirectional dependencies (Devlin et al., 2019); GPT (Generative Pre-trained Transformer) series, which demonstrated the capacity of autoregressive models to generate coherent and contextually rich text (Brown et al., 2020); T5 (Text-to-Text Transfer Transformer), which unified multiple NLP tasks within a single text-to-text framework (Raffel et al., 2020). These models achieve state-of-the-art results across benchmarks, often surpassing human performance in constrained evaluation settings.

Hybrid and Emerging Approaches. Current research increasingly explores hybrid systems that combine symbolic reasoning with deep learning. Such approaches aim to integrate the interpretability and logical structure of symbolic methods with the flexibility and scalability of neural architectures (Marcus, 2020). In parallel, multilingual and multimodal models expand the boundaries of NLP, linking text with speech, vision, and structured knowledge bases. From handcrafted rules to billion-parameter neural networks, the evolution of NLP techniques illustrates a dynamic interplay between linguistic theory, statistical modelling, and computational innovation. Each paradigm has addressed limitations of its predecessors while introducing new challenges, reflecting the complexity of modelling human language computationally.

Application Domains of NLP

The advancement of Natural Language Processing has enabled its application across a wide range of domains, including information retrieval, healthcare, and psychotherapy (see Figure 5). These domains illustrate the dual character of NLP as both a technical discipline and a socially embedded practice, with implications that extend far beyond computational efficiency.

One of the earliest applications of NLP lies in information retrieval (IR). Systems like web search engines rely on tokenisation, indexing, and semantic ranking to retrieve relevant documents from vast corpora. The incorporation of latent semantic indexing (Deerwester et al., 1990) and transformer-based models has significantly improved retrieval performance by capturing semantic relationships.

Knowledge management systems extend these principles to organisational contexts, enabling the automatic extraction and organisation of information from unstructured data such as reports, emails, or technical documentation (Manning, Raghavan, & Schütze, 2008). In healthcare, NLP has emerged as a transformative tool for both research and clinical practice. Applications include a) Electronic Health Record (EHR) analysis: NLP enables the extraction of clinically relevant information from unstructured notes, such as comorbidities, adverse events, or treatment outcomes (Meystre et al., 2008); b) clinical decision support: systems integrate NLP-derived insights into diagnostic and treatment recommendations (Demner-Fushman et al., 2009); c) biomedical literature mining: large-

scale text mining supports drug discovery, pharmacovigilance, and systematic reviews by identifying patterns across thousands of publications (Cohen & Hersh, 2005). The intersection of NLP and mental health research is particularly notable. Studies applying sentiment analysis and topic modelling to psychotherapy transcripts have provided new ways to track therapeutic alliance, emotional dynamics, and indicators of clinical progress (Imel et al., 2015; Atkins et al., 2014).

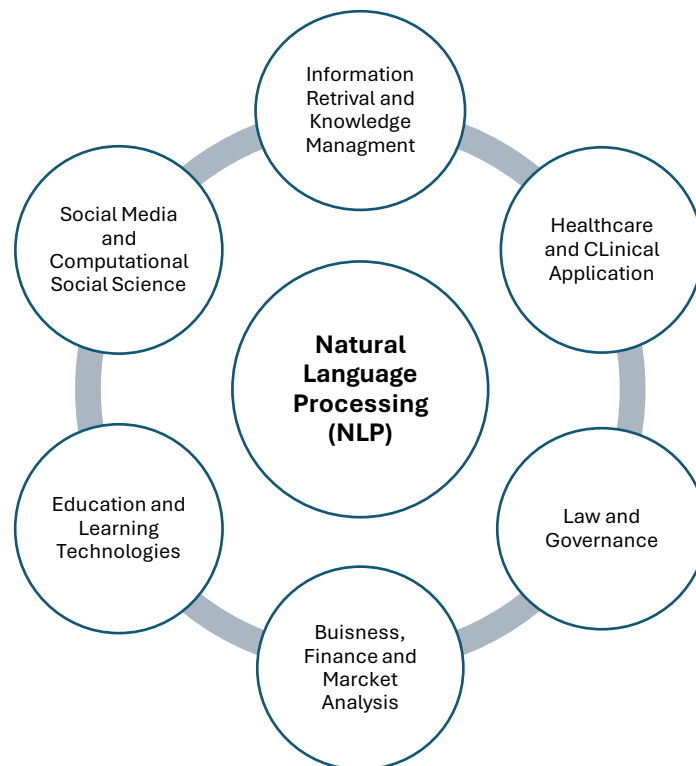


Figure 5. The significant applications of NLP across domains. [Image created by me].

The legal and governmental domains present distinct challenges due to the complexity and formality of legal language. NLP techniques are increasingly employed to: a) automate contract analysis and compliance checking; b) assist in e-discovery, where relevant documents are identified within large datasets during litigation; c) support policy analysis by identifying themes, sentiments, and discursive frames in parliamentary debates or regulatory texts (Ashley, 2017). Here, interpretability and transparency remain critical. Hybrid approaches that combine symbolic reasoning with machine learning are often preferred to preserve accountability.

In business contexts, NLP underpins applications such as customer sentiment analysis, chatbots for customer service, and automated financial reporting. Financial institutions employ NLP to analyse news articles, earnings calls, and social media to predict market movements or assess reputational

risks (Du et al., 2024). The rapid expansion of e-commerce has amplified the relevance of NLP for product recommendation systems, opinion mining, and fraud detection.

Educational technology has increasingly embraced NLP, particularly in developing automated essay scoring systems, intelligent tutoring systems, and adaptive learning environments. Systems such as e-rater, used in large-scale standardised testing, demonstrate the feasibility of automated grading based on syntactic complexity, lexical diversity, and discourse coherence (Attali & Burstein, 2006). NLP supports personalised learning through dialogue-based tutors that adapt feedback to learners' needs, exemplifying the intersection of computational linguistics and pedagogical psychology.

The rise of digital communication has created an unprecedented corpus of naturalistic textual data. NLP methods enable researchers to analyse social media platforms to gain insights into public opinion, political polarization, and the spread of misinformation (Tumasjan et al., 2010). In computational social science, large-scale analyses of online discourse illuminate phenomena such as collective identity formation, protest mobilisation, and mental health expression in online communities (Guntuku et al., 2019).

Recent advances point toward cross-domain integration, where models pre-trained on massive general-purpose corpora are adapted to specialised domains with relatively little data. For instance, BioBERT (Lee et al., 2020) adapts the BERT architecture to biomedical text, while LegalBERT (Chalkidis et al., 2020) addresses legal corpora. This trend exemplifies the flexibility of transformer-based architectures and their potential for domain-specific adaptation. The breadth of NLP applications illustrates its dual status as a scientific discipline and a transformative technology. While technical advances enable increasingly sophisticated analyses, the implications extend to fundamental questions about human communication, decision-making, and social organisation.

Ethical and Societal Implications

The exponential growth of Natural Language Processing has generated unprecedented opportunities but also complex ethical and societal challenges. These challenges concern not only technical performance but also issues of fairness, accountability, and the societal consequences of deploying language technologies at scale.

Bias and Fairness. NLP systems are profoundly shaped by the data on which they are trained. Since large corpora often reflect social biases related to gender, race, class, or culture, models tend to reproduce and even amplify such biases (Bolukbasi et al., 2016). For instance, word embeddings have been shown to encode stereotypical associations, as linking women with domestic roles and men with professional domains. In contexts such as recruitment, credit scoring, or legal decision-making, these

biases can lead to discriminatory outcomes. Efforts to mitigate bias include debiasing algorithms, fairness-aware training objectives, and careful curation of training datasets. The tension between linguistic representativeness and ethical neutrality remains unresolved (Blodgett et al., 2020).

Privacy and Surveillance. The growing use of NLP in healthcare, education, and workplace monitoring raises urgent privacy concerns. Handling sensitive textual data, such as psychotherapy transcripts or electronic health records, must strictly comply with data protection laws, including the General Data Protection Regulation (GDPR) in Europe and the Health Insurance Portability and Accountability Act (HIPAA) in the United States. Simultaneously, employing NLP in surveillance systems—such as sentiment analysis on social media for political aims—highlights the risk of misuse. These practices blur the line between public communication and private expression, prompting questions about informed consent and the principle of proportionality.

Transparency and Explainability. The rise of deep learning has introduced a paradox: models achieve state-of-the-art performance yet often operate as “black boxes,” making their decision-making difficult to understand. This lack of transparency undermines trust in applications where accountability is crucial, such as healthcare or law. Research into explainable AI (XAI) aims to provide interpretable explanations for model predictions through methods such as attention visualisation, feature attribution, or surrogate models (Ribeiro, Singh, & Guestrin, 2016). While promising, these methods often provide only partial insight into the highly complex decision boundaries of large neural models.

Labour and Economic Disruption. The automation of language-based tasks has significant implications for the labour market. Routine tasks, such as transcription, translation, and basic customer service, are increasingly being automated, potentially displacing workers. At the same time, new opportunities emerge in AI system development, data annotation, and algorithmic auditing. These dual dynamics highlight the importance of reskilling and policy frameworks that anticipate economic transitions driven by NLP technologies (Brynjolfsson & McAfee, 2014).

Societal and Cultural Impact. Language is not a neutral medium but a constitutive element of culture and identity. The deployment of NLP systems at scale significantly impacts how people access information, form opinions, and engage in public discourse. Search engines and recommender systems, for example, shape epistemic environments by privileging specific sources of information over others (Pariser, 2011). The global spread of English-dominated corpora also risks reinforcing linguistic inequalities. Low-resource languages remain underrepresented in NLP research, perpetuating disparities in digital participation and cultural preservation (Joshi et al., 2020).

Ethical Governance. Addressing these challenges requires robust governance frameworks. Initiatives such as the EU Ethics Guidelines for Trustworthy AI (2019) or the OECD Principles on AI (2019) articulate principles of fairness, accountability, and human oversight. However, translating principles into practice remains difficult, given the global nature of data flows and the rapid pace of technological innovation. Proposals for algorithmic impact assessments and audit mechanisms aim to institutionalise ethical oversight, though such mechanisms are still in their infancy.

Current Trends and Future Directions

The field of Natural Language Processing has entered a transformative phase marked by the rise of large-scale pre-trained models, multimodality, and growing interest in aligning computational advances with ethical and societal needs. These trends shape not only the trajectory of research but also the practical applications that are becoming embedded in everyday life.

The most notable development in recent years has been the emergence of large language models (LLMs), including BERT (Devlin et al., 2019), GPT-3 (Brown et al., 2020), and their successors. These models are characterised by a) pre-training on massive corpora, enabling them to capture general linguistic patterns; b) fine-tuning or prompting, which adapts them to specific tasks with minimal additional data; c) transferability, allowing a single architecture to perform across multiple domains and tasks. This paradigm shift towards foundation models has been described as a new “operating system” for AI (Bommasani et al., 2021), consolidating diverse NLP tasks within a unified framework. Generative language models have opened new possibilities for text production, including dialogue systems, summarisation, and content generation. In creative domains, they have been used to co-author fiction, generate poetry, or support ideation in journalism and marketing. At the same time, generative capabilities raise concerns about disinformation, deepfakes, and the blurring of boundaries between human and machine authorship (Floridi & Chiriatti, 2020). Research into factual grounding and hallucination mitigation aims to address these challenges.

A growing trend is the integration of language with other modalities such as vision and speech. Multimodal models (e.g., CLIP by Radford et al., 2021) combine textual and visual inputs, enabling tasks such as image captioning, video retrieval, and cross-modal reasoning. In clinical contexts, multimodal analysis of speech, gesture, and language has potential for enriched assessments of mental health. Equally important is progress in cross-lingual NLP, which aims to reduce inequalities between high-resource and low-resource languages. Approaches such as multilingual BERT (Devlin et al., 2019) and transfer learning strategies show promise in extending NLP capabilities globally; however, challenges related to linguistic diversity and cultural bias persist (Joshi et al., 2020).

Future directions in NLP emphasise the need for interpretability and human-centred design. Beyond achieving performance benchmarks, researchers increasingly prioritise systems that provide transparent reasoning, adapt to user needs, and augment rather than replace human expertise. This point is especially important in psychotherapy research, where NLP tools should function as aids to clinical interpretation rather than as opaque, black-box decision-makers (Xiao, 2016; Malgaroli et al., 2023). NLP is progressively intersecting with disciplines such as psychology, sociology, philosophy, and law. Computational methods are used not only to solve engineering problems but also to investigate human behaviour, discourse, and culture. This convergence exemplifies a broader shift towards computational social science and the digital humanities, in which language technologies become instruments for empirical inquiry into society and subjectivity. Another critical trend is the shift towards open science, as reflected in initiatives such as Hugging Face's model hub and large-scale collaborative datasets like Common Crawl.

Despite remarkable advances, several challenges remain at the frontier of NLP research: a) energy efficiency: the environmental cost of training large models is considerable, necessitating greener approaches to computation (Strubell, Ganesh, & McCallum, 2019); b) robustness: models are often brittle when confronted with adversarial examples or out-of-distribution data; c) ethical alignment: ensuring that NLP systems respect human rights, cultural diversity, and democratic values is a long-term imperative.

2.1. Text Mining and Topic Modelling

The terms "*text mining*" and "*text analysis*" are often used interchangeably, yet they denote related but distinct traditions within computational linguistics and the social sciences. Both approaches share the overarching aim of extracting meaningful information from unstructured textual data, but they diverge in methodological emphasis, epistemological orientation, and applications.

At its most general level, *text mining* refers to the application of computational techniques for extracting structured information from large bodies of unstructured text (Feldman & Sanger, 2007). Its orientation is exploratory and inductive, detecting latent structures, patterns, and relationships within textual corpora. Typical methods include clustering, classification, sentiment analysis, and topic modelling. Early developments in text mining during the 1980s and 1990s were driven by advances in information retrieval and statistical modelling, with milestones such as the vector space model (Salton, Wong, & Yang, 1975) enabling automated indexing and retrieval of digital texts.

By contrast, *text analysis* takes a broader, more interpretive approach. While it may integrate computational tools, it is also closely tied to traditions in linguistics, discourse analysis, hermeneutics, and narrative inquiry. Its primary concern lies in contextualisation, meaning, and interpretive nuance (Gee, 2014; Krippendorff, 2018). As such, text analysis encompasses qualitative approaches such as conversation analysis, which investigates the sequential organisation of talk (Sacks, Schegloff, & Jefferson, 1974), or narrative psychology, which explores how individuals construct meaning through storytelling (Bruner, 1990). In contrast to the algorithmic scalability of mining, text analysis privileges depth, subjectivity, and theory-driven interpretation.

This methodological distinction can be conceptualised along a continuum:

- Text mining enables quantification, scalability, and automation by relying on algorithms to process large datasets with minimal human intervention.
- Text analysis privileges context, meaning, and interpretive depth, requiring human expertise to situate textual patterns within the frameworks of theoretical and empirical traditions.

For example, a text mining study of psychotherapy transcripts may identify clusters of frequently co-occurring words, suggesting latent themes. In contrast, a text analysis study may interpret those themes through hermeneutic or clinical frameworks, such as therapeutic alliance, resistance, or transference. Similarly, in biomedical informatics, text mining is used to process millions of PubMed abstracts, detecting novel associations between genes and diseases (Cohen & Hersh, 2005), while text analysis situates these findings within established clinical knowledge frameworks. In literary studies, text mining enables macro-analyses of genres and stylistic patterns, while text analysis grounds interpretations in historical and cultural contexts (Houston, 2023).

The historical development of these fields reflects their disciplinary origins. Text mining emerged from the fields of information retrieval, computational linguistics, and machine learning in the late 20th century. In contrast, text analysis has deeper roots in the humanities and social sciences, with a strong emphasis on interpretation, context, and meaning-making. The convergence of these traditions became more visible in the 2000s, as interdisciplinary scholars sought to combine computational scalability with interpretative depth. Hybrid methodologies have since expanded across domains: in computational social science, algorithms generate candidate themes that are then qualitatively interpreted (Grimmer & Stewart, 2013); in psychotherapy research, automated topic modelling can uncover recurrent themes in transcripts, which are then validated through process coding and clinical expertise (Atkins et al., 2014; Imel et al., 2015; Cioffi et al., 2022).

This integrative movement echoes Ricoeur's (1976) hermeneutic arc, which emphasises the dialectic between explanation and understanding. Text mining provides explanatory models of linguistic patterns, whereas text analysis reintroduces human interpretation to reconstruct meaning. Across diverse fields, their complementarity has become increasingly apparent: mining offers scalable, data-driven discovery, while analysis provides interpretive frameworks that situate findings within cultural, psychological, or social contexts.

The distinction between text mining and text analysis thus highlights a central epistemological challenge in computational research: the balance between scalability and interpretability. Overreliance on automated mining risks producing “patterns without meaning,” whereas exclusive reliance on interpretive analysis risks idiosyncrasy and lack of generalisability. The future of text-based inquiry, therefore, lies in integrative approaches that harness the strengths of both traditions while remaining reflexive about their limitations.

Core Techniques in Text Mining

Information Retrieval and Pre-processing

A foundational stage in text mining is the retrieval and preparation of textual data, a process that directly conditions the quality of subsequent analyses. Information Retrieval (IR) constitutes the computational discipline dedicated to locating and ranking relevant documents from large collections, a field whose origins date back to the pioneering work of Salton and colleagues in the 1960s and 1970s (Salton, Wong, & Yang, 1975). At its core, IR systems rely on constructing document representations that enable efficient matching with user queries, typically implemented using models such as the Vector Space Model (VSM), probabilistic retrieval models, and, more recently, neural embeddings (Manning, Raghavan, & Schütze, 2008).

In text mining pipelines, IR provides the input corpus, while pre-processing ensures that raw textual data are transformed into a structured form suitable for algorithmic manipulation. This stage often requires multiple layers of linguistic normalisation, including:

- Tokenization or dividing text into minor linguistic units, like words or subwords. Early methods relied on whitespace rules, while modern NLP uses subword tokenizers such as Byte-Pair Encoding (Sennrich, Haddow, & Birch, 2016) and WordPiece (Wu et al., 2016).
- Stop-word removal, whereby highly frequent but semantically light words (e.g., articles, prepositions) are filtered out, based on the assumption that they contribute little to discriminative power.

- Stemming and lemmatisation, techniques designed to reduce word forms to their base or canonical representations. While stemming relies on heuristic truncation (e.g., “running” → “run”), lemmatisation integrates morphological and syntactic information to achieve greater linguistic accuracy (Jurafsky & Martin, 2023).
- Part-Of-Speech (POS) tagging, which labels tokens with grammatical categories, thus enabling more refined analyses of syntactic or semantic roles.
- Named Entity Recognition (NER), used to identify and classify proper names, organisations, and other domain-specific entities.

The importance of pre-processing lies not merely in computational efficiency but in its epistemological implications. Each transformation reflects an interpretative decision about what counts as “meaningful” text. For instance, removing stop-words assumes that function words are irrelevant, yet in discourse analysis or psychotherapy transcripts, these words may carry pragmatic significance (e.g., fillers, hesitations). Hence, methodological reflexivity is required to align pre-processing strategies with research objectives.

Contemporary pipelines increasingly use embedding-based representations during preprocessing. Models such as Word2Vec (Mikolov et al., 2013), GloVe (Pennington, Socher, & Manning, 2014), and later contextual embeddings like BERT (Devlin et al., 2019), map tokens into high-dimensional vector spaces that capture semantic and syntactic relationships. Unlike traditional bag-of-words representations, embeddings preserve word order and contextual nuance, representing a profound shift in the architecture of text mining systems.

Finally, pre-processing is intimately linked to data quality issues. Textual corpora are often noisy, particularly in domains such as social media, clinical notes, or psychotherapy transcripts. Misspellings, incomplete sentences, and idiosyncratic jargon require sophisticated cleaning methods. Techniques such as normalisation, spelling correction, and domain-specific lexicon construction are therefore indispensable. In clinical research, for example, adapting NER to identify therapeutic constructs (e.g., “alliance,” “rupture”) illustrates how domain expertise must guide technical implementation (Imel et al., 2015).

Statistical and Machine Learning Approaches

Following preprocessing, text mining applies statistical and machine learning techniques to transform textual data into mathematically tractable representations. These approaches constitute the core of computational text analysis, enabling classification, clustering, prediction, and the extraction of latent structures from large corpora.

1) Bag-of-Words and Term Weighting. The earliest and most influential statistical framework is the Bag-Of-Words (BoW) model, in which documents are represented as unordered collections of word counts. Although simplistic in its disregard for syntax and word order, BoW provides a robust baseline that has underpinned decades of research in text categorisation and information retrieval (Manning, Raghavan, & Schütze, 2008). A critical refinement of BoW is the Term Frequency-Inverse Document Frequency (TF-IDF) weighting scheme (Jones, 1972). TF-IDF enhances models' ability to distinguish documents based on salient vocabulary.

2) Document Classification and Supervised Learning. Supervised machine learning has been widely employed to classify documents into predefined categories. Classic algorithms include Naïve Bayes classifiers, which assume conditional independence among features and remain surprisingly effective in text domains (McCallum & Nigam, 1998). Other algorithms, such as Support Vector Machines (SVMs) (Joachims, 1998), have demonstrated strong performance in high-dimensional spaces typical of textual data, owing to their ability to maximise separation margins between classes. In psychotherapy research, for example, SVMs and logistic regression have been used to automatically code therapist behaviours, such as motivational interviewing strategies, achieving accuracy comparable to that of trained human coders (Atkins et al., 2014). Such applications illustrate the translational potential of supervised learning for scaling behavioural coding in clinical contexts.

3) Clustering and Unsupervised Learning. Beyond supervised approaches, unsupervised learning methods play a pivotal role in discovering latent patterns in text without predefined labels. Clustering algorithms, such as k-means and hierarchical clustering, group documents based on lexical similarity, facilitating the exploratory analysis of large corpora (Steinbach, Karypis, & Kumar, 2000). These methods are particularly valuable in domains where taxonomies are underdeveloped or categories are emergent, such as patient narratives or social media data. Dimensionality reduction techniques, including Latent Semantic Analysis (LSA) (Deerwester et al., 1990), extend clustering by uncovering hidden semantic dimensions. By decomposing term-document matrices using singular value decomposition, LSA captures synonymy and polysemy more effectively than surface-level co-occurrence, paving the way for later probabilistic topic models.

4) Transition to Neural Architectures. While traditional machine learning relies heavily on feature engineering, the advent of deep learning shifted the paradigm towards representation learning. Neural architectures such as Convolutional Neural Networks (CNNs) (Kim, 2014) and Recurrent Neural Networks (RNNs) (Mikolov et al., 2010) have been adapted for text classification, sentiment analysis, and sequence labelling. These models learn hierarchical representations of text directly from data,

thereby reducing the dependence on manual feature design. This transition has been further accelerated by the rise of transformer-based models, which will be discussed in detail in later sections on generative AI. For text mining specifically, embeddings derived from transformers (e.g., BERT, RoBERTa, GPT-based models) have achieved state-of-the-art performance in clustering, classification, and retrieval tasks (Devlin et al., 2019; Liu et al., 2019).

The integration of statistical and machine learning methods has profoundly influenced how textual meaning is operationalised. Whereas earlier approaches privileged word frequency and statistical salience, contemporary neural methods emphasise contextual representation and dynamic semantics. Nevertheless, each method embodies trade-offs between interpretability and predictive accuracy. For instance, TF-IDF and Naïve Bayes are transparent and computationally efficient, yet they overlook context. Deep learning models, on the other hand, capture nuanced semantics but often operate as “black boxes,” raising epistemological challenges for domains such as psychotherapy research, where interpretability remains crucial (Kerz, 2023).

Topic Modelling

Topic modelling represents one of the most influential families of unsupervised learning techniques in text mining. Its objective is to uncover latent thematic structures within large collections of documents, thereby enabling researchers to identify, summarise, and compare topics that emerge across heterogeneous corpora. Unlike supervised methods, topic modelling does not require pre-annotated data, making it especially valuable for domains where labelled corpora are scarce or where exploratory discovery is a primary aim.

Latent Dirichlet Allocation (LDA)

The most widely adopted algorithm in this tradition is Latent Dirichlet Allocation (LDA), introduced by Blei, Ng, and Jordan (2003). LDA is a generative probabilistic model that assumes documents are mixtures of latent topics, where each topic is characterised by a distribution over words. The model posits that:

1. Each document is generated by first sampling a distribution of topics from a Dirichlet prior.
2. For each word in the document, a topic is drawn from this distribution.
3. A word is then sampled from the topic’s word distribution.

This formulation provides a Bayesian framework that is computationally tractable via approximation techniques such as variational inference and Gibbs sampling. One of the strengths of LDA lies in its interpretability: topics can be inspected as ranked lists of words, enabling qualitative analysis of thematic patterns.

Applications of LDA are extensive. In psychotherapy research, it has been used to analyse session transcripts, revealing thematic dynamics such as the distribution of therapeutic techniques, patient concerns, and relational shifts across treatment phases (Imel et al., 2015; Atkins et al., 2012). In other fields, LDA has been employed to map the evolution of scientific disciplines (Griffiths & Steyvers, 2004), to study media framing (DiMaggio, Nag, & Blei, 2013), and to detect emerging trends in biomedical literature (Cohen & Hersh, 2005).

Dynamic and Hierarchical Topic Models

While LDA assumes that topic distributions are static, subsequent innovations extended the framework. Dynamic Topic Models (DTMs) (Blei & Lafferty, 2006) incorporate temporal dimensions, allowing topics to evolve over time. This innovation is particularly valuable for longitudinal corpora, such as psychotherapy sessions or social media discussions, where thematic trajectories and shifts in discourse are central analytic targets.

Hierarchical Topic Models (Griffiths et al., 2004) enable topics to be organised into tree-like structures, capturing both broad thematic categories and more specific subtopics. This hierarchical perspective mirrors the way human knowledge is often structured, offering a more nuanced representation than flat topic spaces.

Neural and Contextualised Topic Models

The rise of deep learning has inspired new generations of topic models that leverage neural architectures and contextual embeddings. Neural Variational Document Models (NVDMs) (Miao, Yu, & Blunsom, 2016) and ProdLDA (Srivastava & Sutton, 2017) combine variational autoencoders with probabilistic modelling to improve both scalability and coherence.

Recent developments have integrated transformer-based embeddings into topic modelling. For instance, BERTopic (Grootendorst, 2022) employs BERT-derived embeddings, dimensionality reduction techniques (UMAP), and clustering algorithms (HDBSCAN) to derive topics that are both semantically coherent and context-sensitive. This hybrid approach overcomes some of the limitations of LDA, such as its reliance on bag-of-words representations and demonstrates strong performance in capturing nuanced themes across corpora.

Assessing topic model quality requires both quantitative metrics and qualitative interpretability. Quantitative measures include perplexity (a probabilistic measure of model fit) and topic coherence scores (Mimno et al., 2011), the latter correlating more strongly with human judgments of interpretability. However, no single metric captures the full complexity of thematic coherence, making human inspection indispensable. Qualitative evaluation often involves examining top words per topic, generating word clouds, or annotating topics with expert-driven labels. In psychotherapy applications, interpretability is particularly crucial, since topics must correspond to clinically meaningful categories rather than artefacts of statistical modelling.

Topic modelling exemplifies the tension between exploration and explanation in computational text analysis. While it offers a powerful tool for discovering hidden structures, its results are inherently shaped by modelling assumptions, parameter choices, and pre-processing decisions. For example, the decision to lemmatise or not can drastically alter topic word distributions, while the choice of topic number (k) introduces arbitrariness that requires both statistical tuning and substantive justification (Wallach et al., 2009). Furthermore, the interpretive labour in topic modelling is non-trivial: researchers must resist the temptation to reify statistical clusters as “real” topics and instead situate them as heuristic constructs that require contextual validation. In psychotherapy research, this entails triangulation with clinical expertise, qualitative coding, and process measures to ensure that discovered topics align with therapeutic phenomena (Atkins et al., 2014).

Sentiment Analysis

Sentiment analysis, also known as opinion mining, is a prominent branch of text mining that aims to detect, classify, and quantify subjective information expressed in written or spoken language. Whereas topic modelling uncovers latent thematic structures, sentiment analysis focuses on the affective valence of text, distinguishing whether expressions are positive, negative, or neutral, and in more advanced forms, capturing complex emotions and attitudes. The origins of sentiment analysis can be traced to early applications in computational linguistics and information retrieval (Pang, Lee, & Vaithyanathan, 2002). Classical approaches relied on lexicon-based methods, in which dictionaries of sentiment-laden words (e.g., “happy,” “angry”) are matched to texts, yielding polarity scores. While interpretable and domain-flexible, these methods often fail to capture context, irony, or polysemy.

Subsequent advances adopted machine learning techniques, where classifiers such as Support Vector Machines or logistic regression are trained on labelled corpora of reviews, comments, or clinical notes (Turney, 2002; Pang & Lee, 2008). More recent work leverages deep learning architectures, including Recurrent Neural Networks (RNNs), Long Short-Term Memory networks (LSTMs), and, most effectively, transformer-based models such as BERT (Devlin et al., 2019). These models excel at capturing context-dependent semantics, thereby overcoming many limitations of lexicon-based approaches.

Whereas sentiment analysis traditionally emphasises a binary or ternary classification of polarity, emotion mining seeks a more granular understanding of affective states. Frameworks such as Ekman's six basic emotions (Ekman, 1992) or Plutchik's wheel of emotions (Plutchik, 1980) provide taxonomies that guide annotation and model training. Computational models can thus classify texts into discrete emotional categories (e.g., joy, fear, sadness, anger) or map them onto dimensional affect models (valence, arousal, dominance). For example, Mohammad and Turney (2012) developed the NRC Emotion Lexicon, which has become a widely used resource for multi-emotion annotation across domains. More advanced models employ multi-label classification, recognising that texts often express overlapping emotions.

Sentiment and emotion analysis have been extensively applied in commercial, political, and biomedical domains. In marketing, they are used to monitor consumer opinions; in political science, to analyse voter sentiment during elections (Tumasjan et al., 2010); and in public health, to track mental health trends through social media discourse (Coppersmith et al., 2014). In psychotherapy research, emotion mining has been employed to analyse therapeutic dialogue, providing insights into the emotional tone of interactions and its relation to therapeutic alliance and outcomes (Xiao et al., 2016). For instance, patterns of positive affect in client speech may correlate with therapeutic progress, whereas persistent negative affect may signal treatment resistance or a risk of relapse.

The deployment of sentiment and emotion mining raises several ethical and methodological challenges. First, affective states expressed in language are often ambiguous, culturally situated, and context-dependent; as such, computational classification risks reductionism. Second, in clinical contexts, sentiment detection must be handled with caution, as algorithmic misclassifications could lead to inappropriate interventions or misinterpretation of patient distress.

Ultimately, the application of emotion mining in psychotherapy intersects with broader discussions surrounding privacy and informed consent. Anonymised analysis of therapy transcripts can yield valuable insights at the aggregate level, but the individual sensitivity of such data necessitates robust safeguards and transparent communication with stakeholders (Darby &

Weinstock, 2018; Straw, 2021). Sentiment and emotion analysis illustrate the dual trajectory of computational text analysis: the pursuit of increasingly accurate models on the one hand, and the necessity of grounding these models in theoretical and clinical relevance on the other. While state-of-the-art systems achieve impressive performance benchmarks, their practical value depends on careful integration with disciplinary frameworks.

2.2. Generative AI and Prompt Engineering

Generative Artificial Intelligence (Generative AI) refers to a class of computational models designed not merely to analyse or classify data, but to create new content that resembles human-generated artefacts. Unlike traditional discriminative models, which learn to map inputs to outputs (e.g., classifying an image as containing a cat or not), generative models learn the underlying probability distribution of the training data, thereby enabling the generation of novel instances that maintain statistical and structural coherence with the source corpus (Goodfellow et al., 2014; Bommasani et al., 2021).

At its core, Generative AI is distinguished by its ability to synthesize text, images, audio, and other modalities in ways that can appear creative, context-sensitive, and adaptive. This capacity arises from deep learning architectures – most notably Generative Adversarial Networks (GANs), Variational Autoencoders (VAEs), and transformer-based language models – which have transformed the landscape of artificial intelligence over the past decade. While GANs and VAEs dominate image and multimedia generation, transformer-based models underpin contemporary advances in natural language generation, enabling machines to produce coherent paragraphs, dialogue, and narratives at scale (Vaswani et al., 2017; Radford et al., 2019).

Generative AI differs from earlier symbolic or rule-based approaches in that its outputs are not explicitly pre-programmed but rather emerge from patterns learned across massive datasets. This makes it both more flexible and more unpredictable. For instance, where a rule-based chatbot could only deliver predefined responses, a generative model such as GPT-4 can compose an essay, draft computer code, or translate text in a style that adapts dynamically to the user's input.

Generative AI systems share several defining features:

1. Creativity through Probabilistic Modelling – These models generate new outputs by sampling from probability distributions rather than retrieving memorised templates.

2. Multi-modality – Contemporary generative systems operate across modalities, from text and images to speech and video, with increasing emphasis on multimodal architectures (e.g., OpenAI’s GPT-4o or Google’s Gemini).
3. Scalability – Generative performance improves significantly with scaling laws in data, model parameters, and computational resources (Kaplan et al., 2020).
4. Adaptivity – Generative systems can contextualise responses in relation to user prompts, which enables interactive and task-specific applications.

Initially, generative models were conceived as assistive tools – augmenting human creativity in art, design, and language. However, their capacity has rapidly expanded to roles that approach autonomous content generation, blurring the boundary between human and machine creativity (Floridi & Chiriatti, 2020). For example, text-to-image models such as DALL·E 2 can create visual compositions that imitate artistic styles, while large language models generate full-length essays, policy drafts, or computer code with minimal human supervision. Generative AI thus represents a paradigmatic shift in the philosophy of artificial intelligence. Rather than simply encoding human knowledge, it operates as a knowledge-producing system, capable of generating novel hypotheses, design options, or discursive possibilities. This capacity has prompted intense debate over questions of authorship, originality, and the status of machine-generated artefacts (Boden, 2016; Marcus, 2022).

Large Language Models (LLMs)

Large Language Models (LLMs) represent the most influential and widely deployed subclass of generative AI systems. They are built on transformer-based architectures (Vaswani et al., 2017) and trained on vast corpora of textual data, enabling them to capture complex statistical regularities in language. By predicting the probability distribution of the next token in a sequence, LLMs generate coherent, contextually relevant, and often stylistically adaptive text across a wide range of domains.

The trajectory of LLMs can be traced through successive generations of increasingly sophisticated models:

- **Early Precursor Models:** Word embeddings such as Word2Vec (Mikolov et al., 2013) and GloVe (Pennington et al., 2014) introduced distributed semantic representations, laying the foundation for contextual language modelling.
- **Contextual Representations:** Models such as ELMo (Peters et al., 2018) and BERT (Devlin et al., 2019) demonstrated the power of contextualised embeddings, capturing polysemy and nuanced meaning.

- **Autoregressive Generative Models:** The GPT series developed by OpenAI marked a decisive shift toward generative capabilities. GPT-2 (Radford et al., 2019) demonstrated the potential for coherent multi-paragraph generation, while GPT-3 (Brown et al., 2020) scaled to 175 billion parameters, exhibiting emergent abilities in reasoning, translation, and coding without requiring task-specific fine-tuning.
- **Frontier Models:** More recent systems such as GPT-4 (OpenAI, 2023), Anthropic’s Claude, and Google DeepMind’s Gemini represent a new paradigm of multimodal and instruction-tuned LLMs, integrating text, vision, and speech while demonstrating strong performance in knowledge-intensive reasoning.

The success of LLMs derives from three interrelated principles:

1. **Transformer Architecture** – The self-attention mechanism allows models to weigh the relative importance of words in a sequence, enabling them to capture long-range dependencies and discourse-level coherence (Vaswani et al., 2017).
2. **Pre-training on Massive Corpora** – LLMs are exposed to terabytes of text from diverse sources (web data, books, academic articles), which allows them to approximate a general-purpose model of language.
3. **Fine-tuning and Alignment** – After pre-training, models are fine-tuned on curated datasets and aligned using techniques such as Reinforcement Learning from Human Feedback (RLHF) to optimise usefulness, safety, and adherence to human instructions (Ouyang et al., 2022).

The impact of LLMs is reflected in their exponential growth in scientific publications and market adoption. A bibliometric analysis by Bender et al. (2021) highlights the surge of papers addressing ethical, technical, and societal implications. Commercially, the release of ChatGPT in late 2022 catalysed mass adoption, reaching over 100 million users in a matter of weeks (Hu et al., 2023). This unprecedented uptake underscores both the transformative potential of LLMs and the urgency of addressing their risks.

Despite their achievements, LLMs face significant limitations.

- **Hallucination of Facts:** models can produce plausible yet false information.
- **Opacity:** their internal decision-making processes remain mostly uninterpretable.
- **Bias and Fairness:** training data contain social biases, which are often reflected and sometimes worsened in outputs (Bolukbasi et al., 2016).

- **Environmental Costs:** the energy required to train and operate large-scale models is considerable (Strubell et al., 2019).

These limitations highlight the importance of responsible deployment and ongoing research into model interpretability, efficiency, and alignment.

Applications of Generative AI

The rapid advancement of generative AI has led to a wide range of applications across research, healthcare, education, creative industries, and commercial sectors. Its adaptability across domains stems from its ability to generate text, images, code, and other multimodal outputs, making it a versatile cognitive assistant.

1) **Research and Knowledge Discovery.** In science, generative AI accelerates the production and dissemination of knowledge. Large language models facilitate literature review, summarisation of research findings, and hypothesis generation (Hope et al., 2022). For instance, LLMs have been used to automatically extract structured data from biomedical literature, thereby aiding drug discovery pipelines and protein design (Stokes et al., 2020).

2) **Healthcare and Clinical Practice.** Generative AI has shown promise in clinical documentation, decision support, and patient interaction. Models can generate discharge summaries, streamline electronic health records, and provide draft diagnostic impressions (Meskó, 2023). In mental health, conversational agents powered by LLMs support therapeutic dialogue and psychoeducation, though their integration raises ethical concerns around safety and oversight (Miner et al., 2016). Beyond text, multimodal generative systems contribute to medical imaging analysis by synthesising radiological data for training and anomaly detection.

3) **Education and Pedagogy.** Educational contexts are increasingly experimenting with generative AI as a tool for personalised learning. Systems can adapt explanations to a learner's level, provide automated feedback, and generate practice problems across subjects (Farhud & Zokaei, 2021). Virtual tutors based on LLMs can simulate Socratic dialogue, fostering critical thinking skills. At the institutional level, generative models automate administrative tasks, freeing educators to devote more time to human-centred teaching.

4) **Creative and Cultural Industries.** One of the most visible impacts of generative AI lies in the creative industries, where text-to-image, text-to-video, and language models augment artistic production. Models such as DALL·E 2 and Stable Diffusion generate visual artefacts in diverse artistic styles, while LLMs are deployed for screenwriting, advertising copy, and poetry (Elgammal et al., 2017). These tools blur the distinction between human and machine creativity, simultaneously

expanding creative possibilities and raising questions of authorship, originality, and intellectual property rights (Floridi & Chiriatti, 2020).

5) Software Engineering and Automation. Generative AI has revolutionized programming practices by creating models that generate functional code snippets, refactor legacy code, and identify bugs. GitHub Copilot, powered by OpenAI's Codex model, exemplifies this shift, enhancing developer productivity and lowering the barrier to entry for non-programmers (Chen et al., 2021). In parallel, generative approaches support data augmentation for machine learning pipelines, enabling more robust model training.

6) Commercial and Business Applications. The commercial uptake of generative AI has been exponential. Businesses deploy LLMs to automate customer service through chatbots, generate marketing content, and optimise workflows (Hu et al., 2023). In financial services, models assist in fraud detection and report generation; in law, they support contract drafting and legal research.

7) Cross-Domain and Multimodal Integration. Recent advances emphasise the integration of modalities, where models process and generate across text, image, and audio simultaneously. For instance, GPT-4o and Google's Gemini unify multimodal reasoning, enabling real-time dialogue with textual, visual, and auditory inputs. Such systems herald a shift toward general-purpose AI agents that can function across domains without requiring domain-specific engineering.

Exponential Growth of Generative AI in Research and Commerce

An unprecedented exponential growth has marked the trajectory of Generative AI in both academic research and commercial domains. Bibliometric analyses consistently indicate a surge in scholarly output beginning in the early 2010s, following the introduction of deep learning architectures, and accelerating dramatically after 2018 with the advent of transformer-based models (Vaswani et al., 2017; Radford et al., 2019). According to surveys of AI research, publications explicitly referencing generative models – including GANs, VAEs, and large language models – have increased by more than 400% over the past decade, positioning Generative AI as one of the fastest-growing subfields within machine learning (Bommasani et al., 2021).

Within academia, the exponential rise reflects not only volume but also diversification of disciplines engaging with generative methods. Initially confined to computer science and engineering, research is now pervasive in biomedicine, psychology, linguistics, law, and the social sciences. For instance, biomedical applications include the generation of protein structures (Stokes et al., 2020), while social science research employs text generation to model discourse and explore narrative structures

(Caliskan, 2022). This diffusion underscores the interdisciplinary adoption of generative AI, expanding its role from a technical innovation to a fundamental research paradigm across fields.

Along with academic dissemination, commercial adoption has surged significantly. The release of widely accessible systems – including OpenAI’s GPT series, Stability AI’s Stable Diffusion, and Midjourney – has boosted mainstream awareness and sped up business adoption. Market analysts predict that the global generative AI market will grow from USD 10.5 billion in 2022 to over USD 118 billion by 2032, with a compound annual growth rate (CAGR) of more than 30% (MarketsandMarkets, 2023). Key industries adopting these technologies include marketing, healthcare, financial services, and entertainment, where generative models enhance efficiency in content creation, automation, and product development.

The rapid rise of Generative AI can be linked to the interaction between academic breakthroughs and industry investment. Innovations in neural architectures from academia quickly turn into commercial products, while industry data and computing power accelerate research progress. This feedback loop has led to shorter innovation cycles, with advances in pretraining methods, scaling laws (Kaplan et al., 2020), and reinforcement learning from human feedback (Christiano et al., 2017) being rapidly integrated into research papers and production systems. This growth is not only in terms of size but also has significant social and cultural impacts. The widespread use of generative tools in everyday life has altered public discussions about AI, sparking debates over ethics, creativity, authorship, and job displacement (Marcus, 2022; Floridi & Chiriatti, 2020).

Prompt Engineering

The rapid diffusion of large language models (LLMs) has led to the emergence of prompt engineering as a central methodological practice (Chen et al., 2023). Prompt engineering refers to the deliberate design and optimisation of input queries – or “prompts” – to elicit desired behaviours from generative models (Liu et al., 2021). Unlike traditional programming, where explicit code governs output, prompt engineering operates through linguistic instructions that guide a model’s probabilistic text generation. This technique has emerged as both a scientific discipline and a professional skill, reflecting the shift toward human-AI interaction through language.

Prompt engineering involves creating, organizing, and refining input prompts to improve the effectiveness, accuracy, or creativity of outputs from generative models. It encompasses a range of strategies, from straightforward rephrasing of natural language questions to more advanced techniques, including role specification, few-shot learning examples, and chain-of-thought

scaffolding (Wei et al., 2022). Its scope goes beyond language: multimodal systems also use prompt engineering with images, audio, or symbolic representations.

Although the concept of prompting dates to early rule-based chatbots (Weizenbaum, 1966), its significance has intensified with the advent of transformer-based architectures. As LLMs such as GPT-3 demonstrated unexpected emergent abilities – for instance, solving reasoning tasks without explicit training – the design of prompts became critical to harness these latent capacities (Brown et al., 2020). This marked a paradigmatic shift in human-machine interaction: users transitioned from coding deterministic instructions to crafting linguistic cues that exploit models’ pre-trained knowledge. Several techniques define the practice of prompt engineering:

- Zero-shot prompting: Supplying only a task description, relying on the model’s generalisation abilities (Kojima ET AL., 2022).
- Few-shot prompting: Providing examples of task–response pairs within the prompt to guide output (Mathur & Shrivastava, 2024).
- Chain-of-thought prompting: Explicitly requesting reasoning steps, improving performance on complex tasks (Wei et al., 2022).
- Role prompting: Framing the model as a specific persona (e.g., “act as a clinical psychologist”), aligning responses with domain-specific expectations (PromptLayer, 2023).
- Instruction tuning and reinforcement learning from human feedback (RLHF): Though often model-level interventions, they interact with prompt design by calibrating how models interpret instructions (Christiano et al., 2017).

Prompt engineering has become pivotal in scientific research, where precise formulations determine the reliability of model-assisted analyses (Park & Choo, 2024). In healthcare, carefully engineered prompts are essential for ensuring safety in clinical decision support (Meskó, 2023). In education, prompts guide adaptive tutoring systems in tailoring responses to learners’ needs (Farhud & Zokaei, 2021). Creative industries utilise prompt engineering to generate stylistically constrained outputs, from poetry to visual art prompts (Liu & Chilton, 2023). The business sector applies it to optimise customer support chatbots, ensuring consistent tone and accuracy.

The growing interest in prompt engineering has driven efforts to create systematic frameworks. Recent surveys outline taxonomies of prompt types, benchmark datasets for assessment, and structured guidelines (Sahoo et al., 2025; Zhu et al., 2023). Benchmark datasets and shared tasks are increasingly utilized to enable systematic comparisons among models and techniques (Schulhoff et

al., 2024). Automated prompt optimization methods have also emerged, using reinforcement learning, evolutionary search, or iterative refinement; notable examples include Automatic Prompt Engineer (APE) (Zhou et al., 2023), Automatic Prompt Optimization (APO) (Pryzant et al., 2023), and EvoPrompt (Guo et al., 2023). New strategies, such as automatic prompt generation and meta-prompting, indicate a shift towards reducing dependence on human intuition, transitioning the practice from artisanal experimentation to evidence-based engineering (Liu et al., 2021).

As with all generative AI techniques, prompt engineering introduces ethical questions. Poorly designed prompts can amplify biases, produce unsafe outputs, or create misleading information. Moreover, the unequal distribution of expertise in prompt engineering may exacerbate digital divides, privileging those with access to specialised knowledge. For this reason, research increasingly emphasises the need for transparent prompt design practices and guidelines in high-stakes domains such as healthcare, law, and education (Farhud & Zokaei, 2021).

3. Psychotherapy, Computational Language Analysis, and Semantic Shift

This chapter investigates how psychotherapy, as a dialogical and meaning-making practice, can be analysed through the perspective of computational linguistics. It examines how the language used by patients and therapists reflects and influences therapeutic change, emphasizing the interaction between linguistic markers, clinical constructs, and algorithmic models. Attention is given to the phenomenon of semantic shift, viewed both as a clinical sign of psychological transformation and as a measurable linguistic dynamic. The discussion is organized around several interconnected themes:

- *The potential of topic modelling and sentiment analysis to track therapeutic processes.*
- *The integration of quantitative methods with hermeneutic interpretation.*
- *The construction of the concept of “semantic shift”, based on the dynamic changes of topics during the psychotherapeutic process.*
- *The relevance of cognitive mechanisms, such as bias and decision-making, for both psychotherapy and algorithmic systems.*

By bridging insights from clinical theory, process–outcome research, and Natural Language Processing, the chapter highlights current frontiers and open challenges, preparing the ground for the methodological strategies and empirical analyses presented in the following sections.

Psychotherapy is, at its core, a practice of language. Unlike other medical interventions, which often depend on standardized procedures or pharmacological protocols, psychotherapy unfolds within an intersubjective space constituted by words, silences, and narratives. The therapeutic encounter is not simply an exchange of information but an ongoing negotiation of meaning: through metaphors, reformulations, and discursive shifts, patients and therapists co-construct new perspectives on experiences that were previously opaque or distressing (Watson, 2025). Narrative and dialogical approaches have long emphasized that words are never neutral carriers; instead, they are vehicles that shape emotional atonement, relational positioning, and cognitive restructuring (White & Epston, 1990; Frank & Frank, 1991; Angus & McLeod, 2004).

Language in psychotherapy is also profoundly dynamic. A patient’s repeated use of the same lexical item may acquire radically different connotations across sessions. For instance, the word “fear” may

initially signify vulnerability and avoidance, but later emerges as a marker of awareness, resilience, or even transformation. Computational linguistics conceptualizes this phenomenon as a semantic shift, a process in which the associations of a word evolve as it appears in new contexts (Baes, Haslam, & Vylomova, 2023). From a clinical perspective, attending to such transformations is crucial for evaluating the depth and direction of therapeutic change (Mergenthaler & Stinson, 1992; Van Staden & Fulford, 2004). From a computational perspective, these shifts provide measurable indices of evolving meaning structures that can be systematically tracked (Pennebaker, Mehl, & Niederhoffer, 2003; Tausczik & Pennebaker, 2010).

This convergence highlights why psychotherapy has become a fertile domain for interdisciplinary research. By integrating insights from clinical practice, discourse studies, and computational modelling, researchers can illuminate how subjective change is instantiated and reflected in the microstructure of dialogue. Such a perspective resonates with Daniel Kahneman's (2011) distinction between fast, intuitive thought (System 1) and slower, reflective reasoning (System 2). Psychotherapy often enables a shift from the immediacy of unfiltered, emotionally charged experience to more deliberate, coherent, and narratively integrated accounts. In this sense, language functions simultaneously as the medium of change and as the trace through which cognitive–emotional reorganization can be observed (Stiles, 1992; Angus, Levitt, & Hardtke, 1999).

The advent of computational language approaches has opened novel avenues for examining psychotherapy as a dynamic, meaning-making process. Clinical researchers have long observed that therapeutic change is reflected in the ways patients narrate their experiences and position themselves through language (Stiles, 1992; Mergenthaler & Stinson, 1992). More recently, topic modelling techniques (Blei, Ng, & Jordan, 2003; Blei & Lafferty, 2009) have provided a systematic means of mapping the distribution of thematic content across therapy sessions, enabling researchers to identify which themes gain or lose prominence over time. For instance, longitudinal analyses may reveal a decline in symptom-related topics alongside an increase in themes linked to insight or therapeutic alliance (Imel et al., 2015). Such topic trajectories can be quantitatively associated with clinical outcome measures, thereby offering a novel lens on process–outcome relationships.

Beyond the level of thematic content, computational methods now allow researchers to capture the more subtle phenomenon of *semantic drift* – the gradual change in meaning that words or concepts acquire as they are used across different contexts (Hamilton, Leskovec, & Jurafsky, 2016). Embedding-based models, such as Word2Vec (Mikolov et al., 2013) and BERT (Devlin et al., 2019), as well as more recent large language models, can detect how a patient's recurrent use of terms like "*control*" or "*freedom*" shifts in their semantic associations over the course of treatment. A word

initially tied to anxiety or rigidity may later emerge in connection with autonomy, mastery, or resilience. From a clinical perspective, such transformations are not merely linguistic artifacts but markers of psychological reorganization – an indication that the patient’s internal world and self-narratives are undergoing restructuring in ways that underpin therapeutic progress (Pennebaker et al., 2015; Watson, 2025). Sentiment analysis provides yet another analytic layer. By quantifying the affective valence of linguistic expression, it becomes possible to trace fluctuations in emotional tone across the therapeutic trajectory (Tausczik & Pennebaker, 2010). A progressive increase in positive affect markers may correspond with improved therapeutic alliance, whereas sustained negative sentiment might foreshadow rupture or stagnation (Mergenthaler, 2008). While sentiment extraction cannot substitute for clinical interpretation, it enriches the analytic toolkit, offering complementary signals that may anticipate shifts in process and outcome (Tanana et al. 2021).

Significantly, these computational tools extend beyond descriptive analytics by fostering a deeper integration of qualitative and quantitative perspectives. Psychotherapy has always relied on the interpretive sensitivity of clinicians – the ability to attune to subtle shifts in a patient’s wording or tone. Computational analysis does not replace this hermeneutic stance but amplifies it, offering scalable methods for detecting linguistic transformations that might otherwise remain unnoticed. The investigation of semantic shift exemplifies this interdisciplinary synthesis, marrying the statistical rigor of vector-space models with the qualitative richness of clinical meaning-making (Angus & Kagan, 2013; Salvatore et al., 2017).

In this comprehensive research, I tried to conduct a semantic analysis of the text, considering the patient’s narrative and his experiential modes as interconnected, mutually influencing levels of Self (Ricoeur, 1992; Arciero & Bondolfi, 2012; Liccione, 2019; Demaria et al., 2023). The theoretical idea throughout this work is that human beings are unique in their ability to transform lived experiences into verbal expressions of meaning. Expanding on Ricoeur’s ideas, interpretation entails appropriation, the process by which something initially unfamiliar becomes familiar and one’s own. Therefore, interpretation is the process through which the revelation of new modes of being enhances the client’s capacity for self-knowledge (Ricoeur, 1976; 1981). Semantic analysis rooted in the hermeneutic tradition should prioritize uncovering the patient’s lived experiences rather than creating abstract narratives. This methodology involves examining the themes that organize a text. In linguistics, a topic refers to the subject matter of a conversation and is defined by contextual considerations that give it meaning (e.g., Traini, 2008). Changes in topic trajectories are associated with changes in outcomes (Atzil-Sonim et al., 2021), indicating that analysing topic trajectories during interviews can reveal changes in the patient’s experiential ways of being (Griffiths, Steyvers & Tenenbaum, 2007).

3.1. Text Analysis Tools for Psychotherapy Research

Psychotherapy research has historically relied on the meticulous work of human coders, who manually annotated transcripts to capture the complexity of clinical interactions. Early approaches, such as the Coding of Therapist Statements (Stiles & Sultan, 1979) and the Therapeutic Cycles Model (Mergenthaler, 1996), offered structured frameworks for linking linguistic forms to psychological change. Similarly, the Vanderbilt Psychotherapy Process Scale (O'Malley et al., 1983) or the Penn Helping Alliance Scales (Luborsky et al., 1996) attempted to operationalize relational constructs such as empathy, alliance, and rupture. These tools were crucial in shaping psychotherapy research, as they provided systematic methods for interpreting the subtle dance of words, silences, and emotional tones that characterizes the therapeutic encounter.

However, manual systems were not without limitations. They required extensive rater training, suffered from inter-rater reliability issues, and were prohibitively time-consuming when applied to large corpora (Hill, 2011). For example, coding even a single session using the Therapeutic Cycles Model could take several hours, rendering large-scale studies nearly impossible. Moreover, human raters inevitably brought their own biases and interpretative frameworks to the task, raising questions about objectivity and reproducibility. These challenges created a methodological bottleneck: psychotherapy research risked being confined to small case studies or limited samples, leaving unanswered questions about generalizability and broader process-outcome relationships.

It was against this backdrop that the first computational attempts emerged. Pioneering efforts such as the Gottschalk-Gleser Content Analysis Scales (Gottschalk & Lolas, 1989) sought to automate the measurement of psychological constructs by using predefined dictionaries of words linked to affective or cognitive categories. These approaches remained constrained by rigid, rule-based definitions and lacked sensitivity to context. For instance, the word '*control*' could be coded as 'dominance' or 'autonomy' depending on the dictionary entry, but its meaning in a specific therapeutic dialogue – whether signalling rigidity, anxiety, or resilience – was often lost.

The computational turn in psychotherapy research, spurred by advances in Artificial Intelligence and Natural Language Processing (NLP), fundamentally transformed this landscape. NLP enables machines to process and interpret human language at scale (Manning & Schütze, 1999), while Text Mining (TM) integrates computational linguistics, statistics, and machine learning to extract patterns from large corpora of unstructured text (Bolasco, 2005; Kherwa & Bansal, 2019). This paradigm shift enabled researchers to move beyond the constraints of small-scale, coder-dependent studies toward scalable, data-driven analyses that can detect patterns invisible to the naked eye.

A good example of this transition can be seen in the evolution from dictionary-based systems such as LIWC (Pennebaker et al., 2015), which counted word frequencies within predefined categories, to probabilistic models such as Latent Dirichlet Allocation (LDA) (Blei, Ng, & Jordan, 2003). LDA permits the discovery of latent themes within psychotherapy transcripts without human pre-annotation. This opened the possibility of identifying recurring motifs in therapy, such as conflict, self-reflection, or alliance, that emerge directly from the data.

This computational expansion did not aim to replace clinical interpretation but to extend it. Just as Daniel Kahneman (2011) distinguishes between the fast, intuitive judgments of *System 1* and the slower, more deliberative reasoning of *System 2*, one could argue that manual coding embodied the reflective stance of *System 2*, while computational methods operate more like *System 1*, rapidly scanning massive amounts of language to detect patterns that would otherwise remain unnoticed. The integration of these two modes – clinical interpretation and computational detection – represents not a rupture but a complementarity, where each corrects the blind spots of the other. As Watson (2025) emphasizes, the central task of psychotherapy process research is to identify the productive in-session processes that drive change. This perspective reminds us that computational methods are tools for detecting and clarifying those micro-processes that support therapist responsiveness.

Among the broad spectrum of NLP techniques, topic modelling has become one of the most influential and widely applied approaches in psychotherapy research. Topic modelling belongs to the family of probabilistic generative models, which represent documents as mixtures of topics, with each topic defined as a distribution of words (Steyvers & Griffiths, 2007; Reuter et al., 2025). The LDA model (Blei, Ng, & Jordan, 2003) has provided researchers with a scalable and theoretically grounded method for uncovering latent themes in clinical dialogue (Atkins et al., 2012; Imel et al., 2015). The central advantage of LDA is its ability to identify patterns of meaning that emerge organically from the data rather than imposing predefined categories. This feature is particularly well suited to psychotherapy, where sessions are often characterized by fluid, nonlinear shifts in narrative and affect. In this sense, topic modelling resonates with the very logic of therapeutic dialogue: just as clients' narratives evolve through free association and reinterpretation, LDA traces the statistical co-occurrence of words to infer latent psychological structures. Several landmark studies illustrate how topic modelling can shed light on psychotherapy processes and outcomes. Atkins et al. (2012) were among the first to apply LDA to transcripts of couple and family therapy, demonstrating that automatically extracted topics corresponded to clinically meaningful categories such as conflict, intimacy, and problem-solving. Their findings showed that computationally derived topics could predict therapist behaviours and relational dynamics, suggesting that LDA could function as a bridge between naturalistic dialogue and process coding systems.

Building on this, Imel et al. (2015) employed topic modelling to analyse thousands of psychotherapy transcripts, showing that the distribution of topics across sessions was predictive of therapeutic outcomes. For instance, a decrease in symptom-focused language combined with an increase in insight-related issues was associated with improved client functioning. This finding demonstrated, for the first time, that unsupervised computational models could track process–outcome relationships traditionally studied through labour-intensive manual coding. More recent studies have combined LDA with advanced statistical frameworks to model temporal trajectories of therapeutic change. For example, Atzil-Slonim et al. (2021) used LDA-derived topics to predict alliance ruptures and repairs, revealing that sessions dominated by negative themes were more likely to precede ruptures. In contrast, sessions emphasizing collaboration and reflection predicted alliance strengthening. Similarly, Lalk et al. (2024) integrated topic modelling with transformer-based embeddings (e.g., BERT), showing that contextual word representations capture nuances in therapeutic discourse that LDA alone might miss, such as irony, metaphors, or subtle affective cues.

A complementary approach was introduced by Liccione and Siciliano in a paper published in the *Journal of Psychotherapy Research* (2025), which combined topic modelling with classical statistical analyses (Generalized Mixed Additive Models) to examine topic trajectories in a single-case psychotherapy study. Moreover, by employing correlation-based measures among LDA topics (Topic Matrix Correlation and qualitative overlap techniques), we aim to highlight how therapeutic transformation can be captured not only in the content of emerging topics but also in their semantic relationships. This approach introduced a novel perspective on topic trajectories, emphasizing that the therapeutic process is not just about which topics emerge but also about how they converge, diverge, and reorganize over time.

One of the most compelling contributions of topic modelling is its ability to represent topic trajectories across therapy. Unlike traditional coding, which captures static snapshots, LDA enables researchers to track the rise and fall of themes on a session-by-session basis (Atkins et al., 2014; Imel et al., 2015; Watson, 2025). For instance, a client beginning therapy might focus heavily on symptoms such as anxiety, loss, or conflict. As therapy progresses, these topics may give way to narratives of insight, resilience, or relational growth. Tracking such transitions provides an empirical counterpart to clinicians' intuitive sense of therapeutic progress. Importantly, these trajectories can be statistically linked to outcome measures, thereby creating a direct bridge between computational linguistics and research on the effectiveness of psychotherapy. For example, Rosen-Zvi et al. (2010) demonstrated that topic proportions could predict patient improvement scores, while Salvatore and colleagues (2012) showed that co-occurrence networks derived from topics mirrored the developmental reorganization of client narratives.

While LDA offers significant advantages, it also faces limitations, particularly in capturing the contextual richness of language. For example, words like '*control*' or '*freedom*' may cluster into the same topic across sessions. Still, the evolving meaning of these terms – whether signalling rigidity or autonomy – requires more context-sensitive models. To address this, hybrid approaches have emerged: combining LDA with Generalized Additive Mixed Models (GAMMs) to model nonlinear change over time (Liccione & Siciliano, 2025) or integrating LDA with embedding-based methods that capture semantic drift (Hamilton, Leskovec, & Jurafsky, 2016). Such extensions point toward a future in which topic modelling is not seen as an endpoint but as a foundation for richer models of therapeutic dialogue can be built. As Wampold and Imel (2015) emphasized, psychotherapy is always a meaning-making practice; computational tools are valuable insofar as they remain anchored to clinical interpretation and contribute to understanding the human change process.

Personality traits constitute a fundamental dimension of how individuals perceive, interpret, and respond to both internal and external experiences. Classical trait models, such as the Five-Factor Model (McCrae & Costa, 1999), have consistently demonstrated that characteristics like neuroticism, extraversion, and conscientiousness shape affective tendencies, interpersonal behaviours, and coping strategies. In psychotherapy, these traits are not static descriptors but dynamic moderators of how patients engage with therapeutic discourse and relational processes (Smith et al., 2020).

Linguistic research has shown that personality is systematically reflected in language use: individuals high in extraversion tend to employ more positive affect words and social references, whereas those high in neuroticism are more likely to use negative emotion terms and self-referential pronouns (Mairesse et al., 2007; Schwartz et al., 2013). Such findings suggest that psychotherapy transcripts offer a valuable window into personality traits as they manifest in vivo, embedded within narratives of distress, agency, and meaning-making. Recent computational studies have extended this perspective, demonstrating that topic modelling and trait-informed approaches (e.g., PT-LDA or KeyVector; Hu et al., 2017) can uncover latent connections between thematic content and stable personality dimensions (Sperandeo et al., 2020).

From a clinical standpoint, understanding personality traits provides insight into patients' decision-making styles, their susceptibility to cognitive biases, and their emotional orientations toward others. For example, high conscientiousness has been linked to more reflective and deliberate reasoning, whereas impulsivity, associated with low agreeableness or high extraversion, may be associated with faster, more intuitive decision-making (Kahneman, 2011). In therapeutic dialogue, these tendencies can shape how themes are maintained, resisted, or reframed. Thus, integrating personality trait analysis into computational models of psychotherapy holds promise for advancing both theory and

practice. By combining trait-informed linguistic markers with process-oriented indices such as topic trajectories, co-occurrences, and sentiment shifts, researchers may better capture the interplay between stable dispositions and dynamic therapeutic change. This provides a natural bridge to the subsequent sections, which focus on computational approaches to modelling therapeutic discourse.

3.2. Tracking Semantic Change in Psychotherapy

Psychotherapy is not only organized around the exchange of words but also a process of meaning-making through language. Each therapeutic encounter involves the negotiation, modification, and reconstruction of meanings, as clients and therapists co-create new narratives (Angus & Kagan, 2013). This dialogical and hermeneutic dimension implies that language in psychotherapy is never static: words, metaphors, and expressions are dynamically reshaped across sessions in ways that mirror the client's shifting self-concept, relational experiences, and emotional states (Mergenthaler, 1996; Salvatore et al., 2012).

The concept of semantic change refers precisely to these evolving meanings. Terms initially associated with distress or avoidance can gradually acquire connotations of resilience, acceptance, or growth. Ricoeur (1986) described this as a process of *re-narration*, in which familiar symbols are reinterpreted within new storylines of the self. Within psychotherapy research, this transformation has been operationalized as a semantic shift, defined as a measurable change in the distribution, co-occurrence, or contextual embedding of words over time (Salvatore et al., 2012; Vanin, Bolshev & Panfilova, 2025).

From a clinical perspective, such language changes are not merely lexical substitutions but signal deeper reorganizations of the patient's psychological world. As Pennebaker and colleagues (2015) noted, words can act as *psychological markers*; tracing their shifting associations provides a window into therapeutic change at both cognitive and emotional levels. For instance, Mergenthaler's Therapeutic Cycles Model has shown how the oscillation of emotional and cognitive word use reflects phases of therapeutic work (Mergenthaler, 1996), while more recent computational approaches have demonstrated how reorganizations in semantic fields are linked to process–outcome dynamics (Salvatore et al., 2017; Imel et al., 2015).

The rise of computational linguistics has expanded the methodological toolkit for investigating semantic change in psychotherapy. Different approaches operationalize this phenomenon in complementary ways, each providing distinct insights into how language evolves across sessions.

1. Co-occurrence and Semantic Mapping.

Early methods focused on identifying how words co-occur and reorganize over time. Techniques such as Automated Co-occurrence Analysis for Semantic Mapping (ACASM; Salvatore et al., 2017) examine changes in the structure of patients' narratives by detecting clusters of terms that emerge, decline, or shift in association. These reorganizations can reveal the reconfiguration of thematic fields, pointing to underlying transformations in meaning-making processes.

2. Topic Evolution and Trajectory Analysis.

Probabilistic models such as LDA (Blei et al., 2003) have been adapted to psychotherapy research to track the evolution of topics across sessions (Atkins et al., 2012; Atzil-Slonim et al., 2021). By modelling which themes gain or lose prominence, these approaches map the temporal trajectory of therapy. For instance, a gradual reduction in symptom-related topics accompanied by an increase in themes of agency or alliance can signal therapeutic progress (Imel et al., 2015).

3. Embedding-based Models and Semantic Drift.

More recent advances, particularly word embeddings and transformer-based models, allow researchers to measure semantic drift with high sensitivity. Static embeddings such as Word2Vec (Mikolov et al., 2013), GloVe (Pennington et al., 2014), and FastText (Bojanowski et al., 2017) capture semantic similarity through distributional patterns, while contextual embeddings like BERT (Devlin et al., 2019) or GPT architectures can distinguish how the same term acquires different meanings depending on its linguistic context (Radford et al., 2018). This is crucial for psychotherapy, where a word like *control* might initially appear in association with anxiety and rigidity but later shift toward autonomy and mastery.

These approaches not only quantify subtle shifts in meaning but also align with clinical interpretations. For example, Tanana et al. (2021) demonstrated that the semantic neighbourhoods of empathy-related words correlated with human ratings of therapist empathy. The phenomenon of semantic change takes on greater significance in clinical practice, as words that recur across sessions rarely maintain fixed meanings; rather, their associations evolve and often mirror the patient's psychological transformation. Consider the term *failure*. In early therapy stages, it may cluster with words such as *shame*, *fear*, and *rejection*, reflecting a rigid, self-critical narrative. As therapy progresses, computational analyses reveal shifts in its semantic neighbourhood, including terms such as *learning*, *growth*, and *opportunity*. This indicates a reframing of personal narratives: failure is no longer viewed solely as evidence of inadequacy but as a catalyst for development (Ricoeur, 1986).

A similar dynamic occurs with the word relationship. Initially, its semantic field may contain terms such as conflict, betrayal, or distance; later, it may co-occur with words like trust, support, and intimacy. While these transformations are often intuitively recognized by therapists, they can now be systematically detected and quantified through embedding models, thereby providing an objective complement to clinical judgment (Mergenthaler, 1996; Salvatore et al., 2017). Furthermore, Pennebaker et al. (2015) highlighted how words function as psychological markers, with their evolving associations offering a window into therapeutic change at cognitive and emotional levels.

The implications for psychotherapy research and practice are considerable. At the micro level, semantic drift in single words reflects shifts in self-concept, emotion regulation, and relational schemas. At the macro level, these changes contribute to broader narrative transformations, supporting the patient's reconstruction of identity and life story (Angus & Kagan, 2013). Moreover, computational tools enable tracking these changes session by session, offering process markers that may anticipate therapeutic breakthroughs or signal areas of stagnation. Semantic change has potential implications for supervision and training. By analysing linguistic transformations across multiple cases, supervisors can identify patterns associated with effective interventions or highlight moments where meaning-making appears to be blocked. This contributes to the development of evidence-based guidelines that integrate linguistic analysis into clinical decision-making.

Despite its promise, the study of semantic change in psychotherapy faces essential conceptual and methodological challenges. First, definitional clarity remains an issue. In linguistics, *semantic shift* traditionally refers to diachronic changes in meaning across historical periods (Geeraerts, 2009); in contrast, in psychotherapy research, it designates micro-level transformations in clients' narratives within the temporal frame of treatment (Salvatore et al., 2012). This dual usage complicates interdisciplinary dialogue and underscores the need for precise operational definitions that reflect both the hermeneutic and computational dimensions of the phenomenon. Second, methodological diversity presents another difficulty. Co-occurrence analyses (Salvatore et al., 2017), probabilistic topic models (Atkins et al., 2012; Atzil-Slonim et al., 2021), and embedding-based approaches (Hamilton, Leskovec & Jurafsky, 2016; Bamler & Mandt, 2017) each operationalize semantic change differently, making it difficult to compare findings across studies. Moreover, the interpretability of computational metrics remains limited: not every detected shift in word usage corresponds to a clinically meaningful transformation, and expert triangulation is essential (Silverman, 2013; Na et al., 2025). Third, most current studies adopt relatively short-term or small-sample designs, which restricts generalizability. Longitudinal research is needed to determine whether specific patterns of semantic drift reliably predict therapeutic outcomes across populations, modalities, and cultural contexts.

A promising direction is the integration of multiple analytic levels: combining co-occurrence mapping, topic trajectory analysis, and contextual embeddings could provide a holistic account of how meanings evolve in therapy. Another approach is to apply time-sensitive models, such as dynamic embeddings (Bamman et al., 2014; Bamblér & Mandt, 2017) or Generalized Additive Mixed Models (GAMMs), which enable the detection of nonlinear and session-specific changes.

As illustrated in Table 3, Liccione and Siciliano (2025) proposed a set of ten transversal themes common to all psychotherapies. We have referred to these topics as nuclear because, in some respects, they are discussed in therapy; what changes is the amount of time each patient spends on each topic, as well as the semantic content and emotional value each topic holds. These nuclear themes – such as Suffering/Pathology/Symptoms (SPS), Family (FAM), Affective Relationships (AR), Ways of Being (WB), Projectuality (PRJ), and Therapeutic Refiguration and Insight (TRI) – represent fundamental dimensions of the patient’s narrative and of the therapeutic process. The table also includes categories related to social relations, occupational or academic contexts, and macro-existential references.

This categorization aims to provide a clinically meaningful, cross-therapeutic framework to guide computational analyses. By relying on these shared themes, supervised systems can be trained to reliably detect and track their occurrence across large corpora of psychotherapy data. This approach ensures interpretability and continuity with clinical expertise. At the same time, unsupervised or generative models can remain valuable for uncovering case-specific or emergent themes that enrich, rather than replace, this core taxonomy. In this way, integrating expert-defined transversal categories with machine learning techniques enables more robust, clinically grounded monitoring of psychotherapeutic discourse.

Finally, future work should emphasize translational value. By embedding semantic analyses into feedback systems, clinicians could be provided with process markers that highlight subtle shifts in patient narratives – signals that might otherwise go unnoticed. Such tools would not replace clinical expertise but rather augment it, offering a bridge between computational rigor and therapeutic meaning-making. In sum, the study of semantic change in psychotherapy represents a fertile ground for interdisciplinary collaboration. By clarifying definitions, refining methodologies, and aligning computational findings with clinical realities, researchers can move toward a more integrated understanding of how language both reflects and facilitates psychological transformation.

N°	LABEL	THEME	DESCRIPTION
1	SPS	Suffering/Pathology/ Symptoms	Information and meanings concerning the patient's past and/or current suffering, symptomatology, psychopathology, existential distress.
2	FAM	Family	Information and meanings related to the patient's family, including relations with them.
3	AR	Affective Relationships	Information and meanings relating to past and current emotional relationships. All relationships with romantic or sexual involvement.
4	OS	Occupation/Study	Information regarding the patient's occupation, educational and/or academic background. And relations with colleagues and employments.
5	SR	Social Relations	Information and meanings concerning the patient's friendships and social network.
6	WB	Ways of Being	Information on personality traits, temperament, habitual behaviour and emotional tendencies that define how the patient perceives and interacts with the world.
7	PRJ	Projectuality	Information and significant about the patient's plans, future wishes and expectations.
8	TRI	Therapeutic Refiguration and Insights	The label encompasses both the therapist's references and reformulations concerning various themes addressed in therapy (ways of being in situations, values, and behaviours of oneself and others), as well as the patient's responses when these are congruent with the semantic message conveyed by the therapist's reformulations.
9	OTH	Other	Unspecified information and meanings (tended to be pleasantries, organization of the next session and greetings).
10	MEC	Macro Existential Context	Personal information about the patient and his/her social network (family, partner, colleagues, friends, etc.). Includes information on age, occupation, city where one lives, origins, etc.

Table 3. Nuclear Topics: themes relevant in psychotherapy with varying degrees of intensity and significance.

3.2.1. Toward a definition of *Semantic Shift*

The concept of *semantic shift* has increasingly attracted attention in psychotherapy research; yet it remains marked by definitional ambiguity. In linguistics, semantic shift is generally understood as the diachronic evolution of word meanings across historical periods and cultural contexts (Geeraerts, 2009). In computational linguistics, the notion has been adapted to capture changes in word usage or contextual embedding across corpora, often quantified through co-occurrence statistics or vector-space models (Hamilton, Leskovec & Jurafsky, 2016). Within psychotherapy, however, the term acquires a broader significance: it does not merely denote lexical variation but reflects a psychological and relational transformation, whereby clients reorganize meanings in the dialogical space of therapy (Salvatore et al., 2012; Angus & Kagan, 2013).

From a hermeneutic standpoint, semantic shift can be conceived as the reconfiguration of lived experience through narrative reconstruction. Ricoeur (1986) argued that personal identity is continually shaped by reinterpretation, as familiar symbols acquire new resonances and revised self-understandings emerge. In psychotherapy, such transformations often become visible when patients articulate past traumas or interpersonal conflicts using alternative metaphors, novel causal frameworks, or modified emotional tones. This process represents embodies a restructuring of the individual's relation to self and others (Gadamer, 1975; Stanghellini, 2010, 2016).

Empirical research has operationalized semantic shift through diverse methodologies. Co-occurrence analysis highlights reorganizations in patients' semantic networks by mapping how word clusters change across sessions (Salvatore et al., 2017). Topic modelling enables researchers to trace the emergence, decline, or persistence of thematic content, revealing how therapeutic concerns shift in salience throughout treatment (Atkins et al., 2012; Atzil-Slonim et al., 2021). More recently, embedding-based models have allowed the detection of fine-grained contextual shifts, capturing how the meaning of key terms evolves in dialogue with the therapist (Lalk et al., 2024; Periti & Tahmasebi, 2024). These approaches differ in scope and sensitivity, underscoring the importance of establishing a common conceptual framework.

For this work, a semantic shift in psychotherapy can be defined as: the transformation of meanings that patients and therapists attribute to words, topics, and narratives over the course of therapy, observable both hermeneutically – as the appropriation of new meanings through dialogue – and computationally – as measurable changes in linguistic usage, distribution, or contextual embedding. This definition emphasizes the multidimensional nature of the phenomenon. On the one hand, it acknowledges the subjective and relational aspects of meaning-making within the therapeutic encounter; on the other, it foregrounds the potential of computational tools to detect and quantify

these transformations. Crucially, semantic shift is not limited to the emergence of new topics but also encompasses modifications in how existing terms are framed, valued, and interconnected over time.

Conceptualizing semantic shift in this way creates a bridge between clinical observation and computational measurement. For researchers, it offers a coherent framework for testing whether linguistic transformations predict therapeutic outcomes or process variables. For clinicians, it draws attention to the evolving meanings embedded in patient narrative – signals that may provide insight into therapeutic progress beyond symptom reduction. Ultimately, semantic shift captures the essence of psychotherapy as a process of narrative and semantic reconfiguration, aligning linguistic evidence with the therapeutic aim of fostering new ways of understanding and inhabiting one’s life.

Method	What It Measures	Strengths	Limitations	Clinical Relevance
Co-occurrence and Semantic Mapping (e.g., ACASM; Salvatore et al., 2017)	Tracks how words cluster together and reorganize across sessions	Simple, transparent; highlights reorganizations of patients’ semantic networks	Sensitive mainly to frequent words; limited contextual nuance	Detects shifts in thematic associations (e.g., “failure” moving from <i>shame</i> to <i>learning</i>)
Topic Modelling and Trajectory Analysis (e.g., LDA, DTM; Atkins et al., 2012; Atzil-Slonim et al., 2021)	Identifies latent themes and tracks their salience over time	Captures macro-structure of therapy, interpretable themes	Treats words as fixed within topics; struggles with polysemy and contextual subtlety	Reveals whether therapeutic themes (e.g., <i>family</i> , <i>control</i>) gain or lose centrality
Embedding-based Models [e.g., BERT (Devlin et al., 2019), GPT (Radford et al. 2018), dynamic embeddings (Hamilton et al., 2016)].	Represents words in vector space, sensitive to contextual meaning	Captures fine-grained semantic drift; distinguishes meaning shifts within same word	Computationally intensive; requires large corpora; interpretation less transparent	Detects micro-level semantic changes (e.g., “control” shifting from <i>anxiety</i> to <i>autonomy</i>)

Table 4. Computational Approaches to Detecting Semantic Shift in Psychotherapy.

As shown in Table 4, different computational approaches capture distinct dimensions of semantic shift, from lexical clustering to contextual embeddings. Yet, what unites these perspectives is their ability to illuminate how therapy fosters the reconfiguration of meaning. In other words, semantic shift emerges as both a linguistic signature of therapeutic change and a process marker that bridges narrative theory with computational methods. This dual focus underscores psychotherapy as a practice where subjective meaning-making and objective linguistic transformations converge, allowing researchers to trace clinical progress while preserving its dialogical and experiential core.

3.3. Research Frontiers and Open Challenges

A key aim of this dissertation is to identify the “hot topics” and future lines of research that, in my view, emerge after an extensive review of the literature and through the trajectory of this doctoral project. Psychotherapy research has increasingly embraced computational approaches, yet much remains to be explored regarding how linguistic indicators can capture therapeutic change.

Among the many avenues of computational psychotherapy research, the movement of topics across sessions has emerged as a particularly fertile domain. Unlike static measures of word use, topic dynamics capture the evolving interplay of themes that structure therapeutic dialogue. Several studies have shown that topics are not fixed entities but processes unfolding over time: their frequency may increase or decline, their associations with other topics reorganize, and their lexical realizations can drift, reflecting underlying semantic transformations (Atkins et al., 2012; Imel et al., 2015; Tanana et al., 2016; Atzil-Slonim et al., 2021).

Within this framework, three dimensions of topic movement deserve special attention: occurrence, co-occurrence, and distance. Topic occurrence indicates the salience of themes over time, highlighting whether symptomatic narratives persist or give way to more adaptive frames (Reich et al., 2023). Topic co-occurrence offers insight into how themes cluster, indicating whether therapeutic work promotes integration across cognitive, emotional, and relational domains (Angus & Kagan, 2013). Topic distance, in turn, quantifies how far apart semantic fields remain in a patient’s discourse, offering a proxy for dissociation, avoidance, or conversely, the construction of bridges between previously disconnected aspects of experience (Tasca et al., 2020).

Together, these dimensions provide a multidimensional map of therapeutic discourse, enabling the linkage of computationally derived indices to clinically relevant processes (Lukianova-Holden & Cameron, 2021).

Future research should increasingly focus on harnessing these indices to provide clinically actionable insights. Tracking topic occurrence, co-occurrence, and distance in real time could provide therapists with structured feedback on whether sessions are progressing toward insight, integration, and emotional differentiation, or remain confined within maladaptive narratives. Retrospectively, analyses of topic dynamics may illuminate what distinguished successful from unsuccessful therapies, thereby serving as markers of therapeutic mechanisms. This dual application positions topic dynamics at the intersection of process-outcome research and clinical practice, aligning computational metrics with the therapeutic aim of fostering psychological transformation.

Another promising direction in psychotherapy research concerns the application of sentiment analysis to therapeutic dialogue. Recent advances in computational linguistics have shown that beyond thematic content, the affective tone expressed in language can be systematically quantified and tracked over the course of therapy (Tausczik & Pennebaker, 2010; Calvo & Mac Kim, 2013; Tanana et al., 2021). This approach has gained increasing attention because emotional valence, whether patients' narratives carry a predominantly positive, negative, or ambivalent tone. It serves as a meaningful indicator of therapeutic processes.

Many researchers are beginning to explore how sentiment metrics can be used as process indicators, highlighting critical junctures such as alliance ruptures, moments of therapeutic progress, or risks of premature dropout. Moreover, integrating sentiment trajectories with topic modelling offers a novel perspective: rather than treating topics as static, one can examine how their affective coloration evolves over the course of therapy (Tanana et al., 2021). For instance, discussions of "symptoms" may gradually shift from a predominantly negative sentiment to a more positive or hopeful tone, potentially reflecting processes of re-narration and meaning reorganization (Angus & Kagan, 2013; Atzil-Slonim et al., 2021; Tanana et al., 2021).

Finally, sentiment analysis is moving beyond simple polarity detection to capture fine-grained emotional differentiation, distinguishing, for example, between sadness, anger, hope, or relief. This granularity resonates with clinical theory, which suggests that the capacity to articulate and differentiate affective states is itself a marker of therapeutic progress (Greenberg, 2012). As such, analysis represents a growing frontier in computational psychotherapy research, with the potential to complement qualitative interpretation and provide scalable indicators of how clients' emotional landscapes evolve within the therapeutic process (Tanana et al., 2021).

The emergent frontier from topic modelling to generative AI systems exemplifies both extraordinary potential and significant caveats. Large Language Models (LLMs) offer novel capabilities (for instance, automatic labelling of therapeutic themes, semantic summarization of

transcripts, and detection of emerging or under-recognized topics) that could substantially broaden analytic scope in psychotherapy (Tian et al., 2023; Abdalla et al., 2024). For instance, recent applications of LLMs in mental health systems have shown promise in tasks such as developing rating scales to assess latent constructs (e.g., patient engagement) with strong psychometric properties (Xu et al., 2024). Other studies have demonstrated that LLMs may outperform traditional methods in semantic annotation tasks, including topic modelling evaluation and optimizing label assignment for text (Klein et al., 2024). Moreover, LLMs have been applied to detect emotional states in therapy transcripts, correlating detected emotions (e.g., approval, anger, fear) with symptom severity and alliance quality (Zhang et al., 2023).

The fluency and coherence of LLM-generated outputs can produce an “illusion of validity” (Kahneman, 2011); that is, models may appear convincing while failing to capture the nuanced meaning embedded in therapeutic dialogue. The scoping review of LLMs in mental health highlights systemic gaps in model evaluation, such as inconsistent metrics, reliance on proprietary prompt tuning, and limited transparency (Biswas et al., 2024). Furthermore, technological fixes like prompt engineering or synthetic adjustments (e.g., dataset balancing) risk introducing new biases or distortions if not grounded in domain knowledge and interpretive nuance (Liebrenz et al., 2023).

The research challenge, then, lies not just in leveraging LLMs but in critically evaluating whether they can reliably replicate or complement human coding in psychotherapy. Future avenues should include: (i) assessing LLMs’ accuracy in identifying topic trajectories compared with expert coding; (ii) exploring their ability to classify therapeutic approaches from transcripts (e.g., CBT, psychodynamic interventions); and (iii) employing generative capabilities to propose novel topics or patterns. These efforts must be accompanied by rigorous evaluation frameworks that ensure transparency and guard against overreliance, such as structured audit trails, fairness assessments, and the embedding of models within clinician-in-the-loop paradigms (Abdalla et al., 2024). In conclusion, the guiding principle remains: “to know to control.” LLMs should not supplant human expertise but serve as augmentative tools that expand analytic reach while anchoring innovations in clinical theory and safeguarding interpretability (Tian et al., 2023).

An interesting research frontier involves the extraction of personality traits directly from psychotherapy transcripts. Personality has long been shown to influence both linguistic style and decision-making tendencies. For instance, individuals high in extraversion or openness tend to use more positive affect words and broader semantic fields, whereas those high in neuroticism display more negative affectivity and self-referential language (Pennebaker & King, 1999; Mehl et al., 2006). More recent computational models confirm that traits shape the way individuals select topics and

linguistic features, supporting the feasibility of predicting personality from text (Schwartz et al., 2013; Park et al., 2015; Hu et al., 2017). From a decisional perspective, personality traits influence whether people rely on heuristic, fast judgments or more deliberative, reflective reasoning (McCrae & Costa, 1999; Kahneman, 2011). These tendencies affect not only everyday choices but also the therapeutic exchange: for example, patients higher in conscientiousness may display greater coherence and deliberation in discourse, while those higher in openness may explore alternative meanings more readily (Mairesse et al., 2007).

In computational modelling, approaches such as PT-LDA (Liu et al., 2016) and topic-personality frameworks (Hu et al., 2017) demonstrate how latent semantic structures can be linked to the Big Five dimensions, highlighting the bidirectional influence between personality and topic expression. Other techniques, such as tax2vec, enrich short texts with interpretable features, enabling more robust trait prediction in data-sparse conditions (Škrlić et al., 2021). Integrating these findings within psychotherapy research opens the possibility of automatically detecting how personality traits manifest in clinical dialogue, and how they shape not only affective tone (sentiment) but also decision-making patterns.

From our perspective, such decision-making mechanisms directly connect to the study of biases. If personality traits predispose patients or therapists toward specific reasoning shortcuts or affective framings, this will likely condition both the emergence and maintenance of cognitive biases in therapeutic discourse. Understanding this interplay is thus crucial before moving into a systematic analysis of biases in psychotherapy and in algorithmic models of language.

3.3.1. Bias in Psychotherapy and Bias in Algorithms

Bias in psychotherapy operates on multiple, interrelated levels: the therapist, the patient, the therapeutic setting, and increasingly the algorithmic systems used for analysis (Watson, 2025). Cognitive and affective biases are known to account for a large proportion of diagnostic and therapeutic errors in medicine. In psychotherapy, these biases emerge in subtler yet equally impactful ways. For example, therapists may anchor excessively on an early impression, while patients may selectively attend to confirming evidence of pre-existing beliefs, a dynamic particularly evident in paranoid ideation.

Kahneman (2011) proposed a dual-process model of cognition, distinguishing between System 1, which is fast, automatic, associative, and effortless, and System 2, which is slower, deliberate, and capable of logical reasoning. System 1 generates impressions and intuitions that are

often coherent and compelling, but also prone to systematic errors. By contrast, System 2 requires effort and attentional resources, and is therefore less frequently activated, earning the characterization of being “lazy” (Kahneman, 2011). This dual architecture has significant implications for psychotherapy. Therapists frequently rely on intuitive judgments when forming initial diagnostic impressions, often influenced by salient cues or the narratives from the first session. Without subsequent correction by System 2, such impressions may solidify into premature diagnoses or foster confirmation bias throughout treatment (Westen & Weinberger, 2004). Similarly, patients often employ intuitive reasoning in ways that reinforce maladaptive schemas. For example, individuals with paranoid ideation may interpret neutral behaviours as threatening, reflecting the dominance of System 1-driven inference without reflective evaluation (Garety & Freeman, 2013).

Empirical evidence suggests that both clinicians and patients are vulnerable to heuristic shortcuts. Yager, Kay, and Kelsay (2021) emphasized that psychotherapists’ intuitions are highly variable and frequently shaped by systematic distortions. Research in medical decision-making has also demonstrated that reliance on intuitive reasoning is a major source of diagnostic error (Croskerry, 2009), and similar mechanisms likely operate in psychotherapy. Dual-process models in clinical reasoning thus highlight the need for deliberate engagement of System 2 to counteract intuitive overreach (Evans & Stanovich, 2013).

In clinical practice, the balance between intuitive fluency and reflective correction is context dependent. Under time pressure, cognitive load, or heightened emotional salience, the likelihood that therapists default to System 1 increases, with corresponding risks of biased interpretation. Patients, too, may produce coherent but distorted narratives shaped by associative reasoning, making reflective processing more difficult (Kahneman, 2011; Wilson & Dunn, 2004). Understanding these dynamics provides a crucial framework for analysing not only therapeutic discourse but also the vulnerabilities of computational models that attempt to simulate human judgment.

A central insight from the heuristics-and-biases tradition is that intuitive reasoning, while efficient, is systematically prone to error. Kahneman and Tversky (1974) demonstrated that judgments under uncertainty are often guided by mental shortcuts – availability, representativeness, and anchoring – which provide rapid but potentially misleading answers. These heuristics are not confined to laboratory settings; they pervade clinical decision-making and therapeutic encounters.

Availability heuristic refers to the tendency to judge the likelihood of an event based on how easily examples come to mind. In psychotherapy, this may lead a clinician to overestimate the prevalence of trauma-related disorders if several recent patients presented with PTSD, even when epidemiological evidence does not support such frequency (Garb, 1998). Patients, likewise, may

retrieve emotionally salient experiences that dominate their narratives, such as recalling episodes of rejection while neglecting countervailing instances of acceptance. This selective recall reinforces maladaptive schemas and biases the therapeutic dialogue.

Representativeness heuristic involves evaluating similarity to a prototype rather than considering base rates. For example, a therapist might infer that a patient with obsessive traits must suffer from obsessive-compulsive disorder, overlooking the possibility of alternative explanations such as anxiety or perfectionistic personality style (Westen & Weinberger, 2004). Such prototypical reasoning can obscure nuanced clinical realities and result in premature closure. Patients, too, may apply representativeness when interpreting others' behaviour – viewing a friend's delay in replying to a message as evidence of hostility because it resembles earlier episodes of social exclusion.

Anchoring effects are particularly insidious in psychotherapy. An initial diagnostic hypothesis, once formed, often exerts disproportionate influence on subsequent clinical reasoning. Therapists may unconsciously seek confirmatory evidence while discounting disconfirming data, thereby perpetuating a biased formulation (Garb, 2005). Patients mirror this tendency: an individual convinced of being “unlovable” may anchor on early life experiences of neglect, interpreting current relationships through that fixed lens.

Confirmation bias further compounds these distortions. Both therapist and patient tend to search for evidence that validates their existing assumptions. In paranoid ideation, for instance, patients selectively attend to cues that support persecutory beliefs while ignoring contradictory signs (Garety & Freeman, 2013). Therapists, if unaware, may inadvertently collude with such biases by failing to challenge them explicitly.

These heuristics demonstrate that intuitive reasoning in psychotherapy is not merely a matter of efficiency but a source of potential error. The clinical anecdotes highlight how both therapists and patients employ shortcuts that can perpetuate maladaptive narratives or diagnostic inaccuracies. Systematic reflection and supervision are thus critical for identifying and counteracting such biases (Groopman, 2007).

Another recurring theme in cognitive psychology is the overconfidence bias, the systematic tendency to overestimate the accuracy of one's own judgments. Kahneman and Tversky (1979) showed that confidence levels often exceed actual predictive accuracy, particularly in uncertain domains. In psychotherapy, this illusion manifests when therapists trust their intuitive impressions despite limited evidence, a phenomenon sometimes referred to as the illusion of validity (Kahneman,

2011). Clinicians may feel assured that their diagnostic formulation is correct because it “fits” the patient’s presentation, even though statistical base rates or alternative hypotheses suggest otherwise.

This scepticism toward the reliability of clinical judgment has a long tradition in psychological science. Meehl (1954) famously demonstrated that statistical prediction consistently outperforms clinicians’ intuitive assessments, showing how confidence in clinical impressions often reflects subjective conviction rather than predictive accuracy. Later, Klein (1998), in his work on naturalistic decision-making, observed that experts tend to rely on rapid pattern recognition and familiar scripts. While these strategies can be adaptive in highly structured or repetitive domains such as the military or firefighting, psychotherapy operates in inherently ambiguous and idiosyncratic contexts, where reliance on intuitive coherence can translate into premature diagnoses and overly rigid narratives.

Clinical anecdotes illustrate this vividly. A therapist who has repeatedly treated depressive disorders may confidently (but erroneously) classify a patient’s withdrawal as depression, overlooking early signs of a bipolar spectrum condition. Similarly, in supervision, trainees often defend initial case conceptualizations with high certainty, even in the face of contradictory data – a form of anchoring reinforced by professional identity and the need for coherence (Garb, 2005).

The illusion of validity is compounded by the human preference for coherent stories. Patients and therapists alike construct narratives that “make sense,” smoothing inconsistencies in ways that reinforce initial beliefs. For example, a patient who believes in their chronic inadequacy may interpret sporadic successes as “exceptions,” thereby maintaining a biased self-schema. The therapist, if overly confident, may also accept this narrative rather than probing for disconfirmatory experiences. Such mutual reinforcement can create what Kahneman (2011) describes as the “confidence trap,” where subjective certainty replaces objective evidence.

Empirical studies further confirm these tendencies. Research on medical and psychological decision-making shows that even experienced clinicians are susceptible to overconfidence, leading to both false positives (overdiagnosing based on intuition) and false negatives (discounting conditions that fall outside the expected prototype) (Garb, 1998; Spengler et al., 2009). Overconfidence is particularly hazardous in psychotherapy because the therapeutic process itself is interpretative: misjudgements are not easily corrected by immediate feedback, unlike in more technical fields such as radiology or laboratory testing.

The implications are twofold. First, clinicians must cultivate metacognitive awareness, explicitly question their certainty, and invite supervision or peer consultation as corrective mechanisms (Royce, Hayes, & Schwartzstein, 2019). Second, structured tools such as checklists, rating scales, and

standardized coding systems can serve as counterbalances to intuitive overconfidence, ensuring that clinical impressions are tested against external anchors. In psychotherapy, where narratives are inherently complex, such measures are indispensable in avoiding the pitfalls of expert intuition.

One of the most influential contributions of behavioural economics to psychology is prospect theory, developed by Kahneman and Tversky (1979). The central insight is that individuals evaluate outcomes not in absolute terms, but relative to a reference point, with losses weighted more heavily than equivalent gains. This asymmetry produces systematic distortions in judgment, often referred to as the loss aversion principle. In psychotherapy, these dynamics manifest in both how patients assess their own progress and how therapists present interventions and feedback.

Framing effects occur when the presentation of equivalent information leads to different decisions. Kahneman and Tversky (1984) demonstrated that people tend to choose risk-averse options when outcomes are framed as gains but shift toward risk-seeking behaviour when outcomes are framed as losses. In clinical contexts, framing can significantly influence therapeutic engagement. For example, a therapist who frames an intervention as “an opportunity to enhance coping skills” may elicit more openness than if the same strategy is described as “a way to prevent further deterioration.” Similarly, patients evaluating their progress may perceive a 20% symptom reduction as encouraging when framed as improvement but discouraging when framed as “still having 80% of the symptoms.”

The clinical stakes of framing are evident in treatment adherence. Research in health psychology has shown that patients are more likely to follow medical advice when benefits are emphasized (gain framing) but are more likely to engage in screening or detection behaviours when the risks of inaction are highlighted (loss framing) (Gallagher & Updegraff, 2012). In psychotherapy, parallel effects emerge. Positively framed narratives foster therapeutic alliance and motivation, while negatively framed messages may mobilize urgency but also increase resistance or dropout.

For instance, a patient with social anxiety may respond differently depending on whether exposure exercises are described as “building resilience and confidence” versus “avoiding future relational breakdown.” The content of the task remains the same, but the motivational response shifts depending on whether the patient perceives the intervention as a potential gain or as a means of protection from loss. Likewise, therapists’ own decisions may be swayed by framing. Faced with a patient ambivalent about pharmacotherapy, a clinician emphasizing the chance to feel more in control” may reach a different recommendation than one highlighting “the risk of relapse if untreated.

The broader implications of prospect theory for psychotherapy lie in recognizing that both therapists and patients are boundedly rational agents, influenced by contextual cues and affective framing. The

therapeutic dialogue itself becomes a space where frames are negotiated and reframed. Effective therapy may thus involve deliberately shifting patients' reference points – for example, redefining setbacks not as failures but as temporary fluctuations in a broader trajectory of growth. This approach aligns with cognitive-behavioural techniques, specifically cognitive reframing, which explicitly targets biased evaluations and catastrophic thinking (Beck, 2011).

However, framing can also perpetuate biases if not critically examined. A therapist who consistently frames progress narrowly (e.g., symptom reduction alone) may reinforce a deficit-based identity, overlooking gains in relational functioning or self-efficacy. Conversely, overemphasis on positive frames may minimize genuine distress, risking invalidation. Awareness of these dynamics underscores the importance of metacognitive supervision: therapists must reflect not only on the content of what is said but also on the implicit frames shaping therapeutic choices.

Ultimately, framing effects and prospect theory highlight that decision-making in psychotherapy is not purely rational but shaped by perception, context, and narrative construction. Recognizing these mechanisms can enhance both clinical accuracy and empathy, while also reminding researchers to account for similar biases when interpreting therapeutic data or designing algorithmic models.

The classic debate between clinical judgment and statistical prediction has profound implications for the practice of psychotherapy. Paul Meehl's landmark book *Clinical versus Statistical Prediction* (1954) systematically reviewed 20 studies comparing the accuracy of expert clinicians with that of simple statistical formulas. His conclusion was striking in nearly 60% of the studies; algorithms significantly outperformed clinicians, despite their lack of "insight" or "experience." Subsequent meta-analyses confirmed this finding, showing that formulas – often based on a few weighted variables – tend to be more reliable and less expensive than expert judgment (Ægisdóttir et al., 2006).

One key reason for clinicians' relative inferiority is cognitive bias. As Meehl argued, experts often "try to be clever," constructing overly complex mental models that introduce noise and inconsistency. Therapists, for instance, may overemphasize subtle cues in sessions, weaving them into elaborate explanatory narratives that feel convincing but lack predictive validity. Research shows that when presented with identical patient information at two different times, clinicians frequently reach different conclusions – a phenomenon rooted in inconsistency bias (Garb, 1998). In contrast, formulas are perfectly reliable: given the same input, they always yield the same output. For instance, a therapist may intuitively predict that a patient who reports "feeling better" after two sessions will maintain improvement. However, long-term outcomes are far less predictable, and subjective impressions rarely outperform actuarial tools (Dawes, Faust, & Meehl, 1989). This illusion of validity

(Kahneman, 2011) reflects a broader overconfidence bias, where therapists mistakenly equate coherence with accuracy.

Bias also arises from the institutional and relational dynamics of therapy settings. Diagnostic momentum may occur when early hypotheses—formulated in intake sessions or supervision—are carried forward uncritically. Supervisory groups, too, may fall prey to groupthink, suppressing dissenting interpretations in favour of consensus (Janis, 1982). These biases echo Kahneman’s observations from his time in the Israeli Defense Forces. By replacing global, impressionistic interviews with structured ratings on predefined traits, the predictive accuracy of soldier performance improved substantially. Yet when intuition was permitted only after systematic scoring, its value increased, suggesting that structured methods can constrain bias without discarding clinical insight (Kahneman, 2011).

Bias not only affects clinical practice but also psychotherapy research. Investigator allegiance effects – the tendency for researchers to favour their own therapeutic orientation – can skew outcomes (Luborsky et al., 1999). Sampling biases, selective reporting, and confirmation bias further compromise the evidence base. The history of the Apgar score offers a sobering parallel: before its introduction in 1953, physicians relied on intuitive impressions of newborn vitality, often missing critical signs of distress. With Apgar’s simple five-item checklist, mortality rates dropped, demonstrating how standardization can reveal biases previously obscured by “clinical gestalt.” Similarly, structured outcome measures in psychotherapy may help to minimize allegiance-driven distortions and ensure comparability across studies.

Algorithms themselves are not free from bias. Although they avoid inconsistency and overconfidence, they can reproduce dataset biases (Barocas, Hardt, & Narayanan, 2019). Skewed training samples, subjective labelling, or distributional shifts between training and application contexts may lead to systematic misclassification. For example, an algorithm trained on transcripts from cognitive-behavioural therapy may underperform on psychodynamic sessions, not because of patient differences but because of linguistic and stylistic biases embedded in the data. Still, the evidence suggests that in low-validity environments – where prediction is inherently tricky, such as forecasting relapse or long-term therapy outcomes – formulas consistently outperform intuition (Meehl, 1954; Grove & Meehl, 1996). Robyn Dawes’ (1979) concept of the “robust beauty of improper linear models” further highlighted that even equal-weighted formulas often outperform experts, precisely because they are less vulnerable to overfitting idiosyncratic cues.

Many clinicians perceive formulas as “rigid, sterile, and blind,” while viewing intuition as “sensitive, genuine, and holistic” (Meehl, 1954). Yet, as Meehl argued, it may be unethical to rely on intuition

when algorithms demonstrably make fewer errors. This raises a fundamental challenge for psychotherapy: how to integrate the therapist's empathic, relational strengths with the statistical reliability of algorithms.

In contemporary practice, NLP models already support clinicians by analysing transcripts, detecting linguistic markers of distress, and suggesting possible interventions (Elhaddad & Hamam, 2024). Because these tools communicate in natural language, their outputs may be perceived as more dynamic and trustworthy, softening the “horror” clinicians once felt toward algorithmic decision-making. The task ahead is not to replace human intuition but to discipline it – embedding therapists’ insights within structured, algorithmically informed frameworks that maximize accuracy while preserving the humanity of the therapeutic encounter.

4. From Theory to Research: Methods and Tools to Investigate Changes in Psychotherapy

Situated at the core of this dissertation, this chapter consolidates and reports the complete results of the computational-linguistic analysis of psychotherapy transcripts, undertaken with the dual aim of advancing research methodology and innovating clinical practice in psychotherapy. It proceeds cumulatively.

- *The first study delineates recurrent topics and compares their session-level trajectories across two analytic paradigms – classical statistics (GAMMs) and Topic Modelling (LDA) – establishing observable changes in topic prevalence and inter-topic interactions.*
- *The second study compares LDA, BERTopic, and Top2Vec and shows that, beyond core domains, new case-specific topics emerge. This approach allows for the evaluation of how topics evolve across distinct phases of a completed psychotherapy.*
- *The third study models client affective trajectories across sessions, with therapist sentiment treated as neutral by design.*
- *The fourth study codifies the protocolized use of modern LLMs through rigorously engineered prompts to ensure transparent and reproducible analyses.*

Together, the studies trace a coherent path from detecting topic change to demonstrating semantic reconfiguration and operationalizing these insights in reproducible tools for research and clinical decision support. The subsequent section of the dissertation presents a software platform developed directly from the results reported in this chapter.

This chapter presents the results of an investigation into psychotherapies and AI tools designed for both research analysis and clinical utility. The empirical corpus originates from the manual transcription of a complete psychotherapy path (28 sessions) and its English translation via Reverse Translation (back-translation). As generative systems matured over the three years of this research, transcription and translation were substantially accelerated, enabling a greater emphasis on modelling Conversational change. Notably, improved technical tools for automated transcription and editing of damaged audio enabled the recovery of two recordings from two psychotherapy sessions used in this research. In fact, the first study's dataset comprises 26 sessions, while the second study's dataset comprises 28 sessions (see the sample paragraph in subchapter 4.1).

The program of work proceeded in two interconnected phases. First, a taxonomy of recurrent psychotherapy topics was established to delineate core domains and case-specific satellites. Second, algorithms were selected from the existing literature to extract topics and track their trajectories over time, under the assumption that therapeutic change can be operationalized as variation in topic prevalence, interaction, and semantic value. Methodologically, three tools were implemented:

- a) classical statistical approaches (Generalized Additive Mixed Models, GAMMs; and the LIWC branch, whose results are not reported in the present dissertation),
- b) topic-modelling approaches (LDA, BERTopic, Top2Vec), and
- c) generative systems (LLMs such as ChatGPT and Claude).

Quantitative indicators (e.g., GAMMs, a Topic Correlation Matrix in MATLAB, embeddings via RoBERTa, LLM-assisted measures) were complemented by qualitative inspections (lexical overlap, contextual re-use, and drift of salient terms across sessions) to preserve interpretability.

In subchapter 4.1, recurrent topics are first identified and compared in their session-wise trajectories across two analytic paradigms: Generative Additive Mixed Models (GAMMs) and the topic modelling algorithm Latent Dirichlet Allocation (LDA). Findings show that therapeutic change is detectible as systematic variation in topic prevalence and inter-topic interactions over time. A key methodological contribution is the adaptation of GAMMs to psychotherapy process data, enabling flexible, non-linear estimates of topic trends while controlling for session structure. Building on these dynamics, the second study, presented in subchapter 4.2, reframes change as a semantic shift. Topics extracted with LDA are contrasted with those from BERTopic and Top2Vec, revealing two complementary layers: stable cores (e.g., the principal nuclei defined in the methodological chapter, such as TRI and SPS) that organize the discourse throughout treatment, and emergent, case-specific satellites that BERTopic is particularly adept at surfacing. Crucially, the analysis indicates that

specific themes may persist across the therapeutic arc and reappear at restitution, whereas symptom-centred talk typically declines as treatment progresses. The third subchapter (4.3) presented results showing client emotion trajectories across sessions. The goal was to align affective dynamics with topic reconfiguration, testing whether emotional contours covary with the observed semantic shifts. The last subchapter (4.4) concludes by laying the groundwork for a systematic evaluation of modern LLMs. A set of prompt-engineering protocols—developed with technical and clinical input—seeks to make generative outputs transparent, reproducible, and clinically interpretable, turning prior insights (topic dynamics and semantic shifts) into operational tools for decision support.

In parallel with the academic work, these insights informed the development of a software platform that integrates AI tools for psychotherapy practice. The platform and its clinical modules are presented in Section 2 of the thesis, demonstrating how research findings translate into usable features for practitioners. For scope, the LIWC branch of classical statistics, used to detect recurrent linguistic patterns relevant to personality trait inference, is not detailed here. Based on comparisons with NLP methods, future work is outlined to develop an automatic system that estimates proportional personality traits from text, with the explicit caveat that all such outputs require therapist interpretation and consideration of extra-analytic factors.

4.1. The Usefulness of Combining Topic Modelling and Statistical Analysis to Investigate the Therapeutic Process: A Single Case Study

This study investigates the complex relationship between topic trajectories in psychotherapy and therapeutic outcomes, utilizing advanced statistical methods, including Generalized Additive Mixed Models (GAMMs) and Latent Dirichlet Allocation (LDA). The results provide clinicians with a solid framework for tracking patient progress, as shifts in topic prevalence are linked to improvements in therapy. This approach improves understanding of individual cases in psychotherapy and highlights the potential benefits of incorporating AI-driven analyses into therapy. Ultimately, it could enhance clinical practice.

4.1.1. Abstract

This study examines whether patterns in the movement of topics during psychotherapy sessions can provide psychotherapists with actionable insights for single-case analysis. It utilizes both statistical models and AI-driven tools to uncover these dynamics. We transcribed a completed psychotherapy session comprising 26 sessions. First, common topics across all therapies were identified, and then expert psychotherapists labelled each conversational turn of this selected

psychotherapy. As determined by the experts, the topic dynamics were analysed using Generalised Additive Mixed Models (GAMMs), which captured non-linear trends and hierarchical structures within the data. Subsequently, these trajectories, as identified by the experts, were compared with topics extracted unsupervised using the topic modelling algorithm Latent Dirichlet Allocation (LDA). Our findings confirm that topic trajectory analysis reliably indicates therapeutic progress. Specifically, topics related to suffering (SPS) decreased over time, while topics concerning therapeutic refiguration and insight (TRI) increased, reflecting clinical improvement. The study demonstrates that both GAMMs and LDA are helpful tools for understanding how topics in specific psychotherapy are modified in their occurrence during therapeutic work. Combining classical methods of statistical methods with AI-driven topic analysis enhances the sensitivity of assessments, providing insights into how psychotherapy work changes across sessions.

4.1.2. Methods

Our general hypothesis is that analysing topic trends during psychotherapy sessions can provide valuable insights into the therapeutic process (Del Giacco, Salcuni & Anguera, 2019; Hayes & Andrews, 2020; Imel et al., 2021). Combining topic analysis with statistical methodologies can enhance our understanding of the psychotherapeutic process and reveal insights into individual psychotherapies (Gaut et al., 2017). This approach enables us to compare the trajectories of topics selected by expert judges using GAMMs with those identified by a topic modeling algorithm. The second hypothesis is that only specific topics exhibit heightened sensitivity to detecting progress in the psychotherapeutic process. Specifically, we expect to observe a decrease in symptomatology-related topics throughout the psychotherapeutic sessions, in relation to the patient's different existential contexts. At the same time, we anticipate an increase in discussions of "therapeutic relationships" as the psychotherapeutic process progresses. Moreover, we expect that specific topics related to future projects and identity ways of being will emerge during therapy, using tools from the practice in which the psychotherapy was conducted. Additionally, we aim to gain insights into the psychotherapeutic process through LDA analysis of the word composition of different topics. For instance, the more topics share the exact words, the more closely the two themes are semantically related.

The clinical case examined in this research involves a patient, referred to as "I", who presented with symptoms of anxiety and depression, including frequent panic attacks that significantly affected her quality of life. Before the therapy began, the patient provided written consent to use video recordings for training purposes and research. All written material related to the patient was handled in compliance with privacy regulations, and any sensitive data was anonymized before the analysis

to protect the patient's identity. The patient received 18 months of psychotherapy, starting with weekly sessions and transitioning to biweekly and then monthly sessions. A total of 26 sessions were held, all of which were audio- and video-recorded. The therapy took place in a private practice setting and was conducted by a therapist using a hermeneutic phenomenological approach (Arciero et al., 2019; Liccione, 2019; Demaria, 2023).

The Outcome Questionnaire (OQ-45.2) was employed to assess the effectiveness of psychotherapy in this study (Lambert & Finch, 1999). This instrument comprises 45 items that assess various domains of functioning (e.g., symptoms, interpersonal issues, social role performance). This test was administered at two different times: before the start of the first session (T0) and at the end of the final session (T1). As detailed in Table 5, a significant change in the patient's self-reported scores is evident. At the end of the first psychotherapy session, the patient reported a score of 113 out of 180, which exceeds the cutoff score of 63, indicating clinically significant symptoms. After the twenty-sixth and final session, the OQ-45.2 was re-administered, resulting in a score of 44 out of 180. This score suggests that the patient no longer exhibits clinically significant symptoms.

	Symptom Distress (SD) Score	Interpersonal Relations (IR) Score	Social Role (SR) Score	Total Score
T0 – First session	54/100	37/44	22/36	113/180
T1 – Last session	20/100	13/44	11/36	44/180

Table 5. The scores of the three scales and the total score reported by the patient are extracted from the Outcome Questionnaire (OQ-45.2; Chiappelli et al., 2011, Italian version).

It's essential to advise the two of this recording (the session number 17 and the session number 24) are not used for the present analysis for three reasons: 1) the durations were significantly shorter (less than 15 minutes) and with few notable details about the therapy progress, mainly focusing on evaluating the ongoing therapy; b) the transaction was unreliable; 3) for the two reasons before the qualities of the trajectories could present irrelevant noisy. The decision was carefully evaluated.

We employed the descriptive phenomenological method (Giorgi, 1985) to identify the main themes, which we referred to as nuclear topics. These themes are broad and apply to all psychotherapy sessions. To analyse the transcript for this study, we followed the 3-step process (preparation, organization, reporting) suggested by Elo and Kyngäs (2008). This method involves preparing the

data and organizing it into meaningful units, the ones illustrated in Table 3. This approach allowed for a thorough analysis of the language used in therapy.

In the first step, we created a table listing the critical topics in psychotherapy (see Table 3). We considered these topics essential because they are applicable to any form of psychotherapy. Based on the literature, three expert judges compiled a table listing ten relevant topics in psychotherapy for understanding the patient's narrative. The researchers defined each topic in detail, as indicated in Table 2. In the second phase of the study, three experienced psychotherapists transcribed an entire psychotherapy session comprising 26 sessions (Mergenthaler & Stinson, 1992). These therapists were selected based on their experience of 10 years or more. Subsequently, the entire psychotherapy transcript was translated into English to facilitate the use of specific text-mining tools that rely on libraries trained on large English corpora. This translation enhanced the functionality of these tools and was carried out by a qualified professional with several years of experience. The translation was double-checked using the Reverse Translation method (Saldanha & O'Brien, 2014). The next step involved creating a dataset in an Excel file with several columns (refer to Figure 6). The first column contains the interview number, the second column indicates whether it was the therapist (T) or the patient (P) speaking, the third column contains the transcription, the fourth column contains the translation, the fifth column indicates the number of the conversational turn, and the remaining columns allow judges to enter topic labels for each conversational turn (refer to Table 3). Judges receive proper instructions on how to assign topics (as outlined in Table 3) after training and discussions with other experienced psychotherapists.

Session	Charachter	Phrase	Phrase	Conversational Turn	topic 1	topic 2
1	T	Buongiorno.	Good morning	1	OTH	
1	P	Buongiorno.	Good morning	2	OTH	
1	T	Prego, si accomodi pure.	Please, do sit down.	3	OTH	
1	P	Grazie.	Thank you	4	OTH	
1	T	Okay, se vuole può togliere la mascherina, tanto siamo a distanza di sicurezza. Allora, Lei si chiama?	Okay, if you want you can remove the mask, we are at a safe distance anyway. So, your name is?	5	OTH	MEC

Figure 6. An extract from the dataset was constructed to run the analysis of the present study.

Then we proceeded with both statistical and topic modelling analysis. The study used the dataset to implement a text analysis algorithm designed to extract Generalized Additive Mixed Models (GAMMs). These models were used to analyse topics labelled by expert judges during the fourth step of the study. For each topic listed in Table 3, a GAMM was extracted to predict the average trend over time, with time represented by the number of conversational turns across 26 therapeutic sessions.

Additionally, the algorithm was adapted to identify and visualize co-occurrence trends between pairs of topics, where co-occurrence is defined as two topics occurring in the same conversational turn. This analysis focused on topics such as Suffering/Pathology/Symptoms (SPS) and Therapeutic Refiguration and Insight (TRI) concerning various existential contexts, as well as topics related to the patient's projects (PRJ) and ways of being (WB). The complete text dataset was uploaded to MATLAB (MathWorks, 2023). The topic modelling algorithm involved segmenting the text dataset by time into three groups, which we referred to as phases. This segmentation improved the performance of LDA in extracting topics from our psychotherapy transcripts.

The entire English-translated transcript of the 26 therapy sessions was loaded into MATLAB (MathWorks, 2023). Translation into English was necessary because the MATLAB libraries used for topic extraction are trained on a large English vocabulary set and perform better with English text. Before applying the topic modelling algorithm, we performed standard text preprocessing to improve the quality of the analysis (Manning, Raghavan & Schütze, 2008). The transcript was tokenized, and words were then lemmatized to group different forms of the same word (e.g., “walking,” “walked,” and “walk” were all reduced to “walk”). We removed common stop words (e.g., conjunctions, adverbs, and fillers), punctuation, and numerical data such as ages and session dates. To refine the corpus, we excluded very short and very long words (fewer than 4 or more than 15 letters) as well as rare words that appeared fewer than two times. The list of removed words was reviewed and customized by expert evaluators to ensure relevance to the therapeutic context.

To apply the LDA model, it was necessary to determine the optimal number of topics (parameter K). This was done by testing different values and evaluating model performance using the Perplexity function, which measures how well the model predicts new data, lower values indicating a better fit. For our corpus, Perplexity suggested a plausible range between 5 and 10 topics (see Appendix 1). Because Perplexity alone does not guarantee that topics will be meaningful or interpretable, we also considered the semantic coherence of the results. An initial extraction of 10 topics revealed that two were difficult to label due to low internal consistency. Based on this evaluation, we selected 8 topics that offered a good balance between model performance and thematic clarity, enabling a reliable and meaningful analysis. Finally, we conducted a qualitative analysis. The results of the GAMMs, which reflect the judges' topic assignments, were compared with those of the LDA. The topics extracted by LDA were examined for changes over time using the word-overlap technique. Word overlap involves identifying the words shared between two topics and assumes that the strength of the relationship between them is indicated by the number of words they share. This qualitative evaluation was further validated using MATLAB's Topic Correlation function.

Three experienced psychotherapists collaborated on this study. They were responsible for assigning a label to each nuclear topic in every conversational turn of the entire psychotherapy and for identifying the topics extracted using LDA. Each judge has over ten years of experience as a psychotherapist. They worked on therapeutic transcripts while ensuring the privacy of patients and their social networks by either redacting or omitting personal data.

We applied a Generalized Additive Mixed Model (GAMM) (Zuur et al., 2009; Wood, 2017) to analyse the frequency trend of each topic over time. In our model, time (x-axis) is treated as the independent variable, and the dependent variable (y-axis) represents the probability (0 to 1) that a topic appears in each session (0 = not addressed; 1 = addressed). GAMMs were chosen because they allow for the modelling of non-linear patterns over time and can account for random effects, such as variability across sessions. This makes them especially suitable for single-case studies, where fluctuations within and between sessions are common. Compared to other models, such as polynomial regression, which requires manually setting curve degrees and may lead to overfitting, GAMMs offer a more flexible and interpretable approach to visualizing changes. They also go beyond basic spline models by managing session-specific and individual-level variation in a statistically robust way.

We used the Latent Dirichlet Allocation (LDA) algorithm to analyse the transcripts from the psychotherapy sessions (Blei et al., 2003; Rubin et al., 2011; Salvatore et al., 2012). LDA is an unsupervised method that identifies hidden (latent) topics in a set of texts, revealing patterns and themes without predefined labels (Atkins et al., 2012). In our model, each conversational turn was treated as a separate document (Atkins et al., 2012; Imel et al., 2015; Imel et al., 2021). LDA then associated each word with a topic and each topic with different parts of the conversation, allowing us to map the main themes discussed across sessions. The number of topics (K) was chosen by optimizing perplexity scores, and we selected eight topics for our final analysis. Each word was assigned to a topic based on its probability of belonging to that topic, helping us understand how the focus of conversations evolved over time (Blei et al., 2003; Blei & Lafferty, 2009). As shown in Figure 7, the LDA model treats each therapy session turn as a mixture of topics, assigning words to topics based on statistical patterns in the language used.

$$P(\theta, z, w \mid \alpha, \beta) = P(\theta \mid \alpha) \prod_{n=1}^N P(z_n \mid \theta) P(w_n \mid z_n, \beta)$$

Figure 7. Generative model of Latent Dirichlet Allocation (LDA). The formula describes how words in a document are generated. θ represents the topic distribution for a document; z_n is the topic assigned to the n -th word; w_n is the n -th word in the document; α and β are hyperparameters controlling the distribution over topics and words per topic, respectively (Blei et al., 2003).

from 0 to 1 (see Appendix 2). Given the subjective nature of topic assignment in psychotherapy, we assessed inter-rater reliability using Cohen's Kappa (Landis & Koch, 1977). Agreement levels were moderate: 0.50 (rater 1 vs. 2), 0.57 (rater 2 vs. 3), and 0.59 (rater 1 vs. 3) (see Appendix 3). These values are acceptable considering the complexity of the material and the nuanced interpretations involved. To address disagreements, each rater submitted an Excel file with their topic labels. A consolidated dataset was then created, retaining only those topics assigned by at least two out of three judges. This procedure ensured a systematic and balanced resolution of discrepancies, increasing the reliability and consistency of the final analysis.

The analysis using GAMMs enabled the identification and description of temporal trends in the topics discussed during therapeutic sessions. Table 6 summarises the primary statistical parameters for each model, including intercept estimates, standard errors, p-values, and 95% confidence intervals. For the Suffering/Pathology/Symptoms (SPS) topic, the estimated intercept was -1.1184 (95% CI: -1.36, -0.88), with a highly significant smooth term (edf = 18.02, Chi-squared = 225.2, $p < 2e-16$). This indicates an overall reduction in the prevalence of SPS over time, reflective of improvements in the patient's reported symptoms. The Family (FAM) topic showed an estimated intercept of -2.6230 (95% CI: -3.76, -1.49), with a highly significant smooth term (edf = 23.75, Chi-squared = 1490, $p < 2e-16$). These results suggest a complex yet considerable variation in the role of family dynamics during therapy. Conversely, for the Affective Relationships (AR) topic, the estimated intercept of -0.8290 (95% CI: -2.32, 0.66) was not statistically significant ($p = 0.275$), indicating that the frequency of this topic may have remained stable or was not strongly influenced by temporal progression. The Occupation/Study (OS) topic exhibited a significant intercept estimate of -2.4920 (95% CI: -4.18, -0.81, $p = 0.00375$), with a smooth term indicating significant non-linear variation over time (edf = 24.15, Chi-squared = 4522, $p < 2e-16$). This result suggests substantial evolution in themes related to the patient's academic or professional concerns. For the Social Relations (SR) topic, the intercept was estimated at -2.2953 (95% CI: -2.61, -1.98), with a highly significant smooth term (edf = 1.00, Chi-squared = 47.12, $p < 2e-16$). This suggests that social relationships were significantly affected by the progression of therapy. Other topics analysed – and listed in Table 3 –, including WB, PRJ, TRI, OTH, and MEC, displayed a mix of statistical significance and temporal variation. Visual representations of each topic trajectory in Appendix 4.

Topic	Intercept Estimate	Intercept Std. Error	Intercept p-value	Smooth Term edf	Smooth Term Chi.sq	Smooth Term p-value	R-sq. (adj)
SPS	-1.1184	0.1223	<2e-16	18.02	225.2	<2e-16	0.0541
FAM	-2.623	0.5794	5.98e-06	23.75	1490.0	<2e-16	-.222
AR	-.829	0.7588	.275	23.9	4364.0	<2e-16	-.772
OS	-2.492	0.86	0.00375	24.15	4522.0	<2e-16	-.667
SR	-2.2953	0.1611	<2e-16	1.0	47.12	<2e-16	0.0111
WB	-1.6425	0.1194	<2e-16	1.0	28.38	<2e-16	0.00625
PRJ	-3.0641	0.1758	<2e-16	1.0	89.28	<2e-16	0.0156
TRI	-.076	0.1078	1.81e-12	1.0	62.7	<2e-16	0.0135
OTH	-1.5727	0.0996	<2e-16	1.0	51	.821	-.000149
MEC	-3.6482	0.2385	<2e-16	7.914	118.7	<2e-16	.122

Table 6. Results of Generalised Additive Mixed Models (GAMMs) for individual therapeutic topics. The table includes intercept estimates, standard errors, and p-values, along with the significance of the smooth terms (edf, Chi-squared, and p-value). Adjusted R-squared values represent the explanatory power of the models. The 95% confidence intervals (CI) for the intercepts quantify the precision of the estimates. Significant smooth terms indicate non-linear temporal trends in the occurrence of individual topics across therapy sessions.

The analysis using GAMMs was extended to examine the temporal dynamics of topic co-occurrences during therapeutic sessions. Table 7 summarises the main statistical parameters for each model, including intercept estimates, standard errors, p-values, smooth term statistics, and 95% confidence intervals. For the SPS_FAM co-occurrence, the intercept estimate was -3.7529 (95% CI: -4.55, -2.96), with a highly significant smooth term (edf = 18.01, Chi-squared = 398.1, $p < 2e-16$). This indicates a significant non-linear relationship over time, suggesting that the co-occurrence of these topics decreases substantially as therapy progresses. The SPS_AR pair showed an intercept of -2.1536 (95% CI: -2.42, -1.89), with a smooth term that was also significant (edf = 1, Chi-squared = 36.58, $p < 2e-16$). These results reflect temporal variation in co-occurrence, albeit with a simpler pattern than in SPS_FAM. In contrast, the SPS_SR co-occurrence had a significant intercept of -4.0582 (95% CI: -4.46, -3.66), but the smooth term was less pronounced (edf = 2.393, Chi-squared = 7.186, $p = 0.0479$). This indicates a modest yet statistically significant temporal change in the co-occurrence of these topics. The SPS_WB co-occurrence exhibited an intercept of -2.8805 (95% CI: -3.14, -2.62), with a significant smooth term (edf = 1, Chi-squared = 11.58, $p = 0.000667$). These findings suggest that while the topics co-occur less frequently over time, the decline is relatively consistent. Notably, the SPS_PRJ co-occurrence did not show a significant smooth term ($p = 0.274$), suggesting no substantial temporal variation, despite a significant intercept estimate of -4.8742 (95%

CI: -5.28, -4.47). The TRI_FAM and TRI_AR co-occurrences showed highly significant nonlinear patterns (edf = 19.21 and 1, respectively; both $p < 2e-16$), suggesting substantial temporal dynamics in their co-occurrence. In contrast, TRI_OS exhibited no significant smooth term ($p = 0.456$), indicating temporal stability in their co-occurrence. Overall, these results reveal distinct patterns of co-occurrence across different topic pairs, with several showing significant nonlinear dynamics. SPS-related co-occurrences, such as SPS_FAM and SPS_WB, exhibit notable temporal changes, whereas others, such as TRI_OS, show temporal stability. These findings highlight the nuanced and context-dependent nature of topic co-occurrences within therapeutic sessions. Appendix 4 and Appendix 5 present trajectory plots for each Gam extracted from the dataset, for each topic, and for the co-occurrence between topics.

Pair of Topics	Intercept Estimate	Intercept Std. Error	Intercept p-value	Smooth Term edf	Smooth Term Chi.sq	Smooth Term p-value	R-sq. (adj)
SPS_FAM	-3.7529	0.4056	<2e-16	18.01	398.1	<2e-16	-0.0465
SPS_AR	-2.1536	0.1361	<2e-16	1.0	36.58	<2e-16	0.00287
SPS_OS	-3.4329	0.1799	<2e-16	1.002	69.6	<2e-16	0.0232
SPS_SR	-4.0582	0.2028	<2e-16	2.393	7.186	0.0479	0.00387
SPS_WB	-2.8805	0.1336	<2e-16	1.0	11.58	0.000667	0.00392
SPS_PRJ	-4.8742	0.2093	<2e-16	1.0	1.196	274	3.65e-05
SPS_TRI	-2.3694	0.1308	<2e-16	10.3	96.31	<2e-16	0.0179
TRI_FAM	-3.9664	0.4336	<2e-16	19.21	358.1	<2e-16	-0.0858
TRI_AR	-2.2669	0.1262	<2e-16	1.0	15.24	9.43e-05	0.00435
TRI_OS	-3.5767	0.2072	<2e-16	1.0	555	456	-0.000277
TRI_SR	-3.8446	0.2415	<2e-16	1.0	77.92	<2e-16	0.00879
TRI_WB	-2.574	0.1427	<2e-16	1.0	85.86	<2e-16	0.0186
TRI_PRJ	-3.6193	0.1784	<2e-16	1.0	81.63	<2e-16	0.0113

Table 7. Results of Generalized Additive Mixed Models (GAMMs) for the co-occurrence of therapeutic topics. The table presents the intercept estimates with their associated standard errors and p-values, as well as the significance of the smooth terms (edf, Chi-squared, and p-value). Adjusted R-squared values indicate the proportion of variance explained by each model. The 95% confidence intervals (CI) for the intercepts provide an additional measure of uncertainty. Topics with p-values less than 0.05 are highlighted as statistically significant co-occurrences, illustrating key interactions during therapy.

To apply the LDA model developed for our study, we needed only the English-language text and data on patient-therapist interactions. The algorithm was implemented in MATLAB (MathWorks, 2021). To account for changes in topics over time, we decided to divide the text corpus

into three phases: the initial part of the therapy (from the 1st to the 8th session), the middle phase (from the 9th to the 18th session), and the concluding psychotherapy sessions (from the 19th to the 26th session). The English dataset used for LDA implementation initially contained 219,934 words before preprocessing. After preprocessing, the number of words was reduced by approximately 70%. LDA was performed on a dataset containing around 65,980 words, as provided directly by the software. We developed an unsupervised script for the chosen algorithm, i.e., LDA.

Three judges individually labelled the 26 topics according to the labels in Table 3. Topics that received the highest level of agreement were analysed and discussed in this study using the word-overlap technique, a qualitative analysis tool, and the Topic Correlation function implemented in MATLAB. Both methods allowed the identification of shared words across different topics. The distance between two topics indicates how closely or distantly related they are within a given textual corpus. We identified a set of topics through consensus among a team of psychotherapists who reached a high level of agreement in their assessments (Imel et al., 2015; Gelo and Salvatore, 2016).

LDA – First Phase: From 1st to 8th Session. In the first phase of therapy, eight primary topics were identified and assigned specific labels by the three expert judges (Figure 9). Topic 1 was labelled Affective Relationships (AR), Topic 2 was assigned Family (FAM), Topic 3 was labelled Therapeutic Refiguration and Insight (TRI), Topic 4 was again assigned Family (FAM), Topic 5 was labelled Occupation/Study (OS), Topic 6 was labelled Suffering/Pathology/Symptoms (SPS), Topic 7 was assigned Social Relations (RS), and Topic 8 was again labelled Suffering/Pathology/Symptoms (SPS). The agreement between the raters during this phase was measured as follows: 0.52 between rater 1 and rater 2, 0.67 between rater 1 and rater 3, and 0.52 between rater 2 and rater 3. These values indicate a generally good level of agreement, as they fall within the 0.40 to 0.75 range, which is considered "good" according to standard inter-rater reliability guidelines.

Topic 1		Topic 2		Topic 3		Topic 4	
Word	Score	Word	Score	Word	Score	Word	Score
relationship	0,046412	study	0,04810	stefano	0,044818	evening	0,055577
stefano	0,04345	friend	0,043848	person	0,04500	friend	0,048678
aspect	0,032917	start	0,035013	point	0,044241	birthday	0,02990
moment	0,031929	happy	0,030432	understand	0,030893	sense	0,028747
point	0,030612	night	0,026505	question	0,030893	house	0,026447
people	0,030612	happen	0,019961	level	0,024028	girlfriend	0,02530
sense	0,02930	write	0,019306	relationship	0,023646	dinner	0,024531
understand	0,028308	problem	0,01538	emotional	0,022883	leave	0,021464
experience	0,023371	situation	0,01538	people	0,02060	yesterday	0,019548
great	0,018433	morning	0,014725	difficult	0,019451	morning	0,018015

Topic 5		Topic 6		Topic 7		Topic 8	
Word	Score	Word	Score	Word	Score	Word	Score
write	0,041806	person	0,031388	personal	0,031457	mother	0,06703
answer	0,038106	speech	0,026968	close	0,027042	father	0,044142
moment	0,031447	saturday	0,024757	second	0,023731	relationship	0,038147
badly	0,02590	situation	0,023873	person	0,022627	sister	0,037057
sense	0,025157	point	0,02122	valeria	0,022627	study	0,033788
point	0,024787	colleague	0,018568	month	0,02042	school	0,025068
understand	0,020348	behave	0,018568	aggression	0,02042	problem	0,023978
happen	0,01960	them	0,017242	return	0,019868	university	0,019619
period	0,01776	wrong	0,016357	share	0,018212	choice	0,019074
reason	0,01739	story	0,015915	rationally	0,016005	guilt	0,018529

Figure 9. The figure presents the top 10 most representative words for each of the eight topics extracted using the LDA model from the psychotherapy sessions. Each word is accompanied by its score, which ranges from 0 to 1, indicating the probability of the word being associated with a given topic. Higher scores reflect words that are more central to the topic. The topics were labelled by expert psychotherapists as follows: Topic 1: SPS; Topic 2: FAM; Topic 3: TRI; Topic 4: FAM; Topic 5: OS; Topic 6: SPS; Topic 7: RS; and Topic 8: RS. This figure illustrates the lexical components most influential in each topic, providing insight into the patient's evolving narrative.

In this phase, the overlap in word usage between these topics was substantial, particularly between Topic 6 (SPS) and Topic 1 (AR). For instance, words like "problem" and "issue" appeared frequently across both topics, indicating that the patient's suffering was closely tied to affective and romantic issues. The Topic Correlation values show a strong association between SPS and AR with a correlation value of 0.85, and between SPS and FAM (Topics 2 and 4) with a value of 0.78. This suggests that during the initial sessions, the patient's emotional distress was strongly connected to her family and romantic relationship dynamics.

LDA – Central Phase: From 9th to 18th session. As shown in Figure 10, the identified topics remained consistent during the second phase. Still, they were as follows: Topic 1 was labelled Affective Relationships (AR), Topic 2 was labelled Family (FAM), Topic 3 was again labelled Therapeutic Refiguration and Insight (TRI), Topic 4 was labelled Family (FAM), Topic 5 remained Occupation/Study (OS), Topic 6 was assigned Suffering/Pathology/Symptoms (SPS), Topic 7 was labelled Social Relations (SR), and Topic 8 was once again Suffering/Pathology/Symptoms (SPS).

The inter-rater agreement during this phase improved, with values of 0.85 between rater 1 and rater 2, 0.70 between rater 1 and rater 3, and 0.86 between rater 2 and rater 3. These high values indicate excellent agreement among the raters (scores above 0.75 are considered excellent).

Topic 1		Topic 2		Topic 3		Topic 4	
Word	Score	Word	Score	Word	Score	Word	Score
stefano	0,048102	sense	0,059723	write	0,041707	understand	0,080845
sense	0,031721	happen	0,048118	friend	0,036640	mother	0,046456
change	0,024441	house	0,033966	leave	0,029819	sister	0,044646
realize	0,024181	grandmother	0,031135	night	0,027480	friend	0,037104
period	0,021321	moment	0,029437	answer	0,026895	family	0,034691
relationship	0,019241	people	0,022361	study	0,024946	aspect	0,028356
today	0,018981	month	0,019530	spend	0,021049	accept	0,019005
strong	0,018201	worry	0,018964	dinner	0,018515	parent	0,014480
point	0,016901	situation	0,018398	evening	0,018515	wrong	0,013575
month	0,014821	albania	0,017549	friday	0,017930	point	0,012971

Topic 5		Topic 6		Topic 7		Topic 8	
Word	Score	Word	Score	Word	Score	Word	Score
start	0,067450	person	0,043058	valeria	0,066413	sister	0,043093
morning	0,038073	point	0,037717	people	0,044621	problem	0,034233
study	0,032668	problem	0,036382	thought	0,037012	point	0,030608
house	0,031492	understand	0,032377	arrive	0,032169	remember	0,027789
tomorrow	0,023267	emotional	0,029039	question	0,031131	continue	0,024567
Monday	0,023032	relationship	0,026369	story	0,029402	short	0,022151
practically	0,022562	level	0,021362	moment	0,024905	place	0,019734
honestly	0,022092	opinion	0,020361	grandmother	0,022484	mother	0,019332
month	0,021152	manage	0,018024	bring	0,021100	difficulty	0,016915
sleep	0,020917	clear	0,018024	sister	0,018333	marta	0,014901

Figure 10. The figure shows the top 10 representative words for each of the eight topics extracted using the LDA model during the middle phase of the psychotherapy sessions. Each word is accompanied by its score, which ranges from 0 to 1, indicating the probability that the word is associated with a given topic. Higher scores denote words that are more central to the topic. The topics were labelled by expert psychotherapists as follows: Topic 1: AR; Topic 2: FAM; Topic 3: TRI; Topic 4: FAM; Topic 5: OS; Topic 6: SPS; Topic 7: SR; and Topic 8: SPS.

During this phase, the word overlaps between Topic 6 (SPS) and Topic 1 (AR) decreased compared to earlier sessions. Words like "emotion" and "difficulty," which were initially shared, became less frequent, suggesting a shift in the patient's narrative away from relational stressors. The Topic Correlation values support this change, with the correlation between SPS and AR dropping to 0.62, and between SPS and FAM to 0.50. This reduction indicates a differentiation in how the patient began to separate emotional suffering from family and relational conflicts as therapy advanced.

LDA – Final sessions of psychotherapy: from the 18th to the 26th session. In the final phase of therapy (see Figure 11), Topic 1 was labelled Therapeutic Refiguration and Insight (TRI), Topic 2 was labelled Life Projects (PRJ), Topic 3 remained Suffering/Pathology/Symptoms (SPS), Topic 4 was labelled Social Relations (SR), Topic 5 was also Social Relations (SR), Topic 6 continued with Social Relations (SR), Topic 7 maintained the same label of Social Relations (SR), and Topic 8 was once again Occupation/Study (OS). The agreement between raters during this phase was as follows: 0.70

between rater 1 and rater 2, 0.41 between rater 1 and rater 3, and 0.69 between rater 2 and rater 3. Although the overall agreement remained good, some inconsistencies emerged, particularly between rater 1 and rater 3, as indicated by a value of 0.41, which is at the lower threshold of what is considered a "good" level of agreement. Overall, these results suggest a strong consensus among the raters in the classification of topics, with some variation, particularly in the final phase, where social relations became a more complex and prominent theme.

Topic 1		Topic 2		Topic 3		Topic 4	
Word	Score	Word	Score	Word	Score	Word	Score
person	0,071612	change	0,078998	happen	0,062188	friend	0,038893
understand	0,042710	manchester	0,039499	understand	0,052877	write	0,038206
experience	0,035967	place	0,036930	month	0,030928	evening	0,033631
stefano	0,032113	relationship	0,034361	problem	0,029598	night	0,033173
sense	0,030829	aspect	0,025691	thought	0,021616	dinner	0,032945
remember	0,025691	people	0,025369	totally	0,020951	spend	0,029970
emotional	0,025048	sense	0,025048	write	0,019954	weekend	0,023336
point	0,022158	opinion	0,024406	stuff	0,019621	leave	0,021505
story	0,019910	start	0,020874	decide	0,016961	start	0,021505
ilaria	0,017662	position	0,017341	birthday	0,015963	friday	0,018531

Topic 5		Topic 6		Topic 7		Topic 8	
Word	Score	Word	Score	Word	Score	Word	Score
person	0,037886	leave	0,055724	month	0,037554	friend	0,050718
people	0,033898	valeria	0,040618	point	0,031116	interview	0,035458
friend	0,023596	break	0,024841	study	0,027897	close	0,032765
tinder	0,018943	moment	0,024169	sister	0,024678	lawyer	0,024686
start	0,017946	yesterday	0,022155	start	0,020744	experience	0,024237
today	0,016949	realize	0,018127	drink	0,019671	practically	0,023340
stupid	0,016617	today	0,017456	lunch	0,017168	people	0,021993
opinion	0,015952	arrive	0,016113	stefano	0,016452	happen	0,020647
child	0,015952	happy	0,015442	period	0,015379	level	0,019749
associate	0,015952	morning	0,015442	remain	0,015022	recognize	0,019749

Figure 11. The figure presents the top 10 most representative words for each of the eight topics extracted using the LDA model from the final phase of the psychotherapy sessions. Each word is accompanied by its score, which ranges from 0 to 1, indicating the probability that the word is associated with a given topic. Higher scores indicate words that are more central to the topic. The topics were labelled by expert psychotherapists as follows: Topic 1: TRI, Topic 2: PRJ, Topic 3: SPS, Topic 4: SR, Topic 5: SR, Topic 6: SR, Topic 7: SR, and Topic 8: OS.

In this phase, the overlap of words between Topic 3 (SPS) and the other topics, particularly those related to affective and romantic relations (Topics 4-7), decreased significantly. Words such as "plan" and "future" associated with Topic 2 (PRJ) became distinct from terms like "anxiety" and "resolution," which were central to SPS. The Topic Correlation values confirm this reduction: the correlation between SPS and AR falls to 0.38, and between SPS and FAM nearly disappears, with a value of 0.15. These results suggest that the patient's emotional distress was more focused on personal growth and symptom resolution.

4.2. Reconsidering Topic Dynamics in Psychotherapy: A Critical Analysis with LDA and BERTopic

The mechanisms of therapeutic change and the patient's narrative transformation have often been examined through sequential analysis of discourse. Earlier studies have shown that transitions in topic sequences can indicate whether patients move from rigid, repetitive associations toward more flexible and adaptive discourse patterns (Milbrath, Bauknight, Horowitz, Amaro, & Sugahara, 1995). Yet, while such analyses are valuable in highlighting structural change, they remain limited in capturing the evolving semantic content of the topics themselves. Topic modelling extends this perspective by focusing on the reorganization of meaning within discourse. It allows us to observe that therapy does not merely alter the frequency of certain themes or their relational distance from other topics but also reshapes their semantic core. In other words, therapeutic change is visible not only in the emergence or recurrence of themes but also in their semantic reconfiguration. A topic such as “family” may initially be expressed in terms of conflict and burden, but later evolve into terms denoting support, reconciliation, or growth.

This concept of semantic shift is therefore central: the transformation of representative words within the same topic across different stages of therapy reflects not just change in therapy but change due to therapy. Mapping these shifts allows us to track how discourse reorganizes in response to therapeutic work, offering a more nuanced index of narrative transformation. On this basis, comparing computational models (e.g., LDA, BERTopic, Top2Vec) is crucial for evaluating their relative capacity to detect both the recurrence of nuclear topics and the semantic evolution of emergent, case-specific themes. First, we apply three distinct topic modelling algorithms (LDA, BERTopic, and Top2Vec) to the same psychotherapy dataset, directly comparing their performance in reproducing nuclear themes and detecting emergent ones. Next, our analysis revealed themes that could not be labelled within the set of nuclear categories defined by Liccione and Siciliano (2025) and reported in Table 3, underscoring the importance of computational methods in uncovering clinically relevant dimensions.

4.1.1. Abstract

Sequential discourse analyses can detect changes in topic transitions during psychotherapy (e.g., from rigid to flexible associations) but are less able to capture how the semantic content of topics reorganizes across treatment. Topic modelling complements this approach by tracing semantic shift within recurring thematic domains over time. We analysed a single completed cognitive–neuropsychological psychotherapy case (28 sessions), fully transcribed and anonymized, grouped

into four temporal blocks (7 sessions each). Transcripts were segmented into discourse turns and pre-processed in Python 3.10 (tokenization, lemmatization/stemming, customized Italian stopword removal, iterative noise filtering). Three topic-modelling approaches were applied to the same corpus to compare reproduction of nuclear themes and detection of emergent, case-specific topics: LDA (*gensim*; 5 topics/block, 8 top terms), BERTopic (transformer embeddings + UMAP + HDBSCAN + c-TF-IDF; max 8 topics/block, Topic -1 for outliers), and Top2Vec (joint embedding space; adaptive topic number; 15 words/topic cap). Outputs included topic-term rankings, topic summaries, word clouds (BERTopic/Top2Vec), and cross-block similarity routines. Results: Models differed in granularity and temporal sensitivity. LDA yielded a stable structure (5 topics per block; 20 block-specific topics total), supporting the identification of broad, recurrent (nuclear) themes. BERTopic produced variable topic counts: Block 1: 8 topics (incl. Topic -1; 7 substantive); Block 2: 3 (incl. -1); Block 3: 2 (no -1); Block 4: 8 (incl. -1). Top2Vec showed progressive consolidation (Block 1: 5; Block 2: 3; Block 3: 2; Block 4: 2). Across models, the pipeline also highlighted emergent themes not readily mapped onto nuclear categories (Liccione & Siciliano, 2025). Additional *stopword* refinement improved interpretability, especially for bag-of-words outputs. Overall, LDA was most effective for stable, broad themes, whereas BERTopic and Top2Vec were more sensitive to context-dependent and emergent structures, supporting the view that therapeutic change involves not only topic recurrence/transition but also semantic reconfiguration across therapy phases.

4.1.2. Methods

Sample

The present study is based on a single completed psychotherapy case consisting of 28 sessions conducted within a cognitive-neuropsychological framework. All sessions were audio-recorded with the patient's informed consent, fully transcribed, and subsequently anonymized to ensure data protection and compliance with ethical standards. Transcription followed established psycholinguistic criteria, preserving both lexical and paralinguistic features relevant for discourse analysis (e.g., pauses, interruptions). For analytic purposes, the therapeutic process was segmented at two levels. First, each transcript was divided into discourse blocks, corresponding to turns of talk or coherent semantic units. Second, the sessions were grouped into four temporal blocks (seven sessions each), allowing examination of topic trajectories over the longitudinal course of therapy. This segmentation enabled both intra-session analysis of topic structure and inter-session comparison of thematic evolution. The case was selected as a paradigmatic example of a treatment that reached a formal conclusion with evidence of clinical improvement. It thus provides a suitable ground to examine how nuclear topics – those consistently identified across psychotherapy research (e.g.,

symptoms, relational dynamics, therapeutic insight) – interact with emergent, case-specific themes, and how their relationships evolve. The dataset was anonymized, and consent for research use was obtained from the patients. It's essential to note that the present psychotherapy is the same as that used in the first study presented in this chapter (4.1), and that we included two sessions (the 17th and the 24th) that were excluded from the previous research for the reasons presented above. It was decided to use these two sessions in subsequent studies to divide the dataset into four equal blocks, even though they don't introduce any new topics.

Topic Model Tools

To capture the thematic structure of the psychotherapy corpus, three complementary topic modelling algorithms were implemented: Latent Dirichlet Allocation (LDA), BERTopic, and Top2Vec. Each of these models operates on distinct theoretical assumptions and computational mechanisms, thereby offering complementary perspectives on topic identification and evolution.

Latent Dirichlet Allocation (LDA; Blei et al., 2003) is a probabilistic generative model that assumes each document is composed of a mixture of latent topics, and a distribution of words characterizes each topic. Model training estimates two distributions simultaneously: the probability of topics given documents, and the likelihood of words given topics. This approach is well suited to identifying broad, recurring themes across the corpus, particularly those related to nuclear topics established in psychotherapy research. Figure 12 illustrates the functioning of the Latent Dirichlet Allocation (LDA) algorithm.

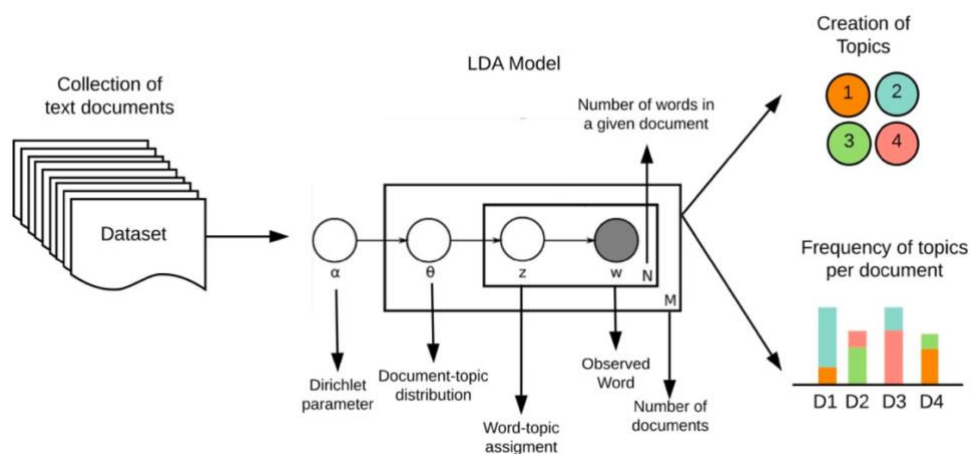


Figure 12. The Functioning of LDA as represented in Buenaño-Fernández et al., 2020. Latent Dirichlet Allocation (LDA) model. Each document (M), composed of a set of words (N), is modelled as a distribution over latent topic. The parameter α represents the Dirichlet prior controlling the distribution of topics per document, θ the document–topic distribution, z the word–topic assignment, and w the observed words. The output illustrates both the creation of latent topics and their frequency across documents.

The process begins by assigning a predefined number of topics to each document, after which every word is iteratively allocated to one of these topics until equilibrium between topics and words is reached. LDA assumes that each document (M), composed of several words (N), can be represented as a Dirichlet-distributed probability distribution over latent topics. In this framework, α denotes the Dirichlet prior that controls the distribution of topics per document, Z denotes the assignment of a word to a topic, and W denotes the observed word in document M . The generative process assumes that each word belongs to a single topic and that each document contains at least one topic. The choice of the parameter α is critical: higher values lead to a more uniform distribution of topics across documents, while lower values constrain the inference process to concentrate probabilities on fewer topics (Buenaño-Fernández et al., 2020).

BERTopic (Bidirectional Encoder Representations from Transformers; Devlin et al., 2019; Grootendorst, 2022) is a modern topic modelling algorithm that integrates contextualized embeddings with clustering and a modified weighting scheme, class-based TF-IDF (c-TF-IDF), to generate interpretable topic representations. The name BERTopic derives from the use of the pre-trained language model BERT (Bidirectional Encoder Representations from Transformers) to generate document embeddings, combined with the model's objective of extracting meaningful topics from documents through clustering and semantic analysis (Egger & Yu, 2022; Grootendorst, 2022).

Unlike standard TF-IDF, which evaluates word relevance at the level of individual documents, c-TF-IDF identifies the most distinctive terms for each cluster of documents, thereby highlighting the lexical items most representative of a given topic. Words with high c-TF-IDF scores are considered more important for defining the thematic content of a cluster. The pipeline consists of three main steps: (a) document embeddings are obtained from pre-trained transformer models (e.g., BERT; Devlin et al., 2019), which capture semantic relations and contextual meaning in the text; (b) dimensionality reduction is performed with UMAP (McInnes, Healy, & Melville, 2018), followed by density-based clustering using HDBSCAN (Campello, Moulavi, & Sander, 2013) to group semantically similar texts; (c) c-TF-IDF values are computed to extract representative terms for each cluster, ensuring interpretability of the resulting topics. Figure 13 illustrates the BERTopic workflow, which includes embedding, clustering, and topic representation using c-TF-IDF.

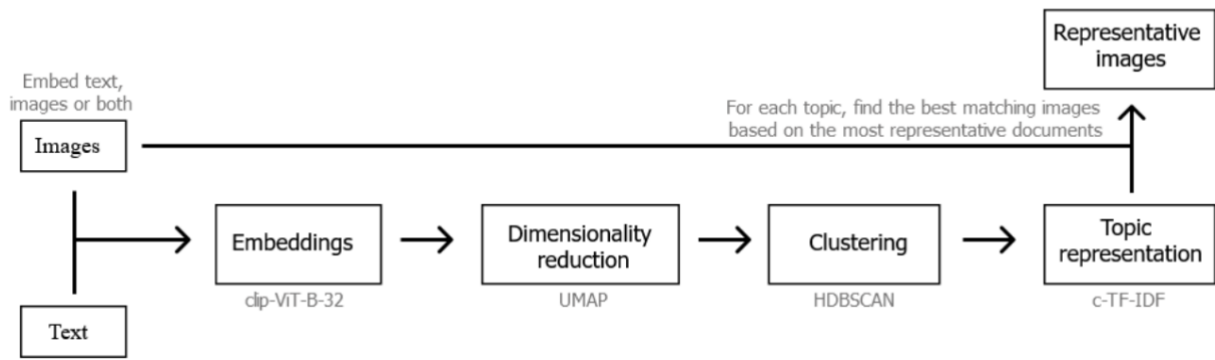


Figure 13. Schematic of the BERTopic workflow. Text or images are first transformed into contextualized embeddings (e.g., using CLIP-ViT-B-32). These embeddings are then reduced in dimensionality with UMAP and grouped through HDBSCAN clustering. Lastly, class-based TF-IDF (c-TF-IDF) is applied to identify the most representative terms for each cluster, resulting in interpretable topic representations. When images are included, the algorithm can also retrieve the most representative images for each topic based on the related documents.

Compared to traditional probabilistic models, such as LDA (Blei, Ng, & Jordan, 2003), BERTopic introduces several innovations. While LDA is based on bag-of-words representations that overlook semantic relationships among words, BERTopic uses transformer-based embeddings that create vector representations of words and sentences, capturing both meaning and context (Grootendorst, 2022). Additionally, BERTopic employs a continuous approach to topic modelling, allowing each document to be associated with multiple topics to varying degrees. In contrast, LDA uses a discrete approach, assigning documents exclusively to one or a limited set of topics (Alcoforado et al., 2022; Egger & Yu, 2022). This architecture enables BERTopic to generate dense, semantically coherent clusters, making it easier to interpret patient discourse while preserving keywords in topic descriptions. Beyond topic discovery, BERTopic incorporates other NLP functions, including sentiment analysis, named entity recognition, semantic role labelling, text classification, co-reference resolution, and next-sentence prediction, thereby broadening its use in psycholinguistic and clinical text analysis (Grootendorst, 2022). However, a major limitation is that, despite using embeddings, BERTopic still employs bag-of-words representations at certain stages, which can limit its ability to fully capture syntactic and semantic dependencies. Therefore, some topics may be lexically coherent but lack deeper contextual understanding.

Top2Vec (topic modelling via joint embeddings; Angelov, 2020) is an unsupervised algorithm for topic modelling and semantic search that simultaneously learns embeddings for documents, words, and topics within a shared vector space. In this space, the distance between vectors represents semantic similarity, such that documents with similar meanings are positioned close to one another

(Angelov, 2020). The core assumption of Top2Vec is that groups of semantically identical documents reflect the presence of an underlying topic. The algorithm proceeds in two stages: a) joint embeddings of documents and words are created, ensuring that both are represented in the same semantic space; b) dense clusters of documents are then identified, and the words most strongly associated with those clusters are used to define the topics. Unlike probabilistic models such as LDA (Blei, Ng, & Jordan, 2003) or hybrid models like BERTopic (Grootendorst, 2022), Top2Vec does not require the researcher to predefine the number of topics. Instead, topics are discovered adaptively, based on the density of semantic clusters in the embedding space. This makes Top2Vec particularly well-suited to detecting novel, emergent themes that extend beyond the nuclear set of psychotherapy topics, capturing unexpected or case-specific patterns in patient discourse.

Together, these three approaches provide complementary insights: LDA emphasizes stability and reproducibility of broad themes, BERTopic highlights contextual nuances, and Top2Vec facilitates the discovery of unexpected or emergent topics. Their joint application thus enables a more comprehensive understanding of how thematic dynamics unfold in psychotherapy discourse.

Hypotheses

Our hypotheses are related to the topic modelling algorithms comparison; we expected that: 1) LDA reliably reproduces broad nuclear topics but has limited capacity in detecting more nuanced or case-specific themes. 2) BERTopic is expected to be more effective in identifying context-dependent and specific satellite themes, due to its embedding-based representation and c-TF-IDF mechanism. 3) Top2Vec is expected to reveal new or unexpected themes by discovering clusters in the semantic embedding space without requiring a predefined number of topics. Overall, we expect to observe that one or more models will prove more suitable for detecting specific and emergent topics beyond the nuclear set, highlighting their relative strengths in psychotherapy process research.

Software, Prompt and Analytical Procedure

All analyses were conducted in Python (version 3.10) using a standardized pipeline developed for the study (part of the code is presented in Appendix 6). The pipeline includes three stages:

1) Pre-processing operations:

- ➔ Tokenization and sentence segmentation using *spaCy*.
- ➔ *Stopword* removal (NLTK and customized stopwords lists adapted for Italian psychotherapy transcripts).
- ➔ Lemmatization to reduce inflected forms to their canonical representation.
- ➔ Construction of the Document–Term Matrix (DTM) as input for statistical modelling.

2) Statistical Models

As we said below, three topic modelling approaches were applied to the same corpus to compare their performance and to examine their ability to reproduce both nuclear and emergent themes:

LDA: implemented via the *gensim library*, with hyperparameters (number of topics, alpha, beta priors) optimized through coherence scores. LDA was chosen for its robustness in detecting broad, recurrent themes that correspond to the nuclear topics identified in psychotherapy research.

BERTopic: applied using the BERTopic Python package, combining sentence embeddings (from transformer-based models) with c-TF-IDF vectorization and hierarchical clustering. BERTopic was expected to identify finer-grained contextual nuances and capture topic satellites related to specific patient concerns.

Top2Vec: implemented via the Top2Vec library, which jointly learns semantic embeddings and clusters documents without requiring a predefined number of topics. This model was particularly well suited to detecting novel, emergent topics beyond nuclear ones.

For each model, topic labelling was guided by a structured prompting strategy (see Appendix 6), ensuring consistency in the generation of descriptive summaries and clinical interpretation.

3) Output Integration

All outputs (topic distributions, word clouds, similarity measures) were automatically stored, indexed by session and block, and later integrated into the analysis of topic trajectories. The workflow also included visualization routines for:

- Word clouds of top terms per topic.
- Similarity indices of words of topics across the four therapy blocks.
- Heat map of BERTopic with similarity matrix between topics.

As mentioned earlier, all analyses in the present study were conducted using Python. This choice represents a methodological difference from the previous work presented in Chapter 4.1, which used MATLAB for topic extraction. Python was preferred for its open-source nature, which allows unrestricted use and distribution, and for its flexibility in integrating multiple libraries for natural language processing and statistical modelling. The language is widely supported by an active developer community, ensuring constant updates and improvements. Python is a high-level, multi-paradigm programming language that combines features from procedural, object-oriented, and

functional paradigms. It is particularly well-suited for scientific applications due to its extensive library ecosystem. In addition to the standard library, which includes over 200 modules, the Python Package Index (PyPI) provides thousands of additional packages, including state-of-the-art tools for machine learning, natural language processing, and data visualization. This extensive environment made Python the most appropriate choice for the present analyses.

4.1.3. Results

In this section, all results extracted from the 28 using Python are presented. Text transcriptions of the entire psychotherapy (see the sample paragraph for more information about the patient). Before running the three topic modelling algorithms, a systematic data cleaning and preprocessing phase was conducted. The first step involved defining a set of *stopwords*, including conjunctions, pronouns, articles, adverbs, and fillers, that are considered semantically uninformative in natural language. In this study, several words and most punctuation marks were manually specified as *stopwords* (e.g., ah, eh, yeah, yes, hum, nt, ., , , ?, !). Additionally, drawing on previous work, a supplementary list of stop words was compiled and stored in a dedicated .csv file.

The second step involved extracting conversational turns from each therapy session using the custom function *extract_conversational_turns_from_session*. These turns were further cleaned by removing superfluous spaces, discarding words shorter than three characters, and converting all uppercase letters into lowercase. Following this stage, the transcripts were segmented into four temporal blocks, each comprising seven sessions. Subsequently, the preprocessing function *preprocess_text* was applied to remove additional natural language elements that could interfere with the analyses. This function implemented several operations: 1) Tokenization: dividing the text into minimal units (tokens) for subsequent numerical representation. 2) Stemming: reducing inflected word forms to a standard stem. The stem does not necessarily correspond to a valid morphological root, but groups related forms under the exact representation (e.g., andare, andrai, andrò → and). 3) *Stopword* removal, performed via the function *no_sw_tokens*. All processed tokens were concatenated to reconstruct the cleaned sentences, which served as input to the subsequent topic modelling procedures. Following preprocessing, the cleaned dataset was saved and subsequently used to construct a Document–Term Matrix (DTM). In this matrix, each row represents a document, each column corresponds to a unique word, and each cell contains the word's frequency in that document. Once the DTM was obtained, the three topic modelling algorithms were applied separately.

It is important to note that, despite extensive preprocessing, the models initially produced topics containing residual noise (e.g., ve, ll, bhe, lot, boh). Some of these artifacts derived from common English contractions (have → 've, will → 'll, a lot → lot), while others represented idiosyncratic

fillers. Following this observation, the dataset was further refined by adding these tokens to the stopword list. This additional cleaning step resulted in substantially cleaner bag-of-words outputs across all three models.

LDA Results

For Latent Dirichlet Allocation (LDA), the model was trained on each therapy block using the DTM as input. Based on prior literature and the relatively small lexical sample, the number of topics was set to five per block, each composed of eight representative words.

The results of applying LDA to the dataset were characterized by the extraction of five topics, each composed of eight words representing the most salient terms within the respective topic. For all LDA outputs, a word cloud was deliberately not generated; instead, the codes as they appear within Python were displayed. The rationale for this choice is that previous studies have shown that LDA results can be visualized as word clouds. Consequently, it was considered more valuable to present the code as it appears in the new programming language selected for this work. Furthermore, for each extracted bag of words, the score of each term was calculated, thereby allowing the data to be evaluated not only qualitatively but also quantitatively.

Figure 14 shows the Bag of Words for the first block, from the first to the seventh session, as it is run through Python. The topic count starts at zero, rather than one.

```
Group: 0
Topic 0: friends mom kind understand mother father sister university
Topic 1: time work things stefano day relationship feel told
Topic 2: point view moment find kind weeks evening law
Topic 3: fact person good people sense - started live
Topic 4: sense going year night call opinion years person
.....
```

Figure 14. The five topics that were extracted from the first block of the psychotherapy case were examined. This image was extracted from the Python code.

In Figure 15, the five topics mentioned above are visualized in tables with two columns, one for the words and one for the associated scores. This score represents the probability that this word is representative of those specific topics. The same word could appear in more than one topic extracted by the LDA algorithm, but with different weights for each topic.

Topic 0		Topic 1		topic 2	
Word	Score	Word	Score	Word	Score
friends	0.029480	time	0.040235	point	0.044758
mom	0.023199	work	0.034299	view	0.025759
kind	0.021266	things	0.034079	moment	0.021114
understand	0.019333	stefano	0.023966	find	0.021114
mother	0.017400	day	0.020888	kind	0.019426
father	0.015467	relationship	0.020229	weeks	0.016892
sister	0.013535	feel	0.020009	evening	0.013937
university	0.011119	told	0.016491	law	0.013515

Topic 3		Topic 4	
Word	Score	Word	Score
fact	0.037688	sense	0.035162
person	0.035725	going	0.023039
good	0.020416	year	0.021826
people	0.018453	night	0.021422
sense	0.016883	call	0.018189
-	0.015313	opinion	0.016977
started	0.014528	years	0.014956
live	0.014528	person	0.014552

Figure 15. The five topics found to be representative of the first block analysed in the present study. The topics are numbered from zero to four. For each topic, it was decided that eight words would be representative. This is a way to visualize each topic along with its score.

For the second block, which spanned from the 8th to the 14th, the same analysis was performed. Five topics, numbered from 0 to 4, were extracted, with each topic represented by its eight most representative words, and the respective scores of each word are shown (see Figure 16 and 17).

```

Group: 1
Topic 0: sense fact told going talk time things point
Topic 1: time people happy years guy phone parents kind
Topic 2: kind things point days day view happened wanted
Topic 3: exam good day house going night week weeks
Topic 4: sister mom friends feel work - family boyfriend
.....

```

Figure 16. The five topics that were extracted from the second block of the psychotherapy case were examined. This image was extracted from the Python code.

Topic 0		topic 1		topic 2	
Word	Score	Word	Score	Word	Score
sense	0.026812	time	0.028680	kind	0.027144
fact	0.026005	people	0.019391	things	0.024505
told	0.024393	happy	0.016160	point	0.023751
going	0.022377	years	0.015352	days	0.022620
talk	0.021167	guy	0.014140	day	0.022243
time	0.020160	phone	0.013736	view	0.016966
things	0.019152	parents	0.013332	happened	0.014328
point	0.017338	kind	0.012929	wanted	0.013197

Topic 3		Topic 4	
Word	Score	Word	Score
exam	0.034061	sister	0.038667
good	0.027357	mom	0.036808
day	0.023870	friends	0.024540
house	0.020920	feel	0.020822
going	0.020652	work	0.020079
night	0.019311	-	0.019707
week	0.016898	family	0.017848
weeks	0.016898	boyfriend	0.013759

Figure 17. The five topics that are found representative of the second block analysed in the present study. The topics are numbered from zero to four. For each topic, it was decided that eight words would be representative. This is a way to visualize each topic along with its score.

In the third block, from the 15th to the 21st, we extract the same number of topics as were established for each phase of the therapy. In Figures 18 and 19, the extracted topics (Python code) and the visualization of scores for the eighth most representative word for each topic are shown.

```

Group: 2
Topic 0: things feel point person understand time friends people
Topic 1: work things years good dinner year change exam
Topic 2: hours exam times day hour fact long month
Topic 3: sense talk fact study going stuff kind opinion
Topic 4: time day mom told house going sister -
.....

```

Figure 18. The five topics that were extracted from the third block of the psychotherapy case were examined. This image was extracted from the Python code.

Topic 0		topic 1		topic 2	
Word	Score	Word	Score	Word	Score
things	0.031552	work	0.035526	hours	0.020481
feel	0.030039	things	0.024899	exam	0.016064
point	0.027230	years	0.020345	times	0.015261
person	0.025501	good	0.018523	day	0.014458
understand	0.020099	dinner	0.015184	hour	0.014458
time	0.019018	year	0.015184	fact	0.014057
friends	0.016857	change	0.014880	long	0.013655
people	0.015561	exam	0.014880	month	0.013254

Topic 3		Topic 4	
Word	Score	Word	Score
sense	0.034514	time	0.030015
talk	0.030736	day	0.029721
fact	0.023178	mom	0.028249
study	0.016125	told	0.023836
going	0.015873	house	0.022070
stuff	0.015369	going	0.020894
kind	0.015117	sister	0.019128
opinion	0.013354	-	0.015303

Figure 19. The five topics found to be representative of the third block analysed in the present study. The topics are numbered from zero to four. For each topic, it was decided that eight words would be representative. This is a way to visualize each topic along with its score.

Finally, the same analysis using the same parameters was run for the third block, which covered the last phase of therapy, from the 22nd to the 28th. In Figure 20, each extracted topic is displayed as it was in Python. In Figure 21, the five topics are presented in a table with two columns: one for the words and one for their associated scores.

```

Group: 3
Topic 0: talk understand night friends phone today change shit
Topic 1: friend stefano mum dinner kind birthday relationship wanted
Topic 2: good edimburgo met experience bad meet call week
Topic 3: sense fact opinion values person father stuff wrote
Topic 4: things going time day feel people point told

```

Figure 20. The five topics that were extracted from the fourth block of the psychotherapy case were examined. This image was extracted from the Python code.

Topic 0		topic 1		topic 2	
Word	Score	Word	Score	Word	Score
talk	0.023628	friend	0.022010	good	0.047572
understand	0.022476	stefano	0.019495	edimburgo	0.018619
night	0.021900	mum	0.015723	met	0.017930
friends	0.018443	dinner	0.015094	experience	0.016551
phone	0.016714	kind	0.014466	bad	0.016551
today	0.016714	birthday	0.013837	meet	0.016551
change	0.016138	relationship	0.013208	call	0.012415
shit	0.015562	wanted	0.013208	week	0.012415

Topic 3		Topic 4	
Word	Score	Word	Score
sense	0.042816	things	0.043374
fact	0.031944	going	0.039391
opinion	0.021751	time	0.037621
values	0.020392	day	0.035408
person	0.019713	feel	0.027885
father	0.019713	people	0.023017
stuff	0.015636	point	0.020362
wrote	0.014956	told	0.019034

Figure 21. The five topics that are representative of the fourth block were analysed in the present study. The topics are numbered from zero to four. For each topic, it was decided that eight words would be representative. This is a way to visualize each topic along with its score.

Top2Vec Results

For Top2Vec, the *preprocessed* conversational turns for each block were provided directly to the model. The outputs were then stored in a dictionary mapping each block to its corresponding model. Word clouds were generated for each block using the function *generate_topic_wordcloud*, providing a visual representation of the terms most strongly associated with each topic.

The data analysis performed with Top2Vec and the corresponding outputs were presented differently from those obtained with LDA. In this case, in addition to printing the similarity indices between the vectors of each word, word clouds were generated to facilitate a more immediate understanding of which terms carried greater weight within a given topic. As shown in the figures below, the larger words have higher similarity indices than the smaller ones. Regarding the number of topics extracted for each group, the model was allowed to determine the most appropriate number independently. However, a manual limit of 15 words per topic was imposed. For this reason, the number of topics varied across the groups.

All the topics extracted using Top2Vec are presented in the following figures (from 21 to 24), and the corresponding word clouds are included in Appendix 7.

Topic 0		Topic 1		Topic 2		Topic 3		Topic 4	
Word	Score	Word	Score	Word	Score	Word	Score	Word	Score
view	0.106552	told	0.134255	point	0.133884	understand	0.094797	fact	0.092558
week	0.073347	going	0.120176	time	0.126264	year	0.090714	understand	0.090349
night	0.072688	week	0.096659	fact	0.089069	stefano	0.086583	year	0.086078
understand	0.055686	person	0.069057	talk	0.079394	work	0.071057	feel	0.083399
point	0.051735	point	0.046009	sense	0.076216	study	0.064057	told	0.081894
fact	0.044169	talk	0.037651	work	0.066509	point	0.058445	going	0.078219
good	0.036095	working	0.031642	people	0.066496	sense	0.057486	day	0.074895
days	0.034537	exam	0.029841	find	0.043784	things	0.053766	people	0.070074
exam	0.034268	years	0.024745	stefano	0.039775	life	0.043079	friends	0.066724
person	0.033418	time	0.022159	feel	0.037624	week	0.035497	stefano	0.063601
hours	0.025715	year	0.017090	relationship	0.028351	working	0.032988	week	0.021971
find	0.022511	kind	0.011741	year	0.026910	relationship	0.026669	hours	0.015901
feel	0.019935	study	0.009381	exam	0.002701	years	0.025611	time	0.015326
sense	0.018276	fact	0.007239	going	0.002179	people	0.025237	moment	0.013563
told	0.017289	days	0.003025						

Figure 21. Five topics were extracted from the first block of the psychotherapy.

Topic 0		Topic 1		Topic 2	
Word	Score	Word	Score	Word	Score
vacation	0.143559	sister	0.131733	told	0.115503
life	0.110290	fact	0.124359	life	0.088859
started	0.102302	days	0.115520	morning	0.096134
moment	0.101099	marta	0.110457	friend	0.094547
wanted	0.094153	point	0.068337	months	0.082141
situation	0.074712	left	0.067996	started	0.080036
morning	0.073072	grandmother	0.050697	week	0.073066
exam	0.072786	weeks	0.050402	grandmother	0.072203
night	0.072676	sense	0.049699	study	0.065938
weeks	0.071647	going	0.049425	feel	0.063583
friends	0.069784	telling	0.047146	people	0.062072
good	0.067018	mother	0.042498	sister	0.056556
telling	0.066106	talk	0.032714	point	0.050130
understand	0.060806	started	0.024108	mother	0.049062
days	0.060214	things	0.023856	exam	0.048461

Figure 22. Three topics were extracted from the second block of the psychotherapy.

Topic 0		Topic 1	
Word	Score	Word	Score
sense	0.220285	year	0.124171
moment	0.134614	people	0.121886
weeks	0.084637	hours	0.106128
week	0.068271	wanted	0.078591
wanted	0.066002	point	0.078212
find	0.063427	night	0.078108
fact	0.050713	kind	0.068508
night	0.047111	days	0.047899
work	0.046860	life	0.045053
study	0.044370	things	0.028051
exam	0.037446	person	0.025443
things	0.034768	time	0.024088
understand	0.032575	talking	0.018557
feel	0.029304	stuff	0.016931
years	0.026612	day	0.016229

Figure 23. Two topics were extracted from the third block of the psychotherapy.

Topic 0		Topic 1	
Word	Score	Word	Score
good	0.080470	day	0.088152
people	0.060958	sense	0.040117
sense	0.050768	good	0.023768
day	0.041822	things	0.018701
things	0.035109	going	0.011723
time	0.018197	time	0.001959
person	-0.024397	person	-0.014332
going	-0.029595	people	-0.023870
feel	-0.076933	feel	-0.055356

Figure 24. Two topics were extracted from the fourth block of the psychotherapy.

BERTopic Results

In the case of BERTopic, preparatory functions were first defined to generate word cloud visualizations. The model was then applied to each therapy block using the *fit_transform* function, which embeds the conversational turns into a high-dimensional semantic space and performs clustering. This procedure generated a dictionary that linked each block to its corresponding topic model. Output summaries were obtained through the *topic_info* function. Consistent with the literature and with the lexical characteristics of the dataset, the maximum number of extracted topics was set to eight. BERTopic, by design, assigns all unclustered terms to topic -1, which functions as a “catch-all” category for words insufficiently close to other topics.

The results extracted through BERTopic were presented in the form of a word cloud, as was done for Top2Vec, so that the contribution of each word to the topic could be immediately recognized.

Additionally, in this case, the similarity scores for each word were reported, allowing a quantitative assessment of each word's representativeness within its topic. Unlike the two previous models, BERTopic specified a maximum number of topics the algorithm could identify within each group of sessions; based on the literature and the dataset size, this maximum was set to eight topics per group. Consequently, the number of topics varies across groups. The number of words characterizing each topic, on the other hand, was kept constant across all groups; in this case, each topic included ten words, arranged in descending order of representativeness. Moreover, as with Top2Vec, the colour of the words in the word cloud was not assigned according to a specific criterion but was randomly determined by the model, while the larger font size corresponded to words that were more representative of the underlying documents.

It is essential to note that BERTopic, unlike the previously described models, categorizes marginal words into a separate cluster called Topic -1, which comprises words that are not sufficiently close to others to be assigned to a specific topic. For this reason, Topic-1 can be discarded. However, for completeness, Topic -1 will also be presented.

The first group, which includes sessions one through seven, is characterized by seven topics, excluding Topic -1. Below are the topics extracted by BERTopic from the first group of sessions, along with the similarity scores for each word. In Figure 25, all topics are displayed, including Topic -1, which contains the ten least significant words for the first group of sessions, along with their similarity scores. In Figure 26, the word clouds of all topics extracted by LDA for the first set of psychotherapy sessions are shown.

Topic -1		Topic 0		Topic 1		Topic 2	
Word	Score	Word	Score	Word	Score	Word	Score
work	0.032090	work	0.031787	friends	0.063428	stefano	0.635368
time	0.031825	time	0.030335	good	0.051544	move	0.078951
sense	0.027595	things	0.030290	time	0.050805	talk	0.075535
fact	0.027025	going	0.026808	work	0.049765	years	0.071291
feel	0.025101	person	0.026130	marta	0.049109	missing	0.065252
day	0.024759	point	0.024008	months	0.046301	relationship	0.064142
things	0.023914	sense	0.022931	sense	0.046085	recognizes	0.062859
badly	0.023326	relationship	0.022743	years	0.038324	abilities	0.059253
years	0.022506	kind	0.020720	manfredi	0.037251	consider	0.053489
point	0.021615	people	0.020491	best	0.036174	aspects	0.051843
Topic 3		Topic 4		Topic 5		Topic 6	
Word	Score	Word	Score	Word	Score	Word	Score
priest	0.149894	law	0.343291	albania	0.598972	diary	0.552715
religious	0.149894	firm	0.303423	italy	0.467617	write	0.493911
orthodox	0.145234	lawyer	0.237651	live	0.335889	written	0.223288
prayers	0.145234	firms	0.118083	milan	0.301891	essential	0.214616
priests	0.124481	touch	0.115234	country	0.293931	meet	0.206198
mary	0.124481	colleague	0.106595	arrived	0.224157	fits	0.192648
belief	0.097777	field	0.090793	square	0.162021	writes	0.155296
ferragosto	0.097777	illegal	0.090793	joined	0.162021	badoo	0.118301
vergin	0.097777	legal	0.090793	location	0.162021	appointment	0.118301
pray	0.097777	impossible	0.083512	garibaldi	0.162021	dwel	0.118301

Figure 25. From the first Block (from the first to the seventh sessions), eight topics were extracted using the BERTopic model. The first topic labelled with -1 represents the 10 words that are less significant for the first group of sessions, while the other topics are represented by the ten words associated with each topic, along with their relative scores.



Figure 26. These are the eighty word clouds corresponding to the eight topics extracted with BERTopic from the first Block of sessions. The words size is proportional to the importance that each word has in a topic.

For the second Block of therapy (sessions from 8th to 14th), the model identified three key topics. The first topic is the number -1, and it shows the ten least representative words for this group of sessions. In Figure 27, these three topics are presented in a visualization that shows, for each topic, the ten words most likely to belong to it, along with their associated scores. The corresponding word clouds are also provided as another type of visual output (see Figure 28).

Topic -1		Topic 0		Topic 1	
Word	Score	Word	Score	Word	Score
hormone	0.288223	methadone	0.164689	things	0.040188
talk	0.129231	sober	0.164689	going	0.039095
friends	0.120117	deadly	0.164689	time	0.037351
ahah	0.113957	stage	0.156246	day	0.033595
friend	0.098142	drug	0.150256	point	0.033081
alessio	0.096074	alcohol	0.150256	told	0.031622
ahaha	0.085925	justify	0.141817	sense	0.030356
worthy	0.085925	cocaine	0.138609	exam	0.029305
rampant	0.078395	drink	0.115775	fact	0.029187
heh	0.078395	stay	0.104036	sister	0.027331

Figure 27. From the second Block (from 8th to 14th sessions), three topics were extracted using the BERTopic model. The first topic labelled with -1 represents the 10 words that are less significant for the first group of sessions, while the other topics are represented by the ten words associated with each topic, along with their relative scores.

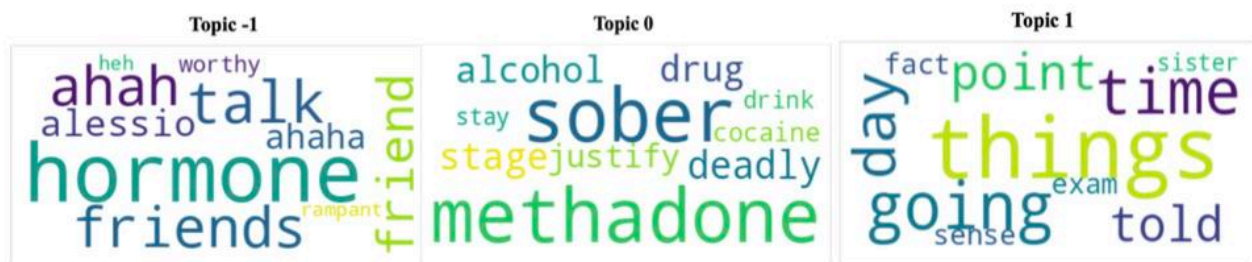


Figure 28. These are the three word clouds corresponding to the three topics extracted with BERTopic from the second Block of sessions. The size of the words is proportional to the importance that each word has in a specific topic.

In the third Block of therapy (sessions fifteen through twenty-one), the BERTopic model identified three topics. The visualizations consist of the top words with their respective scores for each extracted topic (Figure 29) and the corresponding word clouds for each topic (Figure 30). For this group of therapy sessions, no Topic-1 was generated.

Topic 1		Topic 2	
Word	Score	Word	Score
stefano	0.743088	things	0.042960
activates	0.192183	going	0.039899
good	0.117223	time	0.037888
garibaldi	0.104079	sense	0.036876
profile	0.104079	day	0.029995
intelligent	0.100893	feel	0.029939
instagram	0.096091	told	0.028694
powerful	0.089644	fact	0.027756
definitely	0.082626	point	0.027589
game	0.079137	talk	0.027366

Figure 29. From the third Block (from 15th to 21st sessions), two topics were extracted using the BERTopic model. The topics are represented by the ten words associated with each topic, along with their relative scores.

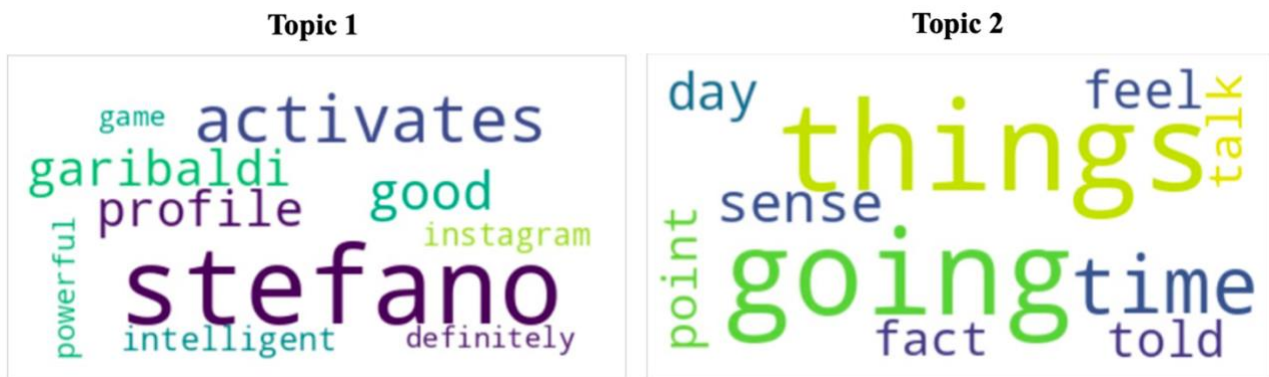


Figure 30. These are the two word clouds corresponding to the three topics extracted with BERTopic from the third Block of sessions. The size of the words is proportional to the importance that each word has in a topic.

The implementation of BERTopic for the final block of therapy (sessions twenty-two through twenty-eight) identified eight topics. The words in each topic are reported with their respective scores, which represent the weight of each term within its topic (Figure 31). The corresponding word clouds for each topic extracted by the model for this final group of psychotherapy sessions are presented in Figure 32. In this group, Topic-1 was also reported.

Topic -1		Topic 0		Topic 1		Topic 2	
Word	Score	Word	Score	Word	Score	Word	Score
things	0.032310	good	0.156340	day	0.035702	values	0.077291
sense	0.027981	understand	0.112676	going	0.033708	opinion	0.056953
time	0.027445	god	0.087539	time	0.031583	understand	0.053853
change	0.026301	game	0.077346	things	0.026795	person	0.053083
people	0.025876	mhm	0.070730	feel	0.026473	sense	0.046438
good	0.025840	andrew	0.070730	told	0.025175	negative	0.043547
person	0.025032	things	0.069366	phone	0.022515	experience	0.040995
going	0.021596	huh	0.066159	night	0.022453	positive	0.039347
happened	0.021261	interested	0.066159	days	0.022451	point	0.037038
feel	0.021040	burnt	0.062919	bad	0.022204	stefano	0.031918

Topic 3		Topic 4		Topic 5		Topic 6	
Word	Score	Word	Score	Word	Score	Word	Score
edimburgo	0.148009	yesterday	0.344581	friend	0.151533	selfish	0.231836
torino	0.083755	news	0.290974	friends	0.116528	psychotherapy	0.216855
law	0.075395	day	0.0260540	match	0.110753	powerful	0.216855
firm	0.067018	wednesday	0.0235282	met	0.102688	nice	0.208645
contacted	0.046446	fourth	0.164932	people	0.084204	suffering	0.191288
going	0.045755	month	0.121833	dinner	0.083878	positive	0.147187
contact	0.043876	fun	0.119916	couple	0.075636	stefano	0.131337
lawyer	0.041886	call	0.114860	super	0.071686	senses	0.128737
positions	0.040060	ago	0.104713	shamelessly	0.069167	antidepressant	0.128737
level	0.038895	finishing	0.097912	pays	0.069167	fucker	0.128737

Figure 31. From the fourth Block (from 22nd to 28th sessions), eight topics were extracted using the BERTopic model. The first topic labelled with -1 represents the 10 words that are less significant for the first group of sessions. In comparison, the other topics are represented by the ten words associated with each topic, along with their relative scores.



Figure 32. These are the eighty word clouds corresponding to the eight topics extracted with BERTopic from the fourth Block of sessions. The size of the words is proportional to the importance that each word has in a specific topic.

4.3. Sentiment Analysis of a Psychotherapy Case Using RoBERTa-GoEmotions

The present study is embedded within a broader research program on the same psychotherapy case. In a previous single-case study, we examined this treatment using human-coded process indicators and topic modelling, showing how thematic trajectories across 26 sessions were associated with a positive outcome (Liccione & Siciliano, 2025). Building on that work, the current study focuses specifically on the dynamics of emotions and sentiment across sessions, using the RoBERTa-GoEmotions model as an AI-mediated lens on therapeutic change. The primary aim is to explore the feasibility and clinical utility of transformer-based sentiment analysis in describing emotional trajectories throughout psychotherapy. More specifically, we (a) examine whether distinct emotional patterns characterize different phases of treatment (initial, middle, final sessions) and (b) investigate correlations between specific emotions across sessions. Our central hypothesis is that automatic emotion analysis can capture meaningful patterns in the emotional climate of therapy and provide quantitative indicators consistent with a positive outcome, thereby advancing process–outcome research through innovative AI-based methodologies.

4.3.1. Abstract

This exploratory single-case study investigates the feasibility of applying automated sentiment analysis to psychotherapy transcripts to describe session-by-session emotional patterns and correlations among discrete emotions. Complete transcripts from 26 sessions of an individual psychotherapy were analysed. A RoBERTa-based transformer model fine-tuned on the GoEmotions corpus was used to classify each text segment into six basic emotion categories (anger, disgust, fear, joy, sadness, surprise) plus neutral. Transcripts segmented into 512-token units, and emotion probabilities were aggregated to obtain session-level scores. Median and interquartile range values were computed for each emotion across three treatment phases (initial, middle, final), and Spearman's rank correlations were calculated to examine associations between emotions across sessions. Joy and neutrality emerged as the most prevalent categories throughout treatment. Joy showed a U-shaped pattern, with median values of 0.290 in the initial phase, 0.169 in the middle phase, and 0.268 in the final phase. Neutrality remained relatively stable (medians between 0.239 and 0.336). Anger increasing from a median of 0.074 in the initial phase to 0.165 in the middle phase and stabilizing at 0.144 in the final phase; sadness followed a more gradual course (medians 0.142, 0.188, and 0.169). Disgust and fear remained consistently low (disgust median < 0.003 across phases; fear medians 0.041–0.053). Significant positive correlations were observed between anger and disgust ($\rho = 0.569$, $p = 0.0024$) and between disgust and sadness ($\rho = 0.561$, $p = 0.0029$). In contrast, strong negative correlations emerged between joy and anger ($\rho = -0.611$, $p = 0.0009$) and between joy and sadness

($\rho = -0.478$, $p = 0.0136$). Although limited to a single case, these findings suggest that transformer-based sentiment analysis may offer a novel tool for Process–Outcome research, providing quantitative indicators of the emotional climate of psychotherapy.

4.3.2. Methods

This study aims to explore the potential of an artificial intelligence tool to analyse emotions expressed during therapy, and to verify the possibility of identifying characteristic patterns across the different stages of treatment and significant correlations between specific emotions. The central hypothesis is that automatic emotion analysis can provide quantitative insights that help understand the dynamics of therapeutic change, thereby advancing process-outcome research.

The hypotheses formulated for this study predict the possibility of observing modulation of emotional intensity during therapy, with characteristic patterns across the different phases of treatment. It is also hypothesized, based on the literature, that significant correlations will be observed among specific emotions, a trend that could reflect psychological dynamics relevant to the therapeutic process (Greenberg & Safran, 1989; Tanana et al., 2021). Negative correlations are expected between contrasting emotions, such as anger and joy, whereas positive correlations are expected between related emotions, such as sadness and anger.

Sample

The case study examined the entire therapeutic process of a 28-year-old individual, referred to as patient I, who signed an informed consent form for audio and video recording of each session. All patient-related materials were managed and stored in accordance with privacy regulations, and all sensitive data were anonymized before analysis to protect the patient's identity. At the start of the therapeutic process, I noted symptoms of anxiety and depression, including frequent panic attacks that significantly impacted her quality of life. The therapy was conducted in a private practice by a therapist experienced in the phenomenological hermeneutic approach (Liccione, 2019).

The therapeutic process initially consisted of weekly sessions, then became biweekly, and finally monthly. During therapy, the patient was assigned therapeutic tasks, including journaling and experiential tasks, according to the therapeutic tools and strategy of the phenomenological approach as stated in Liccione, 2019. The data corpus consists of the complete transcripts of the psychotherapy, organized into 26 sessions, which were examined individually to trace the emotional characteristics that emerged throughout the therapeutic process.

As we mentioned, this dataset of 26 therapies is part of a separate study on therapeutic transcripts that aims to make AI tools reliable for the everyday work of psychotherapists.

Procedure

The RoBERTa-GoEmotions model was implemented in Python using the Hugging Face Transformers library, obtained from the *SamLowe/roberta-base-go_emotions* repository and distributed under the MIT license. The implementation followed a structured process with three main phases, preceded by data preparation.

As a preliminary step, the complete set of 26 recorded sessions was transcribed by experienced psychotherapists following established psychotherapy transcription procedures (Mergenthaler & Stinson, 1992). The transcripts were anonymized and organized at the session level; each session was then exported as a plain-text (.txt) file to ensure compatibility with NLP preprocessing and to preserve the temporal structure of the therapeutic material.

The first step involved tokenizing the text with *Autotokenizer*, which converts natural language into a numerical form that the model can understand. In this context, tokens are discrete units of text turned into numerical vectors that the model can process algorithmically. The analysed corpus consisted of the complete transcripts of 26 therapeutic interviews, each processed individually to maintain the temporal and contextual integrity of the sessions.

To apply the model in the present study, it was necessary to translate the original Italian material into English, as the algorithm was trained exclusively on English-language data. A qualified professional with experience in clinical and research texts carried out this translation. The translated corpus was then double-checked using the Reverse Translation method (Saldanha & O'Brien, 2014) to enhance linguistic accuracy and minimise potential distortions. Despite this procedure, we acknowledge that some loss of meaning may persist, particularly for idiomatic expressions, culturally embedded nuances, and linguistic constructs specific to the Italian therapeutic context.

Afterwards, each transcript was split into 512-token segments, a limit set by the BERT architecture that RoBERTa supports. This process required careful handling of narrative breaks to prevent losing semantic meaning between neighbouring segments. For each segment processed, the model generates a multidimensional output consisting of 28 distinct scores, each corresponding to a specific emotional category in the *GoEmotions* dataset. These values are then normalized using a *softmax* function, producing a probability distribution in which each score ranges from 0 to 1 and sums to 1 (100%).

The methodological choices made in this study led to the use of a classification based on Ekman's six basic emotions: anger, disgust, fear, joy, sadness, and surprise, plus a neutral category. The 27 emotions in the dataset were then grouped into six emotional categories and one neutral category, following the methods used by Burkhardt and colleagues (2022) and Demszky and colleagues (2020).

The process of aggregating the 28 original categories into Ekman's six categories plus one neutral category was carried out as follows: the Anger category includes anger, annoyance, and disapproval; Disgust remains a separate category; Fear includes fear and nervousness; Joy aggregates all positive emotions (admiration, amusement, approval, care, desire, excitement, gratitude, joy, love, optimism, pride, relief); Sadness includes sadness, disappointment, embarrassment, pain, and remorse; Surprise includes all ambiguous emotions (confusion, curiosity, realization, surprise); Neutral remains a distinct category.

This methodological choice was driven by two main reasons: the uneven predictive performance of the RoBERTa-GoEmotions model across emotion labels, and the desire to rely on a well-established theoretical framework. Aggregating into broader categories aims to mitigate variable accuracy, while adopting Ekman's classification facilitates the interpretation and contextualization of the results within the existing psychological literature.

Statistical Analysis

Statistical correlation analyses were performed on the results to examine relationships among emotions detected in the text throughout the entire therapeutic process. The statistical measure chosen was Spearman's rank correlation coefficient, a non-parametric statistical measure of correlation (Spearman, 2010).

The nature of the data obtained determined this methodological choice. The emotional scores generated by RoBERTa-GoEmotions do not follow a normal distribution, a characteristic that classifies them as non-parametric data. Although the model provides numerical outputs, these more appropriately represent an ordinal scale of emotional intensity rather than a regular interval scale.

RoBERTa-GoEmotions assigns numerical probabilities to each emotion, but these scores do not represent direct measurements of emotional intensity: they indicate which emotion is more likely, but not the distance or “how much” more likely it is than another with absolute precision. A higher score for “joy” than “sadness” simply indicates a higher probability, without implying a linear or constant difference in intensity between the emotions.

Spearman's correlation coefficient is particularly well-suited for analysing this type of data because it operates on ranks rather than absolute values. To calculate it, the original values are first converted into ordinal ranks, which represent the position of each value. Next, the correlation between these positions is calculated.

Given the non-parametric nature of the data generated by the RoBERTa-GoEmotions model, appropriate descriptive measures, such as the median and interquartile range (IQR), were used. The median provides a robust measure of central tendency that is not influenced by extreme values, which is important when analysing emotional data that may show isolated peaks. The IQR, which measures the spread of data around the median, provides insight into the variability of each emotion and phase, helping us understand the stability or fluctuation of the different emotional components.

4.3.3. Results

The analysis conducted using the RoBERTa-GoEmotions model across the entire therapeutic process (26 sessions) revealed distinctive emotional patterns and significant correlations among emotional categories. Overall, the trend in emotions during treatment showed that joy and neutrality were the most frequently observed, while disgust and fear were the least. Surprise, anger, and sadness were present at an intermediate level throughout the therapeutic process.

Joy was one of the most constant and prevalent emotions throughout the therapeutic process. As shown by the longitudinal trajectories of session-level scores (Appendix 8.1) and by the phase-specific medians (Appendix 8.2), joy remained relatively high across all treatment phases, with a median value of 0.290 in the initial phase, decreasing to 0.169 in the middle phase, and rising again to 0.268 in the final phase. This pattern suggests a temporary attenuation of positive affect in the central phase of therapy, followed by a substantial recovery towards the end of treatment.

The distribution of joy scores across phases (Appendix 8.3, left panel) further clarifies this trend. Variability, indexed by the interquartile range (IQR), was highest in the initial phase ($IQR = 0.131$), markedly reduced in the middle phase ($IQR = 0.039$), and then increased moderately in the final phase ($IQR = 0.083$). These results indicate that, despite expected fluctuations during psychotherapy, the patient's positive emotional experience is broadly maintained and appears to consolidate again in the concluding sessions.

The qualitative analysis of the material supports this trajectory. Around the middle phase of therapy, moments of genuine joy emerged even amidst stress and uncertainty – for instance, when the patient reacted to good news with visible emotion: “It’s a beautiful thing, I feel like crying just thinking about it, I really wanted to cry out of joy all day today. Moreover, it’s my grandfather’s birthday” (session 16). In the final sessions, expressions of joy appear more composed and reflective, increasingly anchored to perceived progress, restored equilibrium, and future-oriented agency. This late-phase consolidation is also evident in the therapist’s affective atonement and explicit positive appraisal of the patient’s trajectory – for example, “Come on, I’m happy. It all seems to be going well, doesn’t

it?” (session 26) and “I feel like things are definitely going well.” (session 26). In addition, the therapist reinforces the patient’s exploratory stance as a marker of growth (“Well, but you see you are gaining experience. That’s very positive in my opinion” (session 26). Taken together, these examples suggest a shift from episodic positive affect to a more stable, integrated form of joy, consistent with increased self-efficacy and a sense of closure at treatment termination.

Neutrality was the second most common emotional category during the therapeutic process, remaining relatively high across all phases (median values ranging from 0.239 to 0.336; see Appendix 8.1 and Appendix 8.2). The phase-specific distribution of neutrality (Appendix 8.3, right panel) displayed a stable pattern, with interquartile ranges (IQRs) consistently between 0.047 and 0.074, indicating low variability and a steady presence of this emotional state throughout treatment.

The sustained prevalence of neutral affect, especially in the middle and final phases, when its incidence often exceeded the average for other emotions, may reflect distinct aspects of the therapeutic process. On the one hand, it could signal increased emotional stability in the patient; on the other, it may correspond to moments of reflection and cognitive processing typical of psychotherapy, in which overt emotional intensity is attenuated in favor of more rational and integrative work on experience.

Anger showed a particularly dynamic pattern throughout treatment. As illustrated by the longitudinal trajectories (Appendix 8.1) and phase-specific medians (Appendix 8.2), anger started from relatively low levels in the initial phase (median = 0.074), increased markedly in the middle phase (median = 0.165), and then stabilized at 0.144 in the final phase. This pattern suggests that the emergence and elaboration of anger became especially salient in the central phase of therapy.

The distribution of anger scores across phases (Appendix 8.4, right panel) further supports this interpretation. The interquartile range (IQR) increased from 0.067 in the initial phase to 0.080 in the middle phase, then decreased slightly to 0.070 in the final phase, indicating greater dispersion and intermittent intensifications of anger in the middle phase. In line with this, the session-level trajectories (Appendix 8.1) show marked peaks of anger in Sessions 14 and 23, where scores exceeded 0.28, suggesting specific therapeutic moments in which this emotion became particularly prominent and clinically relevant.

The qualitative evaluation of the transcripts corroborates this quantitative profile, capturing both the intensity of anger activation and its progressive regulation. In an early session, anger is described as a high-arousal state with somatic escalation (“On June 23, I had a peak of anger and annoyance, to the point where I was really shaking at one point”), alongside aggressive impulses in response to

perceived disrespect (“So I really wanted to commit a crime... and so it pissed me off”) (session 9). By the later phase, anger is framed more explicitly as a recurrent cognitive–affective process that can be observed and mentalized (“the anger does... when I maybe fall back into the circle of thoughts, I think angrily and say, ‘No because this, this, this...’”). It co-occurs with impulse inhibition (“I stood there wanting to trip him, but I didn’t”) (session 24). These excerpts suggest a shift from acute, physiologically intense anger episodes toward a more reflective and regulated form of anger, consistent with its stabilization in the final phase.

Sadness followed a relatively linear and gradual pattern compared with the other negative emotions. As shown by the session-level trajectories (Appendix 8.1) and phase-specific medians (Appendix 8.2), the median increased from 0.142 in the initial phase to 0.188 in the middle phase, then slightly decreased and stabilized at 0.169 in the final phase. Although sadness was not among the most prevalent emotions, the longitudinal plot shows clear peaks of intensity in Sessions 23 and 24 (Appendix 8.1), suggesting a possible intensification of emotions in the later stages of treatment.

The distribution of sadness scores across phases (Appendix 8.4, left panel) was relatively stable, with interquartile ranges (IQRs) ranging from 0.042 to 0.075 and no significant fluctuations. This pattern indicates a consistent yet controlled presence of this emotional component, which becomes more pronounced in certain advanced sessions rather than showing abrupt global shifts.

The qualitative considerations align with this profile and clarify how sadness was configured across the course of sessions. In the early phase, sadness was already present but often embedded in interpersonal strain, guilt, and emotional burden. For example, the patient described feeling emotionally “heavy” and “trapped” in responsibility when faced with another person’s distress (session 4). In another early session, sadness was explicitly thematized through the patient’s report of maternal invalidation regarding feeling “depressed” and “sad,” highlighting early exposure of depressive-sad affect despite ostensibly “positive” external circumstances (session 9). Then, sadness became more directly articulated as loneliness and emotional isolation (e.g., “a sense of loneliness... the desert, nothingness... I hadn’t cried yet today,” session 18). In later sessions, it also appeared in a somatic-experiential register, described as exhaustion and “emptiness inside,” accompanied by tearfulness under performance demands (session 19). Finally, in an advanced session, sadness was expressed in a more complex relational form – an empathic response to the partner’s perceived loneliness (“it makes me very sad... he is really a super lonely person,” session 24) – suggesting ongoing grief/attachment processing that may help contextualize the late-session peaks observed quantitatively.

Surprise showed a generally decreasing pattern throughout treatment. As indicated by the

longitudinal trajectories of session-level scores (Appendix 8.1) and by the phase-specific medians (Figure 2), the median value declined from 0.158 in the initial phase to 0.108 in the middle phase, then rose slightly and stabilized at 0.134 in the final phase. Although overall levels remained moderate, the time series reveals distinct peaks of surprise in Sessions 19 and 22 (Appendix 8.1), suggesting specific moments characterized by heightened unexpected or disorienting experiences.

The phase-wise distribution of surprise (Appendix 8.3, middle panel) showed relatively low variability, with interquartile ranges (IQRs) of 0.073-0.076. This limited dispersion indicates that surprise remained a relatively stable component of the patient's emotional profile, punctuated by a few salient episodes of increased intensity.

Disgust and fear were the least frequently observed emotions throughout the treatment, remaining consistently low. As shown by the longitudinal trajectories and phase-specific summaries (Appendix 8.1 and Appendix 8.2), disgust was nearly absent, with median values below 0.003 across all phases, indicating that this emotion had only a minimal impact on the patient's experience.

Fear, although more prevalent than disgust, also showed low overall intensity. Its median values ranged from 0.041 in the initial phase to 0.053 in the middle phase and 0.052 in the final phase (Appendix 8.1 and Appendix 8.2). The distribution of fear scores across phases (Appendix 8.5, right panel) showed limited variability, with the IQR decreasing from 0.063 in the initial phase to 0.045 in the final phase. This pattern suggests a constant but modest presence of fear, with no substantial oscillations over the course of treatment.

Consistent with our previous single-case investigation of the same treatment, in which human-coded process indicators and topic trajectories were associated with a positive outcome (Liccione & Siciliano, 2025), the present correlation analysis supports a coherent emotional architecture underlying the therapeutic process. Spearman's tests identified significant associations among emotions across sessions (see Table 8).

A robust positive correlation emerged between anger and disgust ($\rho = 0.5691$, $p = 0.0024$, 95% CI [0.2104, 0.7764]), indicating concurrent activation of these affective states within sessions. A similarly strong positive association was observed between disgust and sadness ($\rho = 0.5611$, $p = 0.0029$, 95% CI [0.2069, 0.8152]), suggesting partially overlapping negative emotional dynamics.

The strongest negative association involved anger and joy ($\rho = -0.6110$, $p = 0.0009$, 95% CI [-0.8338, -0.2649]). Joy and sadness were also inversely related ($\rho = -0.4776$, $p = 0.0136$, 95% CI [-0.7674, -0.0830]), indicating a moderate antagonism between distress-related and positive emotional states across sessions. Disgust and joy showed a borderline negative correlation ($\rho = -0.3894$, $p =$

0.0493, 95% CI [-0.7092, 0.0045]) and should therefore be interpreted cautiously. Finally, disgust and neutral were moderately negatively correlated ($\rho = -0.4151$, $p = 0.0350$, 95% CI [-0.7008, -0.0373]), suggesting that higher disgust co-occurred with reduced neutrality.

Overall, these results indicate a structured pattern in which negative emotions co-vary with one another and show inverse associations with joy, aligning with the expected affective organization in a psychotherapy course with a positive outcome.

EMOTION PAIR	P (SPEARMAN)	P-VALUE	95% CI
ANGER ↔ DISGUST	0.5691	0.0024**	[0.2104, 0.7764]
ANGER ↔ JOY	-0.6110	0.0009***	[-0.8338, -0.2649]
DISGUST ↔ JOY	-0.3894	0.0493*	[-0.7092, 0.0045]
DISGUST ↔ SADNESS	0.5611	0.0029**	[0.2069, 0.8152]
DISGUST ↔ NEUTRAL	-0.4151	0.0350*	[-0.7008, -0.0373]
JOY ↔ SADNESS	-0.4776	0.0136*	[-0.7674, -0.0830]

Table 8. Spearman correlations between emotion scores across sessions. Significant pairwise correlations among emotion dimensions extracted with the RoBERTa-GoEmotions model across sessions (N = 26). Negative emotions showed positive coupling (anger–disgust; disgust–sadness), while joy was inversely associated with anger and sadness. The disgust–joy association was borderline and should be interpreted cautiously given its confidence interval. Note: ρ = Spearman’s rank correlation coefficient; p-values are two-tailed; CI = 95% confidence interval. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

4.4. Are Generative Models Ready for Psychotherapy Research? Evidence from Prompt-Based Topic Modelling Comparison

The present article is situated within a broader methodological effort to understand whether generative language models can be used responsibly and reproducibly for psychotherapy process research – especially when the analytic goal is not free interpretation, but structured, protocol-driven coding. Building on established topic-modelling approaches, we conduct a systematic comparison between traditional Latent Dirichlet Allocation (LDA) and prompt-based topic modelling implemented with two widely used generative systems (Claude 3.5 Sonnet and ChatGPT/GPT-4), applied to a corpus of 26 psychotherapy-session transcripts coded into nine predefined, clinically grounded thematic categories. The primary aim is to assess whether prompt-based topic modelling can achieve meaningful convergence with LDA while offering complementary strengths (e.g., contextual sensitivity) in capturing thematic trajectories and therapist–patient conversational

dynamics over time. More specifically, we evaluate agreement through similarity metrics (e.g., Jaccard overlap and exact concordance), examine longitudinal topic evolution across sequential session blocks, and test whether the two approaches produce clinically coherent asymmetries between therapist and patient discourse. A secondary goal is to foreground methodological reliability as a prerequisite for adoption: we document where generative systems may fail to comply with strict protocols (e.g., format adherence and category constraints), and how such deviations impact the feasibility of comparative analyses.

4.4.1. Abstract

Process-outcome research in psychotherapy requires increasingly sophisticated methodologies for analysing therapeutic transcripts. The emergence of generative artificial intelligence offers new opportunities for textual content analysis, complementing traditional statistical methods. To systematically compare the effectiveness of prompt-based topic modelling (Claude 3.5 Sonnet and ChatGPT/GPT-4) with traditional Latent Dirichlet Allocation (LDA) in identifying thematic patterns in psychotherapeutic conversations. A comparative analysis was conducted on 26 transcripts of individual psychotherapy sessions using the cognitive-neuropsychological approach. The corpus was analysed using nine predefined thematic categories: Suffering/Pathology/Symptoms, Family, Emotional Relationships, Occupation/Study, Social Network, Ways of Being, Planning, Therapeutic Relationship, and Other. Concordance measures were evaluated using the Jaccard index and exact concordance, and the temporal evolution and therapist-patient conversational dynamics were analyzed. The concordance between LDA and Claude 3.5 Sonnet showed Jaccard index values ranging from 47.56% to 57.22% in the four sequential groups analysed (mean: 52.13% $\pm$ 4.05%). Exact agreement recorded lower values (23.88%–34.65%), reflecting the multi-thematic complexity of therapeutic conversations. ChatGPT/GPT-4 exhibited systematic limitations in adhering to the methodological protocol, including the unauthorized introduction of additional categories and deviations in the response format, rendering it impossible to include it in the comparative analysis. Longitudinal analysis revealed distinctive evolutionary patterns: SPS showed concentration in the early groups in LDA and in the central groups in Claude. Distributions by conversational role confirmed clinically significant asymmetries, with RET being prevalent in therapist utterances and SPS/MDE being predominant in-patient utterances. Despite the limitations of this comparison, prompt-based topic modelling appears to be a methodologically valid approach that complements traditional methods for analysing therapeutic discourse. The partial but significant convergence between the two methods, combined with their distinctive interpretive characteristics, supports integrating their use to achieve a multidimensional understanding of the therapeutic process.

4.4.2. Methods

We adopted a comparative research design to evaluate the effectiveness of prompt-based topic modelling relative to the traditional Latent Dirichlet Allocation (LDA) approach in analysing psychotherapeutic transcripts. Our study was divided into three main phases (data preparation, implementation of methodologies, comparative analysis) and followed a systematic methodological approach that enabled rigorous comparison of the different text analysis tools.

In the first phase, we defined the theoretical framework for the concept of topic in the psychotherapeutic context, based on the conceptualization proposed by Traini and Liccione (Traini, 2008; Liccione 2019). We identified topics as pragmatic choices that allow us to identify macro-themes within therapeutic conversations, distinguishing them from isotopies as practical rather than semantic phenomena. This theoretical distinction guided our operationalization of topics as significant existential areas in the patient's narrative, which are subject to magnification or numbing during therapy.

In the second phase, we prepared the dataset by rigorously anonymizing the transcripts. Next, we developed a standardized prompt engineering protocol to guide advanced language models in identifying topics. This protocol was developed through an iterative meta-prompting process, in which the models themselves helped refine the instructions until an optimized version was achieved that balanced clarity, completeness, and computational efficiency.

In the third phase, we implemented three topic modelling methodologies in parallel: Claude 3.5 Sonnet, ChatGPT (GPT-4), and LDA. To ensure the comparability of the results, we used the same set of nine clinically relevant topics as input parameters for all three approaches. This methodological approach allowed us to evaluate the accuracy of topic identification and the ability of each method to capture the temporal evolution of themes throughout the therapeutic process.

The implementation of this comparative methodology required the creation of a text corpus specifically structured for this type of analysis. Our corpus consists of 26 transcripts of psychotherapy sessions, each lasting approximately 50 minutes, from an individual therapeutic program using the cognitive-neuropsychological approach. The clinical case concerns a 28-year-old adult patient who undertook a therapeutic program to address existential difficulties in her emotional, family, and professional life. The sessions, initially weekly and gradually becoming less frequent, were audio- and video-recorded after obtaining informed consent.

To ensure compliance with ethical and legal standards, we implemented a data anonymization process that involved systematically replacing all proper names with pseudonyms, removing specific

identifying details, and altering non-essential demographic information. Attention was paid to removing the existential macro-context, previously identified through LDA analysis, as it contained personal information about the patient and her social network. This approach allowed us to preserve the integrity of the therapeutic content while maintaining high standards of privacy protection, in accordance with Italian professional secrecy regulations and international guidelines for psychological research.

Once the dataset preparation was complete, we proceeded to develop methodological tools for prompt-based analysis. We developed a standardized prompt engineering protocol through a systematic four-step process. Initially, we conducted a literature review to identify the main prompt engineering techniques applicable to generative AI tools. Based on this survey, we developed an initial version of the prompt, which was then iteratively refined using the meta-prompting technique. During this optimization phase, we used the same language models intended for analysis (Claude 3.5 Sonnet and GPT-4) to obtain systematic feedback on the developed prompt. This approach allowed us to identify areas for improvement in the instructions, expand the methodological specifications, and refine the guidelines' clarity. The iterative process continued until a final version was achieved, validated by both models for clarity, completeness, and computational optimization. The final prompt integrates nine clinically relevant thematic categories: Suffering/Pathology/Symptoms (SPS), Family (FAM), Emotional Relationships (RA), Occupation/Study (OS), Social Network (RS), Ways of Being (MDE), Planning (PRG), Therapeutic Relationship (RET), and Other (ALT). Each category was operationally defined with specific inclusion criteria and accompanied by practical examples to guide the systematic identification of topics in the transcripts. With the protocol thus defined, we proceeded to apply the three analysis methodologies.

For prompt-based topic modelling, we used Claude 3.5 Sonnet and ChatGPT (GPT-4) through their standard user interfaces, keeping the default settings to ensure reproducibility. The analysis protocol consisted of four standardized steps for both models. Initially, we provided an introductory message to contextualize the analysis's objective, followed by the complete prompt in a single message. The transcripts were then analysed in sequential groups of 30 lines, a division determined through a collaborative optimization process with the AI models to balance sufficient context with a manageable computational load. After each group of lines was analysed, we sent standard confirmation messages.

Once labelling was complete, we asked each model to summarize the results in an Excel-compatible CSV file, facilitating automatic transcription and subsequent data analysis. This uniform approach ensured methodological consistency across the two tools, enabling a systematic comparison

of the results. In parallel with the analysis using advanced language models, we implemented the traditional statistical approach to complete the comparative framework. Implementing the traditional LDA-based approach required more thorough data preparation than did the advanced language models. We followed a systematic six-step process to transform the raw transcripts into a format suitable for statistical analysis. The first step involved creating a structured dataset in Microsoft Excel, with dedicated columns for interview number, conversation blocks, conversation turns, and textual content. In the second step, we implemented a cleaning process that involved defining and removing *stopwords*, eliminating punctuation, removing unnecessary spaces and overly short words, and converting the text to lowercase.

The third phase involved advanced pre-processing through tokenization, stemming, further removal of *stopwords*, and *recomposition* of tokens into coherent textual units. In the fourth phase, we constructed the Document-Term Matrix, quantitatively representing the frequency of each word in each document. The fifth phase involved converting the matrix into a numerical format compatible with LDA and defining the number of topics to be identified. In the last phase, we applied the LDA algorithm to the prepared matrix to identify latent themes based on word distributions across documents. This systematic approach ensured statistically robust results, providing a reliable basis for comparing the effectiveness of prompt engineering-based methods.

To evaluate the validity and reliability of the results obtained from the three methodologies, we implemented a comparative analysis protocol that considered several performance parameters. We analysed the distribution of topics identified by each approach, evaluating the semantic consistency and clinical relevance of the emerging themes.

The analysis assessed the temporal evolution of topics across the 26 sessions, allowing us to identify patterns of change and thematic stability in the therapeutic process. We also examined the degree of overlap between results from the different methods, using similarity metrics to quantify convergence or divergence between the approaches. Attention was paid to evaluating each method's ability to identify clinically significant topics by qualitative analysis of the content that emerged and its relevance to the clinical case analysed. This multidimensional approach provided a comprehensive assessment of the relative effectiveness of prompt-based topic modelling compared to the traditional LDA-based method, offering valuable insights for future applications of these tools in psychotherapeutic research.

4.4.3. Results

We analyzed a corpus of 26 psychotherapy sessions using a comparative methodological approach between traditional topic modeling techniques (LDA) and innovative approaches based on prompt engineering (Claude 3.5 Sonnet and GPT-4). The dataset was divided into four sequential groups: the first two groups each comprise six sessions (groups 1-2), while the remaining two groups each comprise seven sessions (groups 3-4), for a total of 26 sessions analyzed. The analysis was implemented using automatically generated Python code executed in Google Colab, applying six methodological comparisons: qualitative evaluation of labels, topic distribution analysis, confusion matrix generation, temporal evolution analysis, transition matrix examination, and comparison of therapist-patient conversational roles.

The algorithms were labeled according to a predefined taxonomy that includes nine main categories: ALT (Other), OS (Occupation/Study), RA (Emotional Relationships), RET (Therapeutic Relationship), MDE (Ways of Being), SPS (Suffering/Pathology/Symptoms), FAM (Family), PRG (Plans), RS (Social Relationships). The concordance among the methods was assessed using the Jaccard index and exact concordance measures, given the multi-thematic nature of conversational jokes in the therapeutic context.

Quantitative analysis of the concordance between LDA and Claude 3.5 Sonnet reveals systematic patterns across the four groups of sessions analyzed. Overall concordance measured using the Jaccard index ranges from 47.56% to 57.22%. Specifically, the first group shows a concordance of 52.65% (95% CI: 48.32-56.98), the second group 47.56% (95% CI: 43.15-51.97), the third group 51.10% (95% CI: 46.73-55.47), while the fourth group shows the highest concordance with 57.22% (95% CI: 52.89-61.55) (Table 9).

Groups	Perfect agreement	Adjusted agreement	Jaccard similarity
1	28.62%	30.05%	52.65%
2	23.88%	25.45%	47.56%
3	26.82%	28.35%	51.10%
4	34.65%	37.11%	57.22%

Table 9. Agreement and similarity indices between LDA and Claude 3.5 Sonnet for topic identification across four groups of psychotherapeutic sessions.

Exact concordance, which considers only perfect matches in topic assignment, is significantly lower, ranging from 23.88% to 34.65%. The first group has an exact concordance of 28.62%, the second

23.88%, the third 26.82%, and the fourth 34.65%. The substantial difference between Jaccard concordance and exact concordance ($\Delta = 23.77 \pm 3.42$) reflects the multi-thematic complexity of therapeutic conversations, where a single utterance can contain references to multiple topics simultaneously. Analysis of thematic distributions reveals distinctive patterns in topic identification between the two methods. The RET (Therapeutic Relationship) topic shows a systematically higher prevalence in Claude's identification than in LDA across all groups analyzed. In the first group, Claude identifies 513 occurrences of RET compared to 325 in LDA, maintaining this trend in subsequent groups (426 vs. 224 in the second group) (Figure 33).

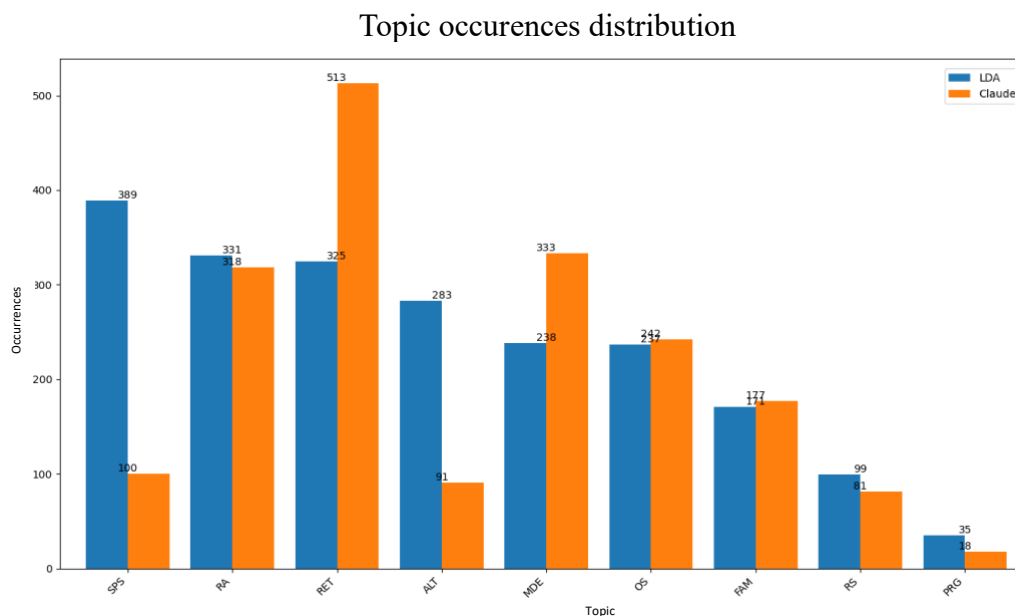


Figure 33. Absolute frequency distribution of topics identified by LDA and Claude 3.5 Sonnet in the first group of psychotherapeutic sessions (sessions 1-6).

The SPS (Suffering/Pathology/Symptoms) topic shows the opposite distribution, with LDA systematically recording higher frequencies: 389 occurrences in the first group compared to 100 for Claude, 250 vs. 77 in the second group. The confusion matrices show significant variability in concordance for individual topics. ALT shows concordance rates ranging from 24.84% to 72.44% in the different groups, with a progressive increase from the initial to the final phase of therapy. OS maintains a relatively stable concordance between 38.27% and 44.21%. SPS shows the lowest concordance values, ranging from 10.84% to 33.59%, suggesting substantial interpretative differences between the two methods in identifying symptomatic suffering (Table 10).

	Group 1	Group 2	Group 3	Group 4
ALT:	27.46%	24.84%	41.86%	72.44%
OS:	43.24%	42.36%	38.27%	44.21%
RA:	41.32%	36.51%	32.73%	44.16%
RET:	44.46%	44.53%	38.68%	43.67%
MDE:	31.45%	30.80%	22.69%	36.68%
SPS:	10.84%	12.00%	19.97%	33.59%
FAM:	42.73%	32.05%	38.25%	40.09%
PRG:	14.81%	16.67%	11.96%	32.92%
RS:	28.64%	28.23%	21.25%	25.13%

Table 10. Summary of confusion matrix values across four session groups.

The analysis of thematic co-occurrences reveals significant patterns of overlap. In the first group, when LDA identifies MDE, Claude classifies the same joke as RET in 21.6% of cases ($n = 47$) and as RA in 23.8% of cases ($n = 52$). This overlap suggests a partial semantic convergence between the categories Modes of Being, Therapeutic Relationship, and Affective Relationships in algorithmic recognition. Longitudinal analysis of thematic evolution across the four sequential groups reveals characteristic temporal patterns. SPS (Suffering/Pathology/Symptoms) shows a concentrated distribution in the first two groups (389 and 250 occurrences, respectively) in LDA, with a progressive decrease in the third and fourth groups (198 and 156 occurrences). Claude presents a different pattern, with a greater concentration in the central groups (100, 77, 89, and 78 occurrences in the four groups, respectively). The transition matrices highlight distinctive sequential dynamics. In the first group, LDA records 122 self-transitions for RET, 121 transitions to RA, and 134 to SPS. Claude, for the same topic, presents 100 self-transitions, 216 transitions to MDE, and 188 to RA (Figure 34).

It is important to note that the differences between SPS and RET in LDA and Claude were analyzed and explained by the researchers as a related issue that can be addressed with an appropriate prompt, also with a view to ensuring validity and replicability. In fact, the prompt in Claude tends to incorporate themes related to suffering as integral to the therapeutic process and, consequently, to the RET framework.

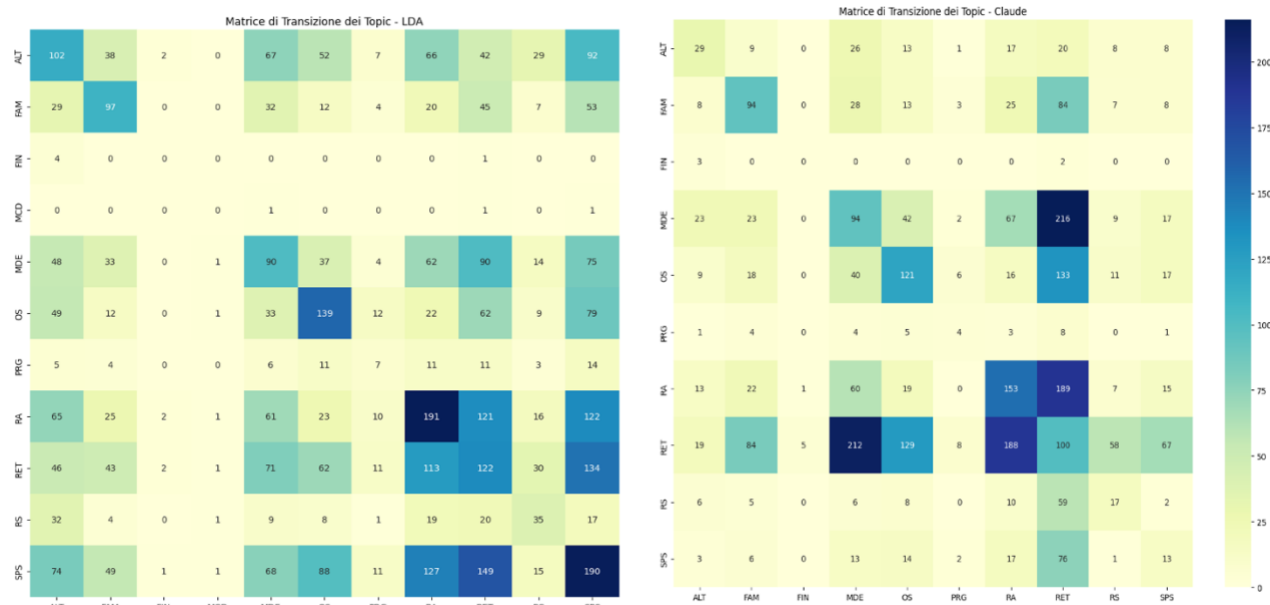


Figure 34. Topic transition matrices for LDA (top) and Claude 3.5 Sonnet (bottom) in the first group of psychotherapeutic sessions (1-6). Note: Cell values represent absolute transition frequencies between topics. Rows indicate source topics and columns indicate target topics.

This pattern suggests that Claude tends to identify narrative sequences that focus more on relational aspects and existential modalities, whereas LDA distributes transitions more evenly across thematic domains. SPS presents differential self-maintenance patterns: LDA records 190 self-transitions in the first group compared to Claude's 83, indicating a greater tendency of the traditional method to maintain symptomatic classification in consecutive lines. Transitions from SPS to other topics show that LDA favors links to RET (149) and RA (127), while Claude shows more moderate transitions (76 to RET, 17 to RA). Topics related to specific areas of life (FAM, OS, RS) show temporal clustering patterns. OS in LDA has 139 self-transitions in the first group, falling to 145 in the second, while Claude records 121 and 82 self-transitions, respectively. FAM maintains similar patterns with 97 self-transitions in LDA in the first group, falling to 63 in the second, while Claude goes from 94 to 41 self-transitions. Analysis of the thematic distribution between therapist and patient reveals systematic asymmetries characteristic of therapeutic roles. RET (Therapeutic Relationship) is markedly prevalent in the therapist's remarks. In the first group, Claude identifies 460 occurrences for the therapist versus 53 for the patient (ratio 8.7:1), while LDA records 248 vs. 77 (ratio 3.2:1) (Figure 35).

Topic distribution comparison: Therapist vs. Patient - Claude

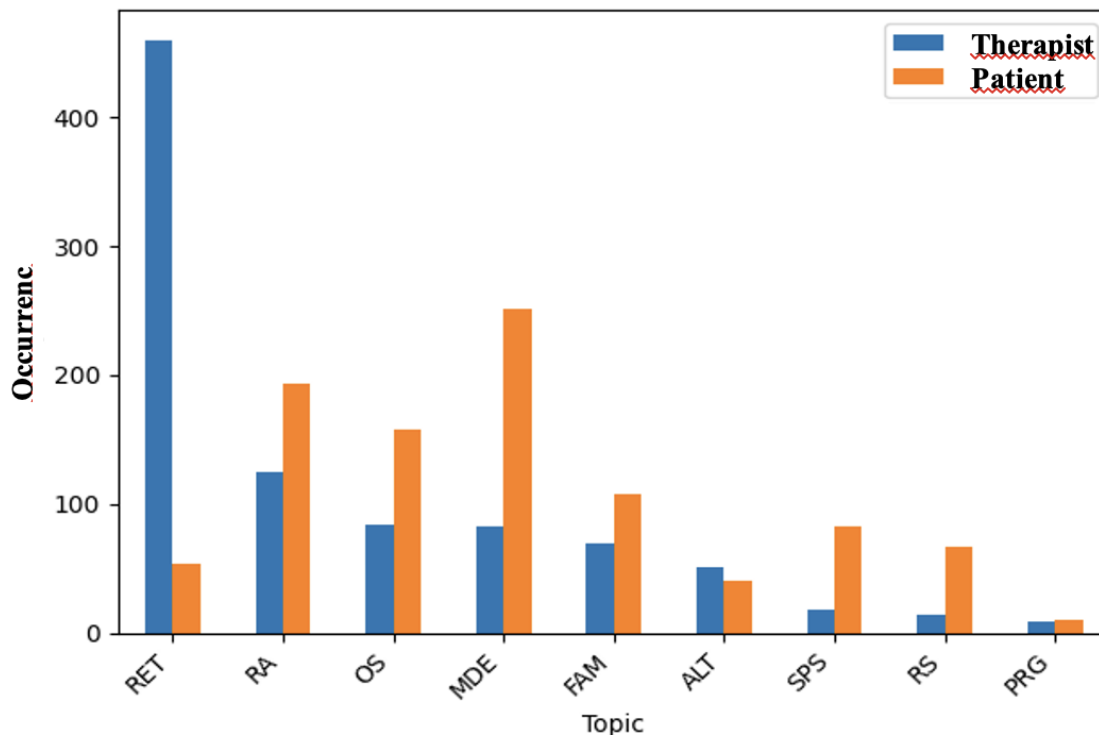


Figure 35. Topic distribution comparison between therapist and patient utterances, as identified by LDA (top) and Claude 3.5 Sonnet (bottom) in the first group of psychotherapeutic sessions (1-6).

This asymmetry remains constant: in the second group, Claude records 400 vs. 26 (ratio 15.4:1), LDA 162 vs. 62 (ratio 2.6:1). SPS (Suffering/Pathology/Symptoms) shows the opposite distribution, with a significant prevalence in the patient's remarks. LDA highlights this trend with 256 occurrences for the patient versus 133 for the therapist in the first group (ratio 1.9:1), maintaining the pattern in the second group (170 vs. 80, ratio 2.1:1). Claude shows the same pattern with lower frequencies: 82 vs. 18 in the first group (ratio 4.6:1), 69 vs. 8 in the second (ratio 8.6:1). MDE (Modes of Being) shows an asymmetrical distribution in favor of the patient, particularly marked in Claude. In the first group, the patient recorded 251 occurrences compared to 82 for the therapist (ratio 3.1:1), in the second 157 vs. 51 (ratio 3.1:1). Topics related to specific areas of life (OS, FAM, RS) consistently maintain a higher frequency in the patient's remarks, reflecting the subject's narrative focus on his own existential domains. The application of prompt-based topic modeling through GPT-4 revealed systematic limitations that compromised the comparative analysis. The main critical issues identified include: (1) failure to adhere to the response format specified in the prompt, with inconsistencies in the transcription of lines (variations from 5 to 15 words instead of the 5 requested); (2) autonomous introduction of unauthorized categories, specifically “Interpersonal Relationships (IR)”, violating the

predefined protocol; (3) frequent interruptions in processing from the third iteration onwards, related to the complete transcription of lines. Analysis of protocol deviations reveals that GPT-4 introduced the IR category in the second iteration and progressively incorporated content from FAM, RS, and RA. This categorical contamination made a methodologically valid comparison impossible, compromising the integrity of the comparative analysis. The persistence of these errors across multiple attempts suggests structural limitations in the tool's ability to handle complex, structured analytical tasks, rather than random or correctable issues.

The comparative analysis demonstrated substantial agreement between LDA and Claude 3.5 Sonnet (mean Jaccard index: $52.13\% \pm 4.05\%$), with consistent evolutionary patterns reflecting clinically meaningful therapeutic dynamics. The differences identified do not indicate classification errors, but rather differential interpretative sensitivities that enrich the multidimensional understanding of the therapeutic process. The inability to include GPT-4 in the comparative analysis underscores the importance of methodological validation when applying generative tools to clinical analysis.

4.5. Discussion

The studies presented in this dissertation converge on a central claim: psychotherapy leaves detectable traces in language, and these traces can be modelled to describe both what changes in the patient's narrative (themes, emotions, relational focus) and how change unfolds over time (non-linear trajectories, clustering, transitions, role asymmetries). Taken together, the results support the methodological value of integrating process-oriented clinical theory with computational approaches. Topic modelling and trajectory analyses capture the reorganization of thematic content across sessions; sentiment modelling adds a complementary layer by quantifying shifts in emotional salience; and prompt-based generative systems offer a new route to interpretive modelling that is closer to clinical reasoning, especially when discourse is context-dependent and pragmatically nuanced. In this sense, the chapter provides a coherent empirical picture of therapeutic change as a dynamic reconfiguration of meaning and affect, rather than as a linear reduction of symptoms alone.

At the same time, several general limitations must guide interpretation. First, all studies rely on a single-case corpus within a specific therapeutic orientation, which limits generalizability. Second, transcript-based analyses necessarily reduce psychotherapy to its verbal surface: nonverbal and paralinguistic cues (prosody, timing, gaze, embodied affect) are only partially preserved, even with detailed transcription conventions. Third, computational outputs remain sensitive to analytic choices – including segmentation into turns or blocks, preprocessing decisions (*stopwords*,

lemmatization/stemming), parameter settings, and the definition of topic taxonomies or emotion categories. Relatedly, topic models may conflate adjacent themes or fragment coherent themes depending on data size and model assumptions. In contrast, sentiment models depend on label mappings that may privilege some affective domains over others. Fourth, the validity of interpretation requires caution: statistical patterns (e.g., topic decline, emotion peaks, transition matrices) may index clinically meaningful dynamics, but they do not automatically explain mechanisms of change without triangulation with qualitative reading and outcome measures.

Finally, the introduction of generative AI adds specific constraints. Prompt-based modelling can increase contextual sensitivity, but it raises issues of replicability, protocol adherence, and dependence on prompt engineering; moreover, these tools must be deployed within rigorous ethical standards for sensitive clinical material. For these reasons, the contribution of this chapter should be understood as methodological and exploratory: it demonstrates the feasibility and utility of computational approaches for capturing psychotherapy dynamics, while indicating the need for replication on larger, multi-orientation datasets and for clinical validation against independent outcome indicators.

This general framework contextualizes the three thematic discussions that follow, which examine 1) topic dynamics as markers of narrative transformation, 2) the modelling of emotional change across sessions, and 3) the methodological transition from topic modelling to generative AI systems.

4.5.1. Topic Dynamics in Psychotherapy

The results presented in Chapter 4.1 confirmed the study's first hypothesis, namely that topic trajectory analysis can serve as a reliable indicator of therapeutic progress. The application of GAMMs proved particularly effective in capturing non-linear trends in the co-occurrence of key topics across sessions (Appendix 5), thus confirming the sensitivity of this method in detecting changes in the patient's narrative that are often overlooked by more static models (Zuur et al., 2009; Wood, 2017). The results showed a marked decline in the topic of Suffering/Pathology/Symptoms (SPS), which paralleled the patient's clinical improvement and reduction in existential distress. At the same time, the increase in the Therapeutic Refiguration and Insight (TRI) topic reflected a growing capacity for cognitive and emotional reorganization, consistent with successful therapeutic outcomes. These dynamics are not linear but recursive, being influenced not only by the patient's evolving narrative but also by contextual variables such as therapist interventions and external life events. While GAMMs can effectively model these trends, they cannot fully account for the bidirectional and recursive nature of psychotherapeutic processes. Future research could address this limitation by

employing advanced dynamic modelling techniques, such as recurrent neural networks or dynamic systems models, which may better capture the complexity of these interactions.

The second hypothesis, that specific co-occurrences between nuclear topics would also be indicative of therapeutic progress, was likewise supported. In particular, the co-occurrence between TRI and Social Relations (SR) increased over time (Appendix 4), suggesting that as the patient developed deeper insights into her relational patterns, her ability to build and sustain meaningful social connections improved. This is consistent with the idea that therapeutic gains extend beyond the clinical setting to foster improved social functioning and resilience (Helm & Jones, 2019; Atzil-Slonim et al., 2021). A further noteworthy finding was the decreasing co-occurrence of SPS with relational and occupational topics in the later phases of therapy, signalling a shift from symptom-driven responses toward more adaptive and resilient modes of functioning (Liccione, 2019). The decline in SPS in relation to work and relationships thus emerged as a crucial indicator of therapeutic progress and was coherent with the non-linear trends identified through GAMMs (Appendix 4).

As therapy progressed, TRI, initially relatively isolated, became increasingly intertwined with broader life contexts, including Projectuality (PRJ) and SR. This pattern suggests that the patient gradually acquired the ability to embed therapeutic insights into daily life, in line with experiential tasks used by the therapist to encourage social engagement (Liccione, 2019; Fuchs, 2021). The results of the LDA analysis further validated these findings. The algorithmically generated topics mirrored those identified by expert psychotherapists, particularly SPS and TRI (Appendix 5), while also uncovering more nuanced aspects of the therapeutic process, such as Affective Relationships (AR) and SR, which became more prominent as therapy progressed.

These results were reinforced by qualitative methods, such as word-overlap analysis. In the initial sessions (1–8), there was significant overlap between SPS and AR, indicating that emotional suffering was strongly bound to relational issues. For instance, words like "problem" and "issue" were frequent across both topics. In later sessions (19–26), these overlaps diminished, and new lexical fields emerged, such as plan and future in PRJ, while terms associated with distress, like anxiety, remained confined to SPS. These qualitative impressions were corroborated by Topic Correlation metrics. For example, the correlation between SPS and AR decreased from 0.85 at the start to 0.38 in the final stage, and the correlation between SPS and Family (FAM) dropped from 0.78 to 0.15. Such results provide quantitative evidence of the patient's movement away from entanglement with relational suffering toward themes of personal growth and reorganization.

Taken together, the convergent evidence from GAMMs and LDA provides a comprehensive view of the therapeutic process. Topic trajectories not only reflect general therapeutic improvement but also

reveal specific dynamics, such as symptom reduction (SPS) and the development of insight (TRI), that are central to understanding psychotherapeutic change. The increasing co-occurrence of TRI with SR and PRJ highlights broader therapeutic gains that extend beyond the clinical setting to improved social engagement and personal development (Atzil-Slonim et al., 2021; Helm & Jones, 2019). In this sense, the study confirms that topic analysis, when combined with nonlinear statistical methods, offers a robust framework for capturing the complexity of therapeutic change in a single case

The results presented in Chapter 4.2, the second study of this dissertation, presented in subchapter 4.2, advance the understanding of psychotherapeutic processes by combining sequential perspectives with contemporary topic modelling techniques. Drawing on the nuclear topics identified by Liccione and Siciliano (2025, see Chapter 4.1) as recurring anchors across psychotherapies – such as Symptom Reduction (SPS), Therapeutic Refiguration and Insight (TRI), Relational Themes (RA), and Occupational/Study (OS) – the study investigated whether computational models could (a) reliably reproduce these core domains and (b) detect emergent, case-specific topics that capture the patient's individual concerns. Such emergent themes, including context-specific anxieties, situational stressors, or existential motifs, may play a pivotal role in the therapeutic process yet often remain invisible to classical coding frameworks or sequential analyses.

The principal contribution of this work lies in the methodological comparison of three topic modelling algorithms – Latent Dirichlet Allocation (LDA), BERTopic, and Top2Vec. Each of these approaches brings distinct advantages: LDA offers a robust probabilistic framework for identifying broad, recurrent patterns; BERTopic, by leveraging transformer-based embeddings, captures more nuanced contextual distinctions and the presence of topic satellites; and Top2Vec, by clustering semantically coherent vectors without prior assumptions, is often capable of revealing unexpected or novel themes. Applied to the same psychotherapy corpus, these models allow for a systematic evaluation of their relative ability to both reproduce established nuclear topics and uncover new, case-specific dimensions of therapeutic dialogue.

Finally, LDA, as an AI-driven tool, would enhance our understanding of latent topics that are not immediately identifiable. The results validated this hypothesis, as the LDA uncovered several themes that were more specific than those identified by the expert judges and listed in Table 3. These topics offer deeper insights into the patient's experiences, capturing details that may have been overlooked during manual topic assignments.

For instance, in the middle phase of therapy, a distinct LDA-derived topic emerged related to the patient's academic and professional concerns. The topic was cantered around words such as "exam," "study," and "career," likely reflecting the patient's anxiety surrounding the state examination she

mentioned frequently. Such specific concerns, although part of the broader Occupation/Study category (OS), are significant markers of personal distress and may indicate key stressors affecting the patient's psychological state. Similarly, topics related to social relations in later sessions appeared more nuanced in the LDA results, distinguishing between general social ties (SR) and more specific interactions with close friends, colleagues, or even peers from educational settings.

In summary, LDA provided additional granularity in the analysis, revealing subtler, context-specific themes that enabled a more personalized understanding of the patient's concerns. This highlights the potential of AI-driven tools, such as LDA, to enhance the sensitivity and precision of psychotherapeutic analysis, providing a detailed understanding of key personal experiences that can inform tailored therapeutic interventions.

Taken together, these two studies offer a conceptual and methodological advance: they bridge sequential discourse analysis with contemporary natural language processing, providing a framework to capture not only the structure of discourse but also the content-specific dynamics of therapeutic change. This dual approach has direct implications for psychotherapy process-outcome research and for the development of clinical tools that provide therapists with real-time, fine-grained feedback on patient narratives.

4.5.2. Tracing Emotional Change across Psychotherapy sessions

The study presented in 4.3 is situated within the framework of Process–Outcome Research (POR), which systematically investigates how procedural dynamics unfolding during psychotherapy sessions relate to clinical outcomes. POR aims to identify how emotional, cognitive, and relational components contribute to therapeutic change, thereby providing an empirical basis for optimizing interventions (Gelo et al., 2015). It is crucial to examine the temporal evolution of these components using methods that capture their complexity across sessions.

In this perspective, natural language processing (NLP) techniques represent a particularly promising methodological frontier. Since Freud, language has been considered a privileged means of accessing unconscious content and intrapsychic dynamics (Smink et al., 2019). Contemporary NLP technologies extend this tradition by enabling the systematic analysis of large corpora of therapeutic text using increasingly sophisticated machine-learning and deep-learning models (Manning & Schutze, 1999; Imel et al., 2015). Recent work has shown that NLP can successfully contribute to the identification of specific psychopathological conditions (e.g., schizophrenia, bipolar disorder), the monitoring of psychotic symptoms, and the analysis of patient-therapist relational patterns (Corcoran

et al., 2018; Alonso-Sanchez et al., 2022; Tsakalidis et al., 2021). The application of sentiment analysis to psychotherapeutic transcripts remains relatively underexplored, despite growing evidence of its clinical potential (Denecke & Reichenpfader, 2023; McCarthy et al., 2017).

In the present study, we applied RoBERTa-GoEmotions, an advanced deep learning model for sentiment analysis, to examine the emotional evolution of a single psychotherapy case across 26 sessions. Each transcript was segmented into 512-token units, and emotion probabilities were then aggregated to yield session-level scores for each emotion category. Descriptive and correlational analyses of these scores revealed distinct patterns in the trajectories and interrelations of emotions over time, which are discussed below considering existing POR literature and our previous single-case findings.

In line with our initial hypothesis, the RoBERTa-GoEmotions model allowed us to derive quantitative indices of the emotions expressed in the therapy transcripts and to describe their evolution across sessions. Joy and neutrality emerged as the most frequent emotional categories throughout treatment (Appendices 8.1-8.3), with relatively high and stable median values across phases. The stability of joy, despite some fluctuations, suggests that positive affect constituted an emotional “baseline” rather than manifesting as isolated peaks of intense positivity. Neutrality, in turn, showed limited variability and remained consistently prevalent, which may reflect moments of emotional regulation and phases of reflective, cognitively oriented work in ongoing psychotherapy.

However, the predominance of joy over the other categories must also be interpreted in light of the methodological choices made. To render the output clinically interpretable, the original 27 GoEmotions labels were grouped into six categories based on Ekman’s model. In this mapping, joy is the broadest label, encompassing a wide range of positive emotions (admiration, amusement, approval, concern, desire, excitement, gratitude, joy, love, optimism, pride, relief). In contrast, disgust is a single category, and fear includes only fear and nervousness. This uneven grouping likely contributed to the relatively low levels of fear and disgust observed in our results (Appendices 8.1, 8.2, and 8.5). Consistently, anger, sadness, and surprise – whose aggregated labels incorporate a comparable number of original emotions (anger + annoyance + disapproval; sadness + disappointment + embarrassment + pain + remorse; and the “ambiguous” emotions confusion + curiosity + realization + surprise) – showed similar frequencies over the course of therapy (Appendices 8.1-8.4). At the same time, it is reasonable to hypothesize that fear and disgust are harder to capture for a purely text-based model or are simply less central in this case and in the linguistic surface of psychotherapy.

In contrast to the relative stability of joy and neutrality, anger and, to a lesser extent, sadness

followed a more discontinuous pattern, with marked fluctuations and specific peaks of intensity during the therapeutic process (Appendices 8.1, 8.2, and 8.4). These oscillations may mirror shifts in the patient’s functioning, with episodes of heightened negative emotionality alternating with phases of greater regulation and integration. Notably, joy exhibited a “U-shaped” temporal pattern, with lower median values in the middle phase and higher values in the initial and final phases (Appendices 8.2 and 8.3, left panel). In contrast, anger and sadness showed their maximum intensities precisely in the middle phase (Appendices 8.2 and 8.4). This configuration suggests that the central phase of treatment may have represented a clinically critical period, characterized by the emergence and working-through of emotionally salient material, with a concomitant temporary reduction in positive affect. The correlation structure further supports this reading: anger and sadness were inversely associated with joy (Table 8; Appendices 8.6-8.8), indicating that increases in negative emotional intensity tended to coincide with decreases in positive affect, especially in those central sessions where anger and sadness reached their highest levels. Qualitatively, this co-occurrence is reflected in mid-treatment narratives marked by pervasive anger and high internal pressure. For instance, the patient described being “very, very angry... angry at the world” while simultaneously emphasizing the need for “peace of mind” in an emotionally and cognitively taxing period (session 14). In the same treatment window, sadness is articulated through themes of depletion and isolation (“it’s such a heavy time... I’m having a sense of loneliness... Otherwise, the desert, nothingness”), suggesting an affective context in which positive states are harder to access and maintain (session 18). More generally, when anger escalated the patient also reported a reduced capacity to resonate with others’ happiness (“I see everybody... they’re happy and happy...,” followed by “I’m kind of turned off, I don’t feel like it anymore”), capturing a subjective dampening of positive affect in proximity to negative affect activation (session 9).

Furthermore, consistent with our second hypothesis, Spearman’s correlation analysis revealed significant negative associations between joy and the emotions of anger and sadness (Table 8; Appendices 8.6 and 8.8). These results can be interpreted through complementary methodological and clinical lenses. From an instrumental standpoint, the negative correlations may partly reflect the operational characteristics of the RoBERTa–GoEmotions model and the way emotions are assigned to text segments. At the linguistic level, segments expressing anger are typically semantically and syntactically distant from segments expressing joy. Transformer-based models such as RoBERTa–GoEmotions learn word–context associations; they are therefore less likely to assign a positive emotion when lexical and contextual indicators of negative affect (e.g., anger, sadness) are prominent. In practice, the sentiment analysis algorithm – trained on datasets where annotators often assign a single prevalent label – tends to favor a single dominant emotion in emotionally dense or ambiguous

passages. This increases the likelihood that joy will be downweighted when anger or sadness is strongly expressed, thereby reinforcing the inverse relationships observed in our data.

From a clinical-theoretical perspective, however, these correlational patterns take on a deeper meaning, especially when viewed through the lens of Emotion-Focused Therapy (EFT). Within EFT, the experience and integration of conflicting emotions are central mechanisms of therapeutic change, and greater integration is viewed as a favorable prognostic indicator (Greenberg, 2012). The negative correlations between joy and the two adverse effects could therefore also signal limited integration of positive and painful emotional experiences, suggesting a tendency toward emotional compartmentalization rather than fluid, co-existing affective states. In this sense, our findings may reflect not only model behavior but also an underlying difficulty in achieving more flexible and adaptive emotional regulation.

In line with this interpretation, positive correlations were observed among disgust, anger, and sadness (Table 8; Appendices 8.6 and 8.7), delineating a cluster of negative emotions that appears to co-activate across sessions. These associations may index specific thematic or narrative configurations in the patient's experience – such as moments of self- or other-directed evaluation, criticism, or moral judgment – in which anger, sadness, and disgust are simultaneously mobilized. The emergence of such emotional configurations suggests the presence of shared “hot spots” in the therapeutic discourse, where complex evaluative and judgmental dimensions of psychological functioning are activated and potentially reworked within the therapeutic relationship. Consistent with this, in an advanced session the patient explicitly described a rapid oscillation between sadness and anger in response to the partner's post-separation condition (“it makes me very sad...”, followed by “But then the anger comes in and you say: ‘Ok, but you did it with your own hands, you asked for it and now you have to take the consequences’”), (session 24). In another late-session narrative, evaluative language becomes overtly moralized and contempt-laden – e.g., the patient reports being called “a whore” and recounts the mother's escalating devaluation (“she's even worse than a whore”), (session 25) – a discursive configuration that is clinically compatible with disgust-like appraisals co-occurring with anger activation and relational pain.

A particularly salient finding concerns the peaks of emotional intensity concentrated in Session 23, where anger, sadness, and disgust simultaneously reached their highest values (Appendix 8.1). This temporal convergence of multiple negative emotions – also reflected in the correlation structure among these effects (Table 8; Appendices 8.6 and 8.7) – suggests that Session 23 functioned as a focal point within the therapeutic process, marked by an intensification of emotional processing and likely by the activation of clinically central themes. Qualitatively, we can read that the session

23 is dominated by an acute rupture narrative centered on betrayal and “contamination”-type appraisals, where the patient repeatedly frames the partner’s sexualized online behavior as “disgusting” and “sickening” (e.g., “It makes me too sick... it’s just too disgusting”), while simultaneously escalating into confrontational anger and moralized demands for accountability (e.g., “you accept the consequences... it just pisses me off even more”). Notably, the transcript also documents a strong somatic loading (nausea, inability to eat, weight loss, nightmares, stomach distress), indicating that this emotional convergence was not only verbally elaborated but also embodied (“I’d put a morsel in my mouth and I’d get nauseous... before I slept three [hours] then I woke up with nightmares”). Within the same session, anger appears to be used deliberately as a regulatory strategy to counteract emerging nostalgia/melancholy (“I’m banking on anger”), which may help explain how anger, sadness, and disgust are simultaneously mobilized within a high-arousal therapeutic “hot spot.”

Identifying such peak moments of emotional intensity may represent an essential contribution to process research, as it allows the detection of sessions that could be considered “hot spots” for therapeutic change and thus merit closer qualitative and quantitative examination. From a methodological perspective, it would be particularly valuable in future studies to link these peak sessions with additional data sources, such as concurrent symptom monitoring using standardized self-report scales or clinician-rated measures. This would allow investigation of whether marked shifts in the emotional profile of sessions are temporally associated with changes in symptom severity or other clinically relevant outcomes.

4.5.3. From Topic Modelling to Generative AI

As shown in the study presented in Chapter 4.4, this thesis examines the methodological transition in psychotherapy process research from traditional, predominantly manual approaches to increasingly sophisticated computational strategies. Early discourse-analytic traditions – often grounded in lexical frequency counts and human coding – made it possible to describe thematic regularities and conversational structures. Still, they were constrained by scalability and by the difficulty of systematically tracking complex topic dynamics across time. The introduction of statistical topic models, such as LDA and related latent semantic approaches, represented a decisive methodological shift by enabling the automated detection of recurring thematic structures in therapeutic dialogue. Building on this trajectory, the more recent emergence of generative artificial intelligence suggests further opportunities: instead of relying solely on bag-of-words distributions, generative systems can interpret language in context, potentially supporting a more clinically nuanced reading of therapeutic discourse. Within this framework, the study in Chapter 4.4 directly compared

a traditional probabilistic model (LDA) with a prompt-based generative system (Claude 3.5 Sonnet) to evaluate whether generative AI can function as a valid alternative – or a complement – to established computational methods. Overall, the findings offered partial support for this hypothesis. Concordance analyses indicated a moderate alignment between the two approaches, with Jaccard overlap values ranging from 47.56% to 57.22% across the four temporal blocks (average $\approx$ 52%). While these values are not “high” in absolute terms, they should be interpreted in light of the intrinsic characteristics of psychotherapy talk: therapeutic discourse is polysemous, context-dependent, and characterized by rapid shifts, implicit meanings, and pragmatic functions that are not always reducible to lexical content. The observed mismatch between Jaccard concordance and lower rates of exact agreement further reinforces this point, suggesting that the two approaches often converge at the level of “semantic neighborhood” rather than at the strict level of one-to-one topic assignment.

Crucially, the study highlights that divergences between methods do not necessarily indicate error; rather, they often reflect clinically meaningful interpretative differences. Confusion-matrix patterns showed systematic overlaps between thematically adjacent categories (e.g., Ways of Being and Therapeutic Relationship, or Affective Relationships and Suffering/Pathology/Symptoms). These “cross-classifications” are best understood as traces of the natural interdependence of patient experience, relational narratives, and symptomatic expression. In this sense, both models appear to capture valid aspects of therapeutic content, but they do so through different lenses: LDA tends to represent discourse through distributions of salient lexical items, whereas generative AI tends to integrate utterances into a more contextually coherent interpretation.

A particularly informative aspect of the comparison concerns how each method handles ambiguity and short conversational units. Brief turns, fillers, and “micro-utterances” (including jokes or minimal responses) systematically affected agreement. Claude more frequently interpreted these segments in continuity with the surrounding context, whereas LDA more often relegated them to a generic residual category (e.g., OTH). This difference suggests a potential advantage of generative systems in psychotherapy corpora, where meaning is often distributed across adjacency pairs and where local coherence may be more informative than isolated lexical content. At the same time, this advantage depends strongly on prompt design and on the clarity and operationalization of topic definitions, making standardization a key methodological priority.

Longitudinal analyses provided further evidence that both approaches are sensitive to clinically relevant change processes, although they may emphasize different temporal patterns. Topic trajectories showed meaningful evolution across therapy blocks, including changes in the distribution of symptom-related material (SPS) and in the presence of therapist-intervention topics (e.g., TRI).

Both models captured systematic role asymmetries: therapist turns were more strongly characterized by therapeutic-process categories, while patient turns showed greater prominence of suffering- and experience-related topics. Such asymmetries are theoretically expected in psychotherapy and thus support the ecological validity of the computational outputs. Moreover, the temporal co-evolution of these patterns – such as increasing intervention-related content in the therapist and decreasing symptom-centered content in the patient – was interpreted as compatible with positive therapeutic progression. However, causal claims remain beyond the scope of a single-case design.

Beyond concordance, the study also points to a substantive methodological insight: topic transitions are not merely technical artifacts but can be clinically informative indicators of how therapy organizes meaning. Differences in transition structure: Claude’s stronger contextual linking between relational and self-experience categories versus LDA’s more uniform distribution across transitions. This suggests that generative systems may be especially suited to capturing the “narrative glue” of psychotherapy: the integrative moves through which experiences, relationships, and interpretations are woven together in dialogue. Conversely, LDA’s strength lies in its stability and reproducibility, providing a robust baseline representation of recurrent thematic architecture.

At the same time, the study underscores that not all generative systems are equally adequate for structured clinical research tasks. The attempted inclusion of another generative model revealed critical protocol violations, including deviations from required formats and the introduction of unauthorized categories. This finding is central for the thesis’ general argument: the usefulness of generative AI in psychotherapy research is not a generic property of “AI,” but a function of model behaviour under constraint, methodological discipline, and replicable prompting protocols.

These results must be interpreted within apparent limitations. The empirical evidence derives from a single psychotherapy case within a specific theoretical orientation, restricting generalizability to other modalities and clinical populations. The topic taxonomy, while clinically informed, may not cover the thematic richness encountered across diverse therapies. Moreover, the analyses primarily focused on quantitative concordance and transition dynamics; deeper qualitative validation – linking extracted topics to clinically grounded interpretations – remains an essential next step. Finally, the dependence of generative approaches on prompt engineering raises reproducibility challenges, reinforcing the need for standardized prompt protocols and transparent reporting.

Taken together, the thesis supports a central conclusion: the methodological future of therapeutic discourse analysis is likely not a replacement of traditional models with generative systems, but an integration of complementary approaches. Mixed pipelines that combine the statistical robustness of LDA with the contextual sensitivity of generative AI may offer the most

promising route toward a richer representation of psychotherapy processes. Future developments should prioritize (a) multi-case, multi-orientation corpora; (b) standardized prompting and benchmarking procedures; (c) clinical validation through correlations with outcome measures; and (d) rigorous ethical frameworks for managing sensitive clinical data. Extending these approaches to multimodal signals (prosody, gesture, and physiological measures) further represents a coherent next frontier for psychotherapy.

In conclusion, as illustrated by the study in Chapter 4.4, the convergence between topic modelling and generative AI marks a significant evolution in psychotherapy methodology. This convergence opens new possibilities for capturing therapeutic change as a dynamic reorganization of themes, roles, and transitions – supporting both scientific understanding and, with due caution, future applications for clinically meaningful process monitoring.

Section 2 – AI-based clinical application: PSICOGENIUS^{AI}

This second section documents the work conducted in collaboration with Magellano S.c. (Society social cooperative), based in the Mi.CAL - Milan Institute for Health Care and Advanced Learning. Beginning in my second year of doctoral studies, I began translating insights from the literature into practical applications to support clinical practice.

The project was developed in close collaboration with psychotherapists with more than ten years of professional experience and a team of software engineers, which enabled the design of not only advanced AI-based analytic tools but also a comprehensive software solution that integrates practice-management functionalities essential for mental health professionals. My primary contribution was to identify, based on research conducted at the University of Pavia and analysis of psychotherapy transcripts, the tools most relevant to psychotherapists and to develop tailored prompting strategies that align generative AI outputs with the actual needs of clinical work.

The overarching goal was to provide specialists with AI-driven tools that are both controlled and profession-specific, thereby reducing the biases typically associated with probabilistic systems and ensuring greater interpretability and reliability. At the same time, the software was designed to consolidate clinical, analytic, and administrative tasks into a coherent digital environment, enabling practitioners to streamline their work within a unified platform.

The research presented in the first part of this dissertation demonstrates how integrating psychotherapy and Artificial Intelligence (AI) can open new perspectives for analysing and understanding clinical processes. Starting from classical statistical models (e.g., GAMMs, LIWC, etc.) and text mining techniques such as topic modelling (LDA, BERTopic, Top2Vec) and moving toward the implementation of state-of-the-art generative systems (LLMs), it has been possible to observe how semantic changes in psychotherapy transcripts mirror profound transformations in patients' ways of experiencing and narrating their lives. The results obtained throughout this doctoral project have not only provided theoretical and methodological advances but also laid the foundation for the design of a practical tool: PSICOGENIUS^{AI}.

PSICOGENIUS^{AI} was conceived to translate research findings into concrete instruments to support psychotherapeutic practice. The software is based on two principles: on the one hand, to capitalize on advances in AI in transcription, translation, and automated text analysis; on the other, to preserve the irreplaceable role of the therapist in interpreting data and providing clinical feedback. Its central idea

is that AI can act as a “helpful ally” capable of accelerating complex operations (such as transcription, topic extraction, sentiment analysis, and session summarization) and offering new exploratory perspectives, while never replacing the clinical judgement and hermeneutic expertise of the psychotherapist. In this sense, PSICOGENIUS^{AI} represents a paradigmatic example of how academic research can generate operational outcomes. The various tools integrated into the software are directly derived from the reflections and results presented in Chapter 4, ranging from topic modelling algorithms to trace the dynamics of therapeutic themes, to metrics for analysing semantic shifts, to experiments with generative models to evaluate their reliability in clinical contexts. These components have been adapted and re-elaborated with attention to the real needs of psychotherapist.

The following section outlines the path leading to the creation of PSICOGENIUS^{AI}, illustrating its underlying rationale and potential clinical applications. Outputs will be presented at both the individual-session level and the overall therapeutic process level, with a specific focus on the roles of prompt engineering and human-machine interaction. Finally, this section will reflect on the significance of conceiving AI as an ally rather than a replacement, and on the future directions for the development of PSICOGENIUS^{AI} and its integration with new analytic and clinical tools.

1. From research to the development of PSICOGENIUS^{AI}

The convergence between psychotherapy and Artificial Intelligence (AI) represents one of the most promising developments in contemporary clinical research and practice. Psychotherapy has traditionally relied on both qualitative and quantitative methods to investigate therapeutic processes and outcomes. In recent decades, advances in Natural Language Processing (NLP) and machine learning have progressively extended into the mental health domain – first in medical diagnostics, and more recently in the field of psychotherapy. This evolution reflects a broader socio-technological trend in which AI is increasingly integrated into professional contexts, while maintaining the need for critical and ethical frameworks that safeguard the human dimension of care (Topol, 2019; Zhang & Wang, 2024; Bloch-Atefi, 2025).

The convergence between psychotherapy and Artificial Intelligence (AI) represents one of the most promising developments in contemporary clinical research and practice. Psychotherapy has traditionally relied on both qualitative and quantitative methods to investigate therapeutic processes and outcomes (Atkins et al., 2014; Llewelyn, Macdonald & Aafjes-van Doorn, 2016). In recent decades, advances in Natural Language Processing (NLP) and machine learning have progressively extended into the mental health domain – not only in medical diagnostics (Topol, 2019) but also in

psychotherapy research and practice (Bedi et al., 2015; Imel et al., 2015; Zhang & Wang, 2024). This evolution reflects a broader socio-technological trend in which AI is increasingly integrated into professional and clinical contexts, while maintaining the need for critical and ethical frameworks that safeguard the human dimension of care (Atkins et al., 2014; Topol, 2019; Russell & Norvig, 2021; Zhang & Wang, 2024).

Within this context, the doctoral research presented in the first section of this dissertation has sought to bridge hermeneutic-phenomenological approaches to psychotherapy with computational methods for text analysis. Techniques such as Latent Dirichlet Allocation (LDA) and Generalized Additive Mixed Models (GAMMs) have been employed to identify thematic structures within psychotherapy transcripts and to track their temporal trajectories, providing empirical evidence that semantic shifts in patient narratives are closely linked to therapeutic change (Liccione & Siciliano, 2025). The notion of semantic shift – the transformation of meaning within the patient’s discourse throughout therapy – emerges as a central construct, enabling the quantification of processes that had previously been accessible only through qualitative interpretation.

Building on these findings, the challenge is not merely methodological but also translational: how to translate research results into tools that directly support clinical practice. The rationale for PSICOGENIUS^{AI} arises precisely from this translational ambition. Rather than designing a system to replace therapists, the aim is to create a digital ally that enhances clinical reflection, supports supervision, and provides therapists with structured yet flexible tools to navigate the complexity of therapeutic processes. By integrating topic modelling, sentiment analysis, and generative AI systems, PSICOGENIUS^{AI} embodies the shift from research outcomes to practical applications, while preserving the interpretative role and ethical responsibility of the clinician.

The development of PSICOGENIUS^{AI} is deeply rooted in the empirical findings obtained during the doctoral research project. The studies presented in Chapter 4 illustrate a progressive trajectory: from the transcription and translation of a complete psychotherapy to the application of advanced topic modelling algorithms, and to the exploration of generative systems for text analysis.

The first step was the systematic analysis of topic trajectories using LDA and GAMMs. These methods demonstrated that specific therapeutic topics – such as Suffering/Pathology/Symptoms (SPS) and Therapeutic Refiguration and Insight (TRI) – tend to evolve in consistent ways across the course of therapy, with a decrease in symptom-related discourse and an increase in reflective or project-oriented language. Such findings confirmed that computational approaches can capture clinically meaningful patterns of change, thereby offering therapists a quantitative framework to complement qualitative expertise (Liccione & Siciliano, 2025).

A second line of research examined semantic shift, emphasizing not only the presence of specific topics but also the transformation of their lexical and semantic content over time. By comparing models such as LDA, BERTopic, and Top2Vec, the research revealed that generative approaches can uncover finer-grained, context-specific themes that may remain invisible to manual coding or classical statistics (see Chapter 4.2 in Section 1). These insights directly informed the design of modules within PSICOGENIUS^{AI} dedicated to topic evolution analysis and semantic monitoring, providing clinicians with visual and textual representations of changes across sessions.

A further step involved experimentation with generative AI systems. Leveraging state-of-the-art large language models (LLMs), it became possible to test the reliability of generative topic assignments compared to human coding. These results underscored both the potential and the limitations of generative systems: while capable of accelerating transcription, translation, and initial coding, they also require careful prompt engineering and expert supervision to ensure interpretative validity. This methodological awareness was integrated into PSICOGENIUS^{AI} by embedding a human-in-the-loop design, in which AI-generated outputs are conceived as aids rather than definitive interpretations.

Finally, the cumulative evidence suggested the feasibility of building a composite platform where different analytic modules could converge, including transcription, topic modelling dashboards, semantic shift trackers, and session summarization tools. Each of these features reflects specific research findings and has been adapted into a clinical tool. Therefore, PSICOGENIUS^{AI} is not just about applying AI techniques but about implementing an entire research program, turning process-outcome evidence into practical insights for clinical use.

The design of PSICOGENIUS^{AI} was guided by a set of theoretical, methodological, and ethical principles that distinguish it from general-purpose AI applications. Rather than being conceived as a diagnostic or prescriptive system, PSICOGENIUS^{AI} was deliberately structured as a clinical support tool, grounded in empirical research and attuned to the complexity of psychotherapeutic practice.

First, the software embraces the principle of AI as an ally, not a replacement. The central role of the psychotherapist remains unchanged: clinical judgement, narrative interpretation, and relational expertise cannot be substituted by algorithms. PSICOGENIUS^{AI} provides complementary insights derived from computational text analysis (topic trajectories, word clouds, or clinical anamneses and summaries) that can stimulate reflection and enrich supervision, while leaving interpretation to the clinician. In this sense, the AI acts as an “artificial third” within the therapeutic triangle, potentially amplifying observation and insight without replacing the therapist's specificity or undermining professional autonomy (Haber et al., 2024).

Second, the platform's architecture was designed in accordance with ethical and regulatory standards, with a particular emphasis on data security and patient confidentiality. In line with the General Data Protection Regulation (GDPR; EU Regulation 2024/1689), all sensitive data processed within PSICOGENIUS^{AI} is encrypted, anonymised where possible, and stored in secure environments. Unlike consumer-grade generative AI systems, which often rely on external cloud infrastructure and retain user interactions for model training, PSICOGENIUS^{AI} ensures that only authorized professionals can access clinical data, thereby preserving trust in the therapeutic setting.

Third, the system incorporates a human-in-the-loop design philosophy that acknowledges both the strengths and limitations of AI. Outputs generated by topic modelling or large language models are treated as provisional hypotheses that require expert evaluation. This design mitigates the risk of overreliance on probabilistic outputs, preventing the passive delegation of interpretive authority to the machine. Instead, the clinician's hermeneutic expertise is explicitly required to contextualise and validate the AI's suggestions, ensuring that the technological layer supports rather than constrains clinical reasoning (Zhang & Wang, 2024).

Finally, the platform was shaped by a commitment to usability and integration. Time pressures and administrative demands often constrain clinical practice; therefore, PSICOGENIUS^{AI} was designed to streamline workflows without adding complexity. The interface provides intuitive visualisations and dynamic summaries, enabling therapists to quickly access analytical outputs for supervision, patient feedback, or research. By combining methodological rigour with practical accessibility, PSICOGENIUS^{AI} aims to operationalise the insights of process–outcome research into tools that are both scientifically robust and clinically meaningful.

2. From session to insight: how and why is it useful?

Psychotherapy, at its core, is a linguistic and narrative process. The therapeutic encounter unfolds primarily through words – spoken, heard, and interpreted – making the transcript of a session not merely a record, but a privileged window into the dynamics of change. Textual material enables the systematic observation of both content (what is being discussed) and form (how it is being expressed), offering researchers and clinicians an indispensable tool for tracing the unfolding of therapeutic work.

Recent developments in computational linguistics have expanded the possibilities of transcript analysis. By applying probabilistic models such as Latent Dirichlet Allocation (LDA) or more recent neural and generative approaches (Blei, Ng & Jordan, 2003), it has become possible to move beyond

purely qualitative interpretations and quantify recurring themes, semantic shifts, and discourse trajectories over time. This methodological innovation has allowed researchers to operationalise constructs that were once considered too elusive to measure, such as the semantic shift – that, as we said, is the transformation in meaning structures within the patient’s narrative over the course of therapy. As demonstrated in the previous chapters, semantic shifts are not only detectable but also correlate with therapeutic improvement, thus providing a measurable index of psychological change.

For therapists, however, the value of transcript analysis extends not only to academic research but also to clinical utility. Having access to syntheses of individual sessions and to longitudinal maps of thematic development can help clinicians maintain perspective across multiple encounters, facilitate supervision, and support the writing of clinical case reports. Yet manual transcription and coding are time-consuming and resource-intensive. It is here that AI-driven tools, such as PSICOGENIUS^{AI}, become indispensable, offering automated yet clinically interpretable ways to transform raw session data into structured insights.

A crucial distinction in this domain concerns supervised versus unsupervised topic modelling approaches (Blei & McAuliffe, 2007). In supervised systems, such as the modules implemented within PSICOGENIUS^{AI}, the AI is trained on fixed labels corresponding to nuclear topics – broad themes that are theoretically grounded and applicable across all psychotherapies. These nuclear topics, defined and validated through prior hermeneutic and clinical research, serve as a stable interpretative grid. They enable the system to categorize conversational turns under categories such as symptomatology, therapeutic reframing, and insight, or life projects. By contrast, unsupervised approaches allow the model to generate topics directly from the data, often identifying session- or therapy-specific satellite topics. While these emergent themes may capture context-specific nuances, they lack the cross-therapeutic comparability ensured by nuclear topics.

The combination of these two approaches – fixed nuclear topics via supervised learning and flexible satellite topics via unsupervised exploration – represents a powerful methodological synthesis. Supervised modules guarantee reliability and alignment with clinical theory, whereas unsupervised modules provide sensitivity to the idiosyncratic elements of a given case. Together, they enable therapists not only to monitor longitudinal trajectories of change but also to attend to the unique narrative features that emerge in a particular encounter.

In this perspective, PSICOGENIUS^{AI} operationalises a long-standing aspiration in psychotherapy research: to move from session to insight. By transforming transcripts into visualisations, topic distributions, and semantic maps, the platform translates the complexity of clinical dialogue into accessible outputs that remain open to interpretation. Rather than reducing the therapeutic process to

numbers, it augments the therapist’s reflective capacity, providing a structured yet flexible means of integrating empirical evidence with hermeneutic understanding.

One of the central design principles of PSICOGENIUS^{AI} is the distinction between nuclear and satellite topics, reflecting both methodological choices and theoretical commitments in psychotherapy research. This dual-level architecture addresses the need for stability in cross-therapeutic comparisons while maintaining sensitivity to the dimensions of individual cases.

Nuclear topics consist of a fixed set of thematic categories that recur across various forms of psychotherapy and therapeutic environments. They result from years of hermeneutic and empirical research and are formalized in the platform through a supervised learning approach. In practice, this means the AI system is trained using predetermined labels, each linked to a broad conceptual domain such as Suffering/Pathology/Symptoms (SPS), Therapeutic Refiguration and Insight (TRI), or Projectuality (PRJ). Within PSICOGENIUS^{AI}, these labels are presented to users via clear clinical descriptions, as detailed in the reference table included in the platform documentation (see Figure 36). The prompting structures used for model training are designed to reflect these descriptions, ensuring consistency between the clinician’s interpretative framework and the AI’s outputs.

Descrizione Topic Nucleari	
Sofferenza/Patologia/Sintomi	Questa categoria raccoglie tutte le informazioni che riguardano il malessere emotivo, psicologico o fisico del paziente. Include riferimenti alla sofferenza personale, alla sintomatologia psicopatologica o ad aspetti di malessere esistenziale passati e presenti. Si riferisce sia a problematiche specifiche (ad esempio, ansia, depressione, attacchi di panico) sia a manifestazioni generali di malessere che influenzano la vita quotidiana del paziente.
Famiglia	Questa categoria riguarda tutte le informazioni e i significati relativi alla famiglia del paziente, comprese le relazioni con i familiari. Include i riferimenti ai rapporti con genitori, fratelli, sorelle, e in caso di matrimonio, anche con il coniuge e i figli. Vengono considerati sia gli aspetti relazionali, emotivi e sentimentali che il paziente sperimenta nei confronti dei familiari, sia le caratteristiche e i comportamenti dei membri della famiglia.
Relazione Affettiva	Questa categoria comprende tutte le informazioni e i significati relativi ai rapporti affettivi romantici, passati e presenti, del paziente. Include la descrizione di sentimenti, emozioni e relazioni con partner romantici, nonché i modi di essere di questi ultimi. Riguarda sia le relazioni in corso che quelle passate, nonché la percezione che il paziente ha di esse.
Occupazione/Studio	Questa categoria include le informazioni riguardanti l'occupazione lavorativa del paziente o il suo percorso di studio. Comprende anche le relazioni con colleghi, superiori o insegnanti, e le eventuali difficoltà o soddisfazioni legate al lavoro o allo studio. Riguarda i modi in cui il paziente si relaziona al proprio ruolo lavorativo o scolastico e le implicazioni emotive connesse.
Rete Sociale	Questa categoria include le informazioni e i significati relativi alla rete sociale del paziente, incluse amicizie, conoscenti e altri rapporti sociali significativi. Si riferisce alle dinamiche relazionali al di fuori del contesto familiare e affettivo-romantico, esplorando il modo in cui il paziente si rapporta agli altri, le emozioni e i sentimenti che vive nelle relazioni amicali e sociali, e le eventuali difficoltà o soddisfazioni in questo ambito.
Modi Di Essere	Questa categoria riguarda i tratti di personalità, il carattere, il temperamento e le abitudini del paziente che definiscono il suo modo di percepire e interagire con il mondo. Include le riflessioni del paziente sul proprio sé, la sua autostima, il senso di corporeità, il rapporto con il proprio corpo. Riguarda anche la descrizione di emozioni ricorrenti, tendenze emotive e comportamentali che influenzano la sua vita quotidiana.
Progettualità	Questa categoria comprende tutte le informazioni relative ai desideri, progetti e aspettative future del paziente. Include riflessioni su obiettivi personali, professionali, familiari o sociali che il paziente desidera raggiungere. Può riguardare sia pianificazioni concrete che desideri più astratti o ideali, e include anche la percezione del paziente rispetto ai propri progressi o ostacoli nel realizzare tali progetti.
Esperienza Terapeutica	Questa categoria si concentra su come il paziente vive la relazione terapeutica e il processo della terapia. Include il modo in cui il paziente percepisce l'alleanza terapeutica, la fiducia, la comunicazione e le proprie emozioni in terapia. Comprende anche i riferimenti a eventuali compiti o esercizi assegnati dal terapeuta e come il paziente li ha sperimentati. Inoltre, rientrano in questa categoria i feedback del paziente sull'efficacia o i progressi della terapia e riflessioni sul proprio coinvolgimento nel percorso terapeutico.
Rifigurazioni Terapeutiche	Questa categoria si concentra sugli interventi terapeutici, quali rifigurazioni, interpretazioni e riformulazioni proposte dal terapeuta durante le sedute. Include le volte in cui il terapeuta aiuta il paziente a rivedere, ristrutturare o comprendere meglio eventi, emozioni o pensieri. Rientrano in questa categoria i momenti in cui il terapeuta introduce nuove prospettive o significati che portano il paziente a riconsiderare la propria esperienza e quando assegna compiti terapeutici.
Sessualità	Questa categoria include tutte le informazioni e i significati relativi alla sessualità del paziente e al suo orientamento sessuale. Comprende la sfera intima, le esperienze sessuali, i desideri, le fantasie, le preferenze sessuali, le eventuali problematiche o difficoltà sessuali e il rapporto con il proprio corpo e con l'intimità. Rientrano in questa categoria anche i pensieri e le emozioni che il paziente sperimenta rispetto alla propria vita sessuale, oltre a eventuali aspetti legati all'identità sessuale o all'esplorazione del sé attraverso la sessualità.
Benessere	Questa categoria raccoglie informazioni e significati legati a esperienze di benessere, soddisfazione, equilibrio emotivo e serenità del paziente. Include riflessioni su stati d'animo positivi, emozioni gratificanti, momenti di soddisfazione nella vita quotidiana e la percezione di uno stato di benessere generale. Può riferirsi anche a pensieri e comportamenti che promuovono uno stile di vita sano e un atteggiamento ottimista verso la vita e il futuro. Include il modo in cui il paziente vive esperienze di appagamento o serenità in vari ambiti della sua vita.
Sistema Valoriale	Questa categoria riguarda tutte le informazioni e i significati relativi ai valori, principi e credenze che guidano il comportamento e le scelte del paziente. Include i valori morali, etici, religiosi, culturali, familiari e personali. Può riferirsi a come il paziente interpreta e vive le proprie convinzioni, nonché alle tradizioni culturali e religiose che influenzano la sua visione del mondo. Rientrano in questa categoria anche i conflitti interni o esterni legati a valori o credenze, così come la riflessione del paziente su questioni morali o esistenziali.

Figure 36. The Italian description of each nuclear topic labelled and displayed on PSICOGENIUS^{AI} in the Analysis AI section. These descriptions give users context about what the percentage of a specific topic represents. The image is taken from PSICOGENIUS^{AI}.

By contrast, satellite topics are generated through unsupervised approaches that enable the system to extract patterns directly from the data, without relying on predefined labels. These topics are inherently more flexible and context-dependent. For example, in a therapy focused on relational difficulties, unsupervised modelling might surface a session-specific topic, such as “conflict with

partner” or “family expectations.” While such topics may not be transferable across therapies, they capture the unique semantic texture of the individual case.

The complementarity between nuclear and satellite topics mirrors the dual requirements of psychotherapy research and practice: 1) comparability and reliability are ensured through nuclear topics, which provide a stable interpretative grid and allow for the quantification of thematic trajectories across different therapies and patients; 2) contextual richness and specificity are ensured through satellite topics, which highlight emergent issues and nuances that might otherwise be overlooked. Importantly, this architecture also reflects the epistemological stance underlying SICOGENIUS^{AI}: psychotherapy is both a generalizable process in which specific themes and dynamics recur and a particular encounter in which unique meanings and contexts emerge. By combining supervised and unsupervised methods, the platform acknowledges this duality, offering therapists an integrated view that is both rigorous and sensitive to individual needs.

In clinical practice, this distinction enables the therapist to track the evolution of nuclear themes throughout the therapeutic process while also attending to the emergence of satellite themes within individual sessions. For example, a decrease in SPS-related discourse over time may indicate progress in symptom alleviation, while the transient appearance of a satellite topic, such as “marital jealousy,” may signal a momentary but clinically relevant concern. The integration of both levels, therefore, enables PSICOGENIUS^{AI} to function as a dynamic map of therapy, guiding clinicians from the macro level of cross-therapeutic trajectories to the micro level of session-specific insights.

Whereas traditional process–outcome studies have often focused on the frequency of certain categories or the presence of specific interventions, semantic analysis investigates how the *meaning* of discourse evolves. In this perspective, psychotherapy is not only a process of symptom reduction but also a transformation in the patient’s way of narrating and interpreting their lived experience. Semantic shift refers to changes in the lexical and conceptual structures that organise a patient’s discourse. These changes may include the gradual movement from pathology-focused language toward reflective or future-oriented expressions, or the reconfiguration of previously fragmented narratives into more coherent and agentic forms. As shown in the longitudinal topic modelling studies presented earlier, such transformations often coincide with clinical improvement, suggesting that semantic trajectories can serve as proxies for psychological growth.

Importantly, as we stated earlier, the software does not claim to provide definitive judgments about progress but rather offers empirical indicators that therapists can integrate with their clinical intuition. For example, a marked reduction in the overlap of symptom-related terms across sessions may suggest a decrease in pathological rumination; however, it is the therapist who ultimately

contextualizes this signal within the broader narrative of the patient's life. In this way, semantic shift becomes not only a research construct but also a practical ally in clinical decision-making. By highlighting transformations in meaning structures, PSICOGENIUS^{AI} enhances the therapist's capacity to monitor progress, tailor interventions, and provide patients with reflective feedback grounded in both hermeneutic insight and computational evidence.

2.1. Prompt Engineering and PSICOGENIUS^{AI}

The implementation of AI-driven analysis in PSICOGENIUS^{AI} relies on the strategic use of prompt engineering, a methodological practice that has become central to the effective deployment of large language models (LLMs). Prompt engineering is the systematic design and refinement of input instructions, or prompts, that direct the outputs of generative AI systems toward clinically meaningful responses. By carefully shaping prompts, developers and researchers ensure that models produce content that is neither arbitrary nor irrelevant, but rather generate structured outputs aligned with the platform's theoretical and clinical goals. In this sense, prompts function as an interface between computational capacities and hermeneutic interpretation, translating the complexity of psychotherapeutic discourse into usable insights (Shah et al., 2024; Wang & Hasan, 2024). As Sahoo et al. (2025) observe, prompt engineering uses task-specific cues to enhance model performance without modifying core model parameters. Similarly, Schulhoff et al. (2024) provide a comprehensive taxonomy of prompting techniques, identifying best practices and challenges that define the field's current and future directions. Variations in prompt structure can significantly affect performance and reliability, making prompt engineering essential for aligning AI-generated outputs with human intent and domain-specific objectives. Through this methodological approach, AI systems can produce high-quality, contextually relevant, and interpretable outputs across both research and clinical domains (Wang et al., 2025).

In PSICOGENIUS^{AI}, prompt engineering underpins the entire analytic pipeline. The first stage involves automatically transcribing therapeutic sessions via an API integration with *Assembly AI*. This service transforms raw audio into accurate text transcripts, handling variations in voice, intonation, and acoustic conditions typical of natural clinical dialogue. These transcripts form the foundational data upon which subsequent generative analyses are built. The text is then processed through APIs that connect to Gemini models, Google's constantly updated family of generative AI systems. Using carefully designed prompts, Gemini produces both single-session outputs, such as structured summaries, nuclear and satellite topics, and word clouds, as well as long-term analyses, including semantic shift trajectories, correlation matrices, and overall topic trends across sessions.

This layered architecture ensures the system can handle both the micro-level details of individual conversations and the macro-level patterns of therapeutic progress. A defining feature of PSICOGENIUS^{AI} is its iterative adaptability. The research and development team continuously monitors the accuracy, interpretability, and clinical relevance of the outputs, refining prompts and adjusting processing parameters as needed. This feedback loop, controlled and guided by psychotherapists and technicians, ensures that the system remains responsive to advances in speech-to-text technology, the evolving capabilities of generative models, and the emerging needs of psychotherapy research and practice.

Through this integrated workflow, which combines transcription, prompt engineering, and generative modelling, PSICOGENIUS^{AI} transforms raw therapeutic dialogue into structured, clinically interpretable knowledge. The pipeline bridges the probabilistic logic of AI and the hermeneutic requirements of psychotherapy, establishing a robust foundation that augments therapists' reflective and interpretative work without supplanting it. Although we are unable to disclose the technical specifications – as they are the property of the company that developed the software (Magellano s.c.) – usability tests were conducted over a two-month period prior to release by a team of psychotherapists and technical experts covering all aspects of the software.

2.2. Individual session outputs

The AI section of PSICOGENIUS^{AI} offers a suite of analytical tools tailored for psychotherapists. At the interface level, this section displays the complete list of patients. By selecting a specific patient, clinicians can access the entire history of their appointments. Within each record, the practitioner can upload an audio or video file of the psychotherapy session, which is then converted into multiple AI-generated outputs within a few minutes. These outputs are designed to streamline clinical insight and support reflective practice. The first step in this pipeline is the automatic transcription of the uploaded session, which is performed through an API integration with Assembly AI. Once the transcript is produced, all subsequent AI-driven outputs are generated by invoking Gemini models via dedicated APIs. Each output is the result of a carefully crafted prompting strategy that I have personally developed and refined, informed by research conducted during my doctoral studies. Examples of these prompts, together with their corresponding outputs, extracted from fictitious patients on PSICOGENIUS^{AI}, are presented in the figures that follow, and in the appendix 9.

From a clinical perspective, the immediacy of these outputs provides a significant advantage for therapists, particularly in the context of high caseloads. Within a few minutes, practitioners can

reacquaint themselves with the main points discussed in prior sessions, thereby entering the therapeutic encounter with a refreshed, accurate understanding of the patient's history. This not only facilitates continuity but also strengthens the therapeutic alliance, as the therapist demonstrates precise recall of previously shared material, conveying attentiveness and reliability that fosters patient trust. Moreover, minimizing the need for patients to repeat information already disclosed avoids redundancy and enhances the efficiency of each session.

It is important to note that clinicians often work with numerous patients in a single week, many of whom present with overlapping or complex life events. Human memory, governed by associative mechanisms, is susceptible to distortions such as the availability heuristic and other cognitive shortcuts described by Kahneman (2011), which can inadvertently lead to confusion or conflation between different cases. In this light, the structured outputs generated by PSICOGENIUS^{AI} serve as an external cognitive aid, supporting therapists in maintaining case-specific accuracy and preserving the integrity of the therapeutic process. Regarding the therapeutic alliance, research has shown that its quality significantly influences treatment adherence and outcomes (Horvath et al., 2011). One critical factor contributing to patient dropout is poor recall of therapeutic content and progress (Beck, 2016). When patients fail to remember prior sessions clearly, their engagement and trust in therapy tend to diminish, increasing the likelihood of premature termination (Simon & Ludman, 2010). By assisting therapists in accurate recall and documentation, AI tools like PSICOGENIUS^{AI} may help mitigate dropout rates and foster stronger, more durable therapeutic relationships.

Transcription

The first output generated for each uploaded session is the verbatim transcription of the therapeutic dialogue. This transcript provides the textual foundation for all subsequent analyses, enabling clinicians to revisit the content in detail while preserving the subtleties of patient-therapist interactions. The transcription is performed through the Assembly AI API, which guarantees high accuracy across diverse acoustic conditions. An example of this output is presented in Figure 37, together with the corresponding prompt in Appendix 9.1.

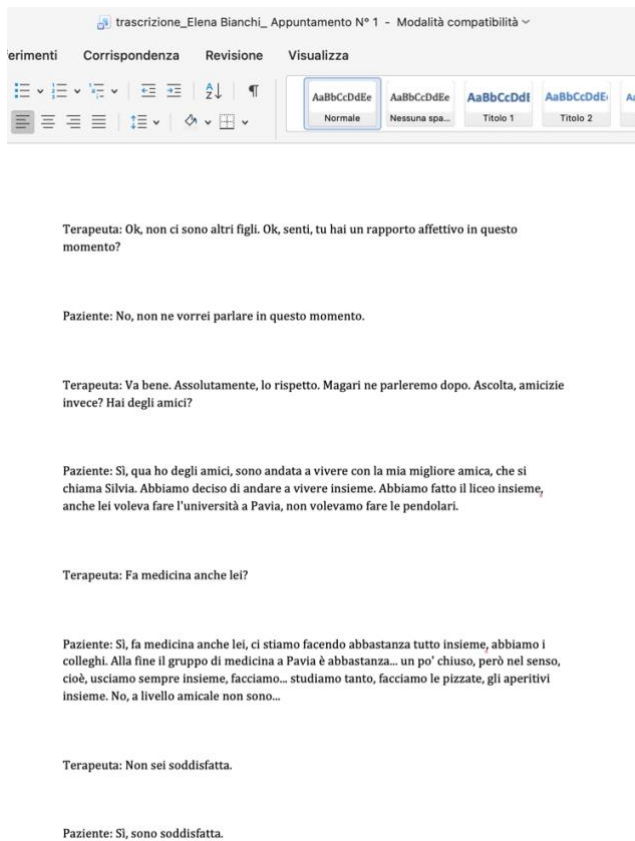


Figure 37. Example of the verbatim transcription output for a single therapeutic session, showing an excerpt of the therapist–patient dialogue in text format. All names and session details are simulated and non-identifiable, and do not correspond to real patients, to protect patient privacy.

Session Summary

The summary section presents a concise synthesis of the session, highlighting the most relevant informational points. This tool enables the therapist to gain an overarching view of the key themes and core data that emerged during the therapeutic exchange, facilitating reflective practice and continuity across sessions. An illustrative example is provided in Figure 38, along with the related prompt, as shown in Appendix 9.2.



Figure 38. Example of the PSICOGENIUS AI interface displaying the Session Summary (Sintesi) output for a single session (Elena Bianchi, Appointment No. 1, 31/03/2025). All personal data shown is simulated and non-identifiable and does not correspond to real patients, to protect patient privacy.

Word Cloud

This section provides a visual representation of the most salient words from the session. A word cloud is generated containing the twenty most significant terms, each associated with a score ranging from 0.0 to 1.0, where higher values indicate stronger semantic importance. The words are displayed in variable sizes and colour gradients, ranging from blue (high relevance) to yellow (low relevance), allowing the therapist to quickly identify dominant themes in the dialogue. A representative word cloud is shown in Figure 39, with the corresponding prompt illustrated in Appendix 9.3.



Figure 39. Example of the Word Cloud output generated for a single therapeutic session, displaying the most salient terms extracted from the dialogue with size and colour indicating their relative semantic importance. All content shown is simulated, non-identifiable, and does not correspond to real patients.

Nuclear Topics

The nuclear topics section presents a graphical visualization of the relative occurrence of each topic within the session. These topics correspond to macro-themes shared across all psychotherapies, and their distribution provides insight into the degree of emphasis given to each theme during the conversation. A supporting table, listed below the graph, outlines the topics and their corresponding proportions. This analysis allows therapists to perceive fluctuations in how these central themes are addressed over time in a patient's therapeutic trajectory. An example visualization is presented in Figure 40, accompanied by the corresponding prompt in Appendix 9.4.

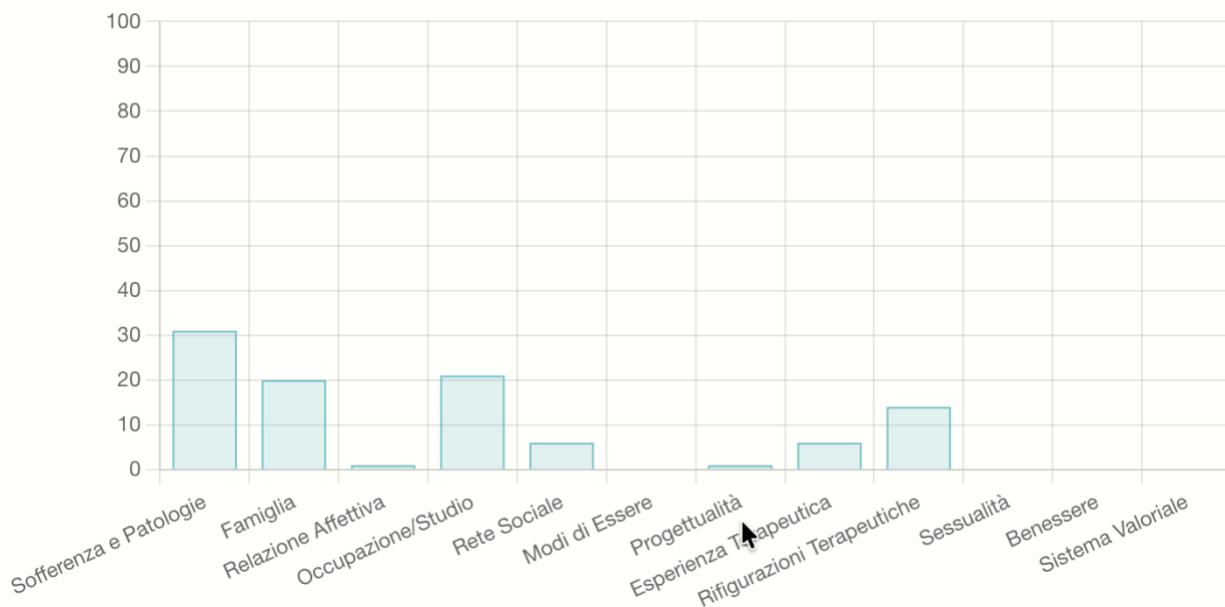


Figure 40. Example of the Nuclear Topics output for a single therapeutic session, showing the relative distribution of predefined macro-topic categories across the dialogue (with the corresponding proportions reported in an accompanying table). All content shown is simulated and non-identifiable and does not correspond to real patients to protect patient privacy.

AI-Derived Topics

The AI-derived topics section examines the main arguments that emerged during the session using advanced AI algorithms. For each identified topic, representative key expressions are provided that reflect the discussion's semantic content. This feature enables therapists to identify themes, reflect on their dynamics, and gain deeper insight into the therapeutic process. The selection of topics is driven by semantic relevance, ensuring that clinically significant issues are clearly surfaced. An example output is shown in Figure 41, along with the associated prompt in Appendix 9.5.



Figure 41. Example of the AI-Derived Topics output for a single therapeutic session, reporting the main topics identified from the session content together with representative key expressions. All content shown is simulated and non-identifiable and does not correspond to real patients to protect patient privacy.

2.3. Overall output analysis

The analytical power of PSICOGENIUS^{AI} extends beyond a single session, offering clinicians outputs that integrate data from multiple therapeutic encounters with the same patient. These tools become available once at least two sessions have been uploaded, enabling the platform to construct longitudinal insights that would otherwise remain inaccessible. By synthesizing information across time, the system allows therapists to monitor patient trajectories, identify patterns of change, and contextualize the therapeutic process within a broader clinical narrative. Importantly, all AI-generated outputs can be downloaded for integration into clinical records, and both the single-session and overall summaries, as well as the dynamic clinical sheet, remain modifiable by the therapist, ensuring professional oversight and contextual refinement.

From a clinical perspective, the availability of longitudinal indicators is particularly valuable because it supports reflective practice and informed decision-making at multiple levels of intervention. Specifically, therapists can:

- Adapt interventions in real time, tailoring therapeutic strategies to the patient's evolving needs based on objective evidence.
- Prepare for supervision and peer consultation, using integrated outputs to present complex cases to colleagues or multidisciplinary teams with clarity and precision.
- Monitor the distribution of topics over time to detect whether specific themes have been under- or overemphasized during therapy.
- Gain a quantitative overview of interactional dynamics, for example, assessing how much the therapist intervenes versus how much the patient integrates and responds to therapeutic inputs.
- Share therapy progress with patients, providing a transparent, data-informed account of the therapeutic journey that can foster collaboration and motivation.
- Reduce cognitive biases in recall, such as the peak-end bias described by Kahneman, whereby clinicians and patients may misremember the salience of a theme because it was discussed at the beginning or end of therapy, or because an emotionally striking event made it more memorable.

- Track salient life events and clinical history through the dynamic sheet, which organizes milestones and contextual factors into an easily navigable format.
- Formulate or refine diagnostic hypotheses, drawing from the longitudinal accumulation of symptoms and thematic patterns across sessions.
- Evaluate therapeutic efficacy retrospectively by analysing concluded therapies to identify which strategies proved functional and which were less effective.
- Consolidate clinical documentation, integrating structured AI outputs with personal clinical notes in a coherent digital environment.

Through these features, PSICOGENIUS^{AI} provides a multi-layered scaffolding for clinical reflection, enabling therapists not only to interpret the therapeutic process more systematically but also to support decision-making, collaborative supervision, and patient engagement.

Dynamic Clinical Sheet

The Dynamic Clinical Sheet provides an integrated overview of a patient's therapeutic history by extracting and organizing information from multiple uploaded sessions. It is structured into six subsections.

- 1) "General Information" presents the patient's demographic data (e.g., age, occupation) alongside relevant details about their immediate social network (e.g., parents, siblings, spouse, children).
- 2) "Symptoms and Distress" describes the primary symptoms and problems reported by the patient, allowing a quick assessment of presenting concerns.
- 3) "Clinical History" outlines a chronological account of clinically relevant events, including previous diagnoses and treatments, situating the current therapy within the patient's broader trajectory.
- 4) "Progress and Treatment Plan" evaluates the progress made over the course of therapy and delineates the current therapeutic strategy.
- 5) "Relevant Events and Dates" list significant life events and key dates that frame the patient's personal and clinical history.
- 6) "Diagnostic Areas of Interest" compiles areas of diagnostic concern, either explicitly formulated by the therapist or suggested by the system based on reported symptoms and contextual information.

This dynamic sheet allows therapists to maintain a comprehensive, continuously updated perspective on the patient's course of therapy, thereby facilitating more targeted and informed interventions. An

example of the clinical sheet is shown in Figure 42, and the corresponding prompt is shown in Appendix 9.6.

The screenshot shows a web application interface for a dynamic clinical sheet. At the top, there are three tabs: "Scheda Clinica Dinamica", "Sintesi Complessiva", and "Grafico Topic". The "Sintesi Complessiva" tab is currently selected. Below the tabs, the content is organized into three main sections, each with a heading and a list of bullet points.

Informazioni generali

- **Dettagli demografici:** Elena, 24 anni, studentessa di medicina a Pavia.
- **Ambiente familiare e sociale:** Ha una coinquilina e amica, Silvia.
- **Contesto familiare:** Famiglia complessa con la perdita del padre, un fratellastro, un fratello minore e un nuovo compagno della madre.
- **Fattori esterni:** Studente universitaria.

Sintomi/sofferenza

- Difficoltà di concentrazione nello studio.
- Pensieri intrusivi.
- Sensazione di tristezza.
- Incapacità di portare a termine gli esami universitari.
- Paura del giudizio.

Storia clinica

- Precedente breve percorso terapeutico nell'infanzia interrotto.
- Episodio di difficoltà alimentare durante il liceo.

Figure 42. Example of the PSICOGENIUS Dynamic Clinical Sheet overall output, illustrating the structured synthesis of patient information across sessions (e.g., general information, symptoms/distress, and clinical history). All content shown is simulated, non-identifiable, and does not correspond to real patients.

Overall Narrative Summary

The Overall Summary offers a discursive synthesis of the patient's therapeutic journey, constructed from the cumulative content of multiple sessions. Each session contributes to a chronologically ordered narrative, which highlights the most salient events, emerging themes, and evolving dynamics over time. The system continuously updates this text as new sessions are added, providing the therapist with an overall view of the patient's developmental trajectory. This feature enhances the clinician's ability to capture changes in relational dynamics, recurring problems, and significant life events, both within and outside the therapeutic setting. A representative example is shown in Figure 43, along with the corresponding prompt in Appendix 9.7.

Elena, una studentessa di medicina di **24 anni**, ha iniziato un percorso terapeutico alla fine di **marzo 2025** a causa di **difficoltà di concentrazione nello studio, pensieri intrusivi, un senso di tristezza generalizzato e incapacità di superare gli esami universitari**. Questo malessere, iniziato a **dicembre** e intensificatosi durante la sessione d'esami, ha compromesso la sua routine di studio e la preparazione per l'esame di **farmacologia**. Nonostante ciò, **Elena** afferma di non avere problemi di socialità e di avere un buon rapporto con la sua coinquilina e amica **Silvia**, anch'essa studentessa di medicina.

Il background familiare di **Elena** è complesso. Ha perso il padre in un incidente stradale quando aveva **12 anni** e ha scoperto l'esistenza di un fratellastro maggiore, **Andrea**, al funerale. Ha anche un fratello minore, **Fabio**, nato da una relazione successiva della madre con un uomo assente, e la madre ha un nuovo compagno. **Elena** ha accennato a un litigio con la madre riguardo alle circostanze dell'incidente del padre, un evento che sembra averla segnata profondamente. Sebbene tenda a minimizzarlo, la complessità della sua situazione familiare sembra essere una fonte di stress. **Elena** ha già affrontato un breve percorso terapeutico nell'infanzia, interrotto a causa della sua reticenza a parlare del lutto. Ha anche menzionato un episodio di difficoltà alimentare durante il liceo, che attribuisce alle pressioni familiari.

Figure 43. Example of the Overall Narrative Summary output, showing a narrative synthesis of the patient's therapeutic course derived from multiple sessions. All content shown is simulated and non-identifiable and does not correspond to real patients to protect patient privacy.

Topic Trajectory Graph

The Topic Trajectory Graph visualizes the occurrence and distribution of therapeutic topics across multiple sessions. On the y-axis, the percentage of occurrence of each topic is represented, while the x-axis indicates the chronological sequence of sessions. Each session is marked by coloured circles, with each colour corresponding to a specific topic. A legend beneath the graph maps colours to topic labels, and a table below further elaborates on each identified theme.

The size of each circle varies with the percentage of topic relevance, with larger circles indicating a stronger thematic presence. Additionally, the interface allows therapists to select specific topics to track their evolution over time, thereby enabling a more detailed exploration of the interplay between themes throughout the therapeutic process. This tool supports clinicians in identifying semantic shifts, recurrences, and emerging concerns, thus aligning computational analysis with reflective clinical practice. An illustrative graph is shown in Figure 44. The underlying prompt is limited to concatenating the individual Nuclear Topics outputs across appointments; therefore, the corresponding code is primarily concerned with front-end visualisation and interaction.

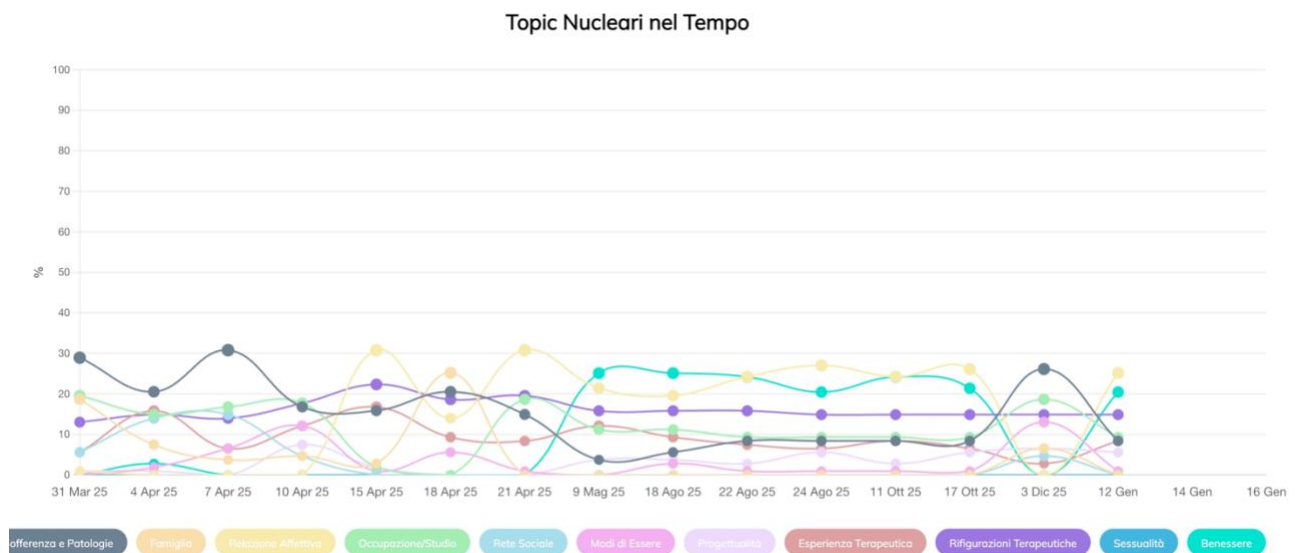


Figure 44. Example of the longitudinal Nuclear Topics over Time visualisation, displaying all selected topic labels concurrently across the therapeutic trajectory. Topics are colour-coded and reported in a legend, and their relative proportions are plotted at each appointment date in chronological order, enabling inspection of thematic changes across sessions. All content shown is simulated and non-identifiable.

In Figure 45, the same graph is shown, but only a subset of topics is selected. Restricting the display to topics of interest reduces visual complexity and facilitates a more fine-grained examination of reciprocal fluctuations and temporal coupling between themes across the therapeutic trajectory. For example, the association between SPS and AF is unclear until the session on 18 April, when their trajectories begin to converge and interact.

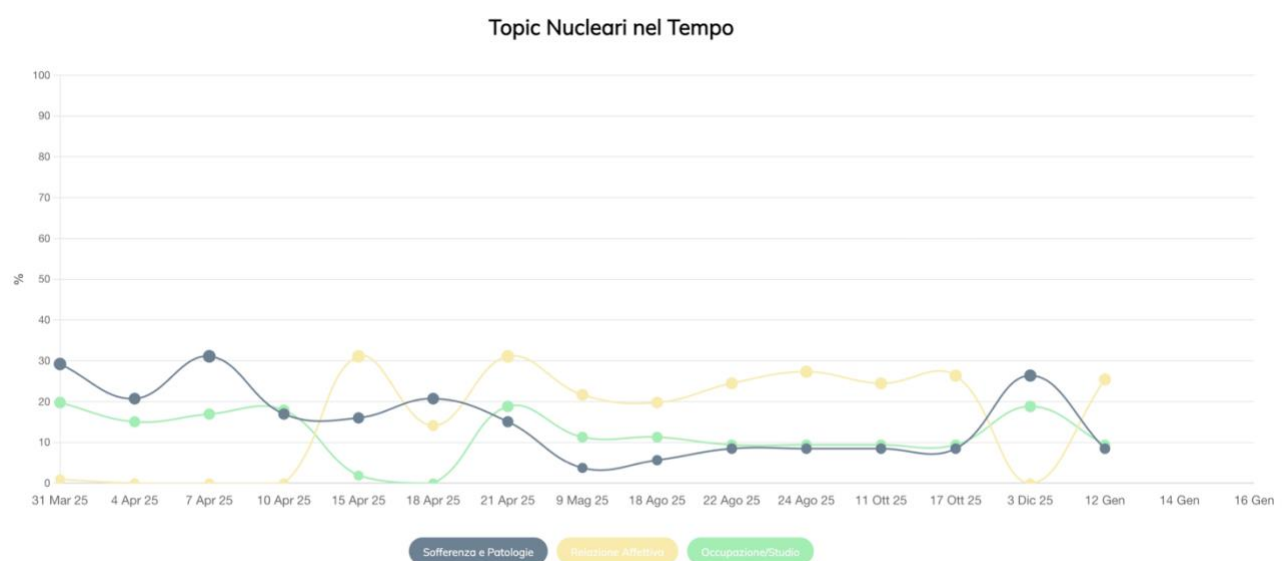


Figure 45. Example of the Nuclear Topics over Time visualisation with only three selected topic labels displayed, enabling a clearer inspection of their relative trajectories and interactions across consecutive session dates. All content shown is simulated and non-identifiable and does not correspond to real patients.

3. AI as an ally, not a replacement

Within PSICOGENIUS^{AI}, generative AI is conceptualized as a clinical resource that augments reflective practice rather than a system that replaces therapeutic judgment. This positioning is particularly important in psychotherapy, where meaning is co-constructed through language, embodiment, and intersubjective atonement. Any computational representation of the session, however sophisticated, captures only a subset of what is clinically salient and therefore risks promoting a reductionist view of the mind if treated as an exhaustive or objective account. For this reason, the outputs generated by PSICOGENIUS^{AI} (e.g., structured summaries, topic distributions, and longitudinal visualizations) are framed as hypothesis-generating artifacts that support the clinician’s interpretive work, not as authoritative conclusions.

The contemporary feasibility of such tools is closely linked to advances in Natural Language Processing, particularly the emergence of transformer-based architectures and large language models. Transformer models have enabled substantial improvements in contextual language representation, supporting a range of downstream tasks relevant to clinical documentation and text-based process analysis. In addition to foundational work on attention-based architectures (Vaswani et al., 2017), a key milestone in 2018 was the consolidation of transformer-based pretraining approaches that demonstrated how large-scale language modelling could be adapted to diverse interpretive and generative tasks (Devlin et al., 2019; Radford et al., 2018). These developments underpin the current generation of LLMs and make it technically plausible to generate coherent session-level summaries and structured analytic outputs from therapy transcripts – provided that their probabilistic nature is explicitly acknowledged and that they are governed by appropriate safeguards.

From a clinical perspective, the primary value of these systems lies in their use. When embedded within a human-in-the-loop workflow, AI can reduce the cognitive and administrative burden associated with transcription and synthesis, support continuity across sessions, and surface patterns that may be difficult to detect through manual review alone, particularly when outputs are paired with data visualisation that supports rapid inspection of trends and fluctuations over time. At the same time, maintaining the “ally-not replacement” stance is essential to mitigate risks such as automation bias (i.e., over-reliance on machine-generated suggestions), category errors (confusing descriptive regularities with clinical explanations), and the loss of phenomenological and relational complexity. In PSICOGENIUS^{AI}, this translates into an explicit division of labour: AI accelerates and structures information, while the clinician retains full responsibility for interpretation, formulation, and decision-making.

A key advantage of integrating AI into clinical workflows is the potential to enhance process-informed practice by making session content more accessible and comparable over time. Tools that support transcription, synthesis, and topic-based monitoring can help clinicians move from fragmented recall to a more systematic overview of what has been discussed, what has shifted, and what persists across appointments. This aligns with a perspective where the central question concerns the relationship between what occurs during psychotherapy and its clinical effects.

In PSICOGENIUS^{AI}, these benefits are operationalised through outputs that prioritise clinical usability over generic text generation: concise session summaries, structured clinical sheets, and longitudinal displays that briefly render topic dynamics interpretable. This is particularly relevant for continuity of care, where therapists may need to reconstruct a patient's trajectory across many sessions and identify clinically meaningful fluctuations (e.g., symptom-related themes that rise in parallel with relational stressors). Importantly, the emphasis on data visualisation is not merely aesthetic: reducing large volumes of narrative content into coherent visual patterns can facilitate hypothesis generation and support reflective clinical reasoning that might not emerge from slow, linear, long audio recordings.

A further opportunity concerns efficiency without standardisation. Digital mental health tools, when well-integrated, have been argued to support service delivery and engagement while preserving the need for support (Ben-Zeev, 2017). In this sense, AI can be leveraged to reduce the administrative and cognitive load associated with documentation – freeing time for clinical work – while still allowing the therapist to maintain authorship through review, correction, and selective integration of generated outputs.

These, however, must be balanced against well-defined limitations. First, there is the risk of reductionism: computational approaches may implicitly privilege what is easily representable (lexical units, topic frequencies) and thereby under-represent phenomenological, embodied, and intersubjective dimensions of experience. This is salient in hermeneutic approaches, where clinical meaning is understood to emerge through an interpretive cycle that cannot be reduced to purely formal textual features (Ricoeur, 1976). Second, generative systems are probabilistic and may produce plausible but contextually inappropriate statements; consequently, uncritical uptake can foster automation bias and a passive delegation of interpretive labour. Recent discussions in the literature converge on the idea that AI may support mental health care while remaining unable to substitute core human functions – therapeutic responsibility, contextual judgement, and the relational dimension of psychotherapy (Zhang & Wang, 2024; Bloch-Atefi, 2025).

The safe deployment of these tools requires attention to data governance and regulatory compliance, particularly where sensitive health data are involved. The slide set highlights practical issues across transcription workflows, data transmission, and data protection (e.g., GDPR, AI Act, encryption, and privacy policies) that must be addressed alongside model performance. Within this framework, prompting and system configuration become part of the safety layer: prompts are iteratively refined to improve output precision and to constrain (e.g., by inventing information), and the workflow is structured so that clinicians can edit outputs before any clinical use.

In PSICOGENIUS^{AI}, the “ally-not replacement” principle is implemented through a human-in-the-loop workflow: AI-derived outputs are treated as provisional hypotheses that must be reviewed, contextualised, and, when necessary, edited by the clinician before any clinical use. This is consistent with the platform’s broader rationale of supporting reflective practice and supervision via complementary computational indicators (e.g., topic trajectories, summaries, and longitudinal dashboards), while keeping interpretative authority and responsibility with the therapist (Haber et al., 2024). From a technical standpoint, the feasibility of these capabilities rests on transformer-based language modelling (Vaswani et al., 2017), but the platform’s emphasis is not on “automation of psychotherapy”; rather, it is on structured, inspectable outputs that can be integrated into clinicians’ reasoning without displacing it.

A second safeguard concerns prompt governance: prompting is treated as a methodological layer that constrains model behaviour towards clinically meaningful, structured outputs, and it is continuously refined based on expert feedback to improve robustness and interpretability. The development roadmap explicitly includes adapting the system to the user/professional’s own dataset to better fit therapeutic style and method; ongoing model selection based on quality benchmarks; creating clinically valid indices through collaboration with therapists and curated libraries, rather than relying on generic AI behaviours. This design directly targets two common failure modes in clinical AI use: over-trusting probabilistic outputs and drifting towards generic, decontextualised text generation.

PSICOGENIUS^{AI} could be an augmentative infrastructure: it can reduce administrative burden, provide more objective recall of session content, and support longitudinal reflection (including through visualisation), while the therapist remains accountable for meaning-making, formulation, and intervention planning.

4. Future directions for PSICOGENIUS^{AI}

Future development of PSICOGENIUS^{AI} involves two complementary strands: 1) continuous operational and legal maintenance of the clinical-management environment, and 2) scientific and product upgrades of the analytic layer that support reflective practice.

On the maintenance side, a core priority is to ensure that the practice-management component (invoicing and administrative workflows) remains aligned with evolving national requirements. This entails ongoing consultation with accountants to keep fiscal and billing procedures up to date, and periodic revision of privacy documentation and terms and conditions, in collaboration with legal counsel and the organisation's Data Protection Officer (DPO). Under the GDPR framework, a DPO is the designated professional who advises on and monitors compliance with data-protection obligations, supports privacy impact assessments, and serves as a point of contact for data-protection matters. This governance layer is essential for maintaining robust compliance with GDPR and for remaining aligned with the EU's emerging regulatory landscape for AI systems (including the AI Act and its emphasis on risk management, transparency, documentation, human oversight, robustness, and cybersecurity).

In parallel, the platform requires continuous improvement of the graphical user interface (GUI) and interaction design. The objective is to further reduce friction in everyday clinical use – supporting faster navigation across sessions, clearer longitudinal inspection (e.g., topic trajectories), and more efficient review/editing of outputs – so that the tool integrates smoothly into routine practice rather than adding cognitive load.

A major direction for upgrade is the development of new clinically meaningful indices to support analysis of the psychotherapy process. The intent is explicitly augmentative: to provide therapists with more objective, trackable descriptors of what occurs in therapy (e.g., changes in themes, conversational dynamics, or affective markers), thereby supporting supervision, reflective practice, and, when clinically appropriate, shared discussion with the patient. These indices are not designed to replace clinical judgement, diagnosis, or formulation; rather, they aim to counteract bias (e.g., selective recall, salience effects, confirmation bias) by providing structured evidence that can be inspected, questioned, and integrated into professional reasoning.

Examples of candidate indices for future implementation include turn-taking balance and interruption patterns, topic volatility and recurrence across sessions, semantic proximity between distress-related content and relational themes, lexical/affective indicators (e.g., shifts in emotional valence or

intensity across a session), and markers of therapeutic focus (e.g., increased mental-state language as a proxy for reflective processing). Importantly, PSICOGENIUS^{AI}'s mission in this area is not merely to “add metrics”, but to validate them: indices should be developed in parallel with research workflows – starting with human coding, inter-rater reliability work, and empirical validation against clinically meaningful anchors – before being deployed in the platform as decision-support artefacts. This research-to-product pathway is also consistent with the AI Act's emphasis on accountability and human oversight for higher-stakes.

A future direction – not planned for near-term implementation – is to integrate psychometric testing with automated test interpretation. At present, PSICOGENIUS^{AI} deliberately avoids this functionality due to substantive ethical, legal, and methodological concerns about using AI for diagnostic inference. Because generative systems are not infallible, erroneous interpretations could lead to inappropriate clinical conclusions; this immediately raises questions of accountability and liability (e.g., who is responsible if an AI-supported interpretation is wrong: the clinician, the developer, or the organisation). Robust test integration typically requires access to full test manuals, scoring rules, and interpretive guidelines – materials protected by copyright and licensing restrictions. For these reasons, the platform's development priorities remain focused on augmentative, transparency-oriented supports (e.g., documentation, longitudinal monitoring, validated process indices) rather than AI-mediated diagnosis or automated psychometric reporting.

Section 3 - Computational Psychotherapy and Suicidal Risk Analysis

During my six-month research stay (April–September 2025) at the University of Utah, I worked in Zac Imel’s lab, a leading group in computational psychotherapy and process–outcome research, which is central to my doctoral trajectory. In this period (and continuing through ongoing collaboration), I contributed to three connected workstreams.

(1) A replication-and-extension program on topic modelling in psychotherapy transcripts, benchmarking classical and transformer-era methods, and emphasizing interpretability and reproducibility.

(2) A hybrid framework combining topic-based representations and large language models with neural prediction to estimate personality-related dimensions at fine temporal resolution in psychotherapy dialogue.

(3) A human–machine collaboration line focused on predicting suicide risk severity from text-based crisis counselling, where we explore message-level risk trajectories and clinically usable decision-support principles.

Because this crisis-text dataset is highly sensitive, I completed the required Institutional Review Board (IRB) training (i.e., human-subjects protection and research privacy/confidentiality policies) prior to accessing it, and all work adheres to strict access-control and de-identification requirements consistent with approved research governance.

1. Generative Topic Modelling in Psychotherapy: Revisiting and Comparing with Established Methods

Abstract

Topic modelling has been used to uncover clinically meaningful themes in psychotherapy transcripts and to support the discrimination of broad therapeutic orientations. Building on the seminal 2015 work by Imel and colleagues, we propose a replication-and-extension study using the General Psychotherapy Corpus (Alexander Street Press; ~1,398 transcripts), focusing on methodological

rigor, interpretability, and reproducibility in the current era of transformer-based NLP and large language models (LLMs). Unlike the original study, we exclude Motivational Interviewing trials to ensure corpus homogeneity and introduce a systematic comparison of multiple generations of topic models, including classical probabilistic and matrix factorization approaches (LDA; NMF) and modern embedding-based methods (BERTopic; Combined Topic Models). A key innovation is the integration of LLM-assisted procedures to (a) resolve ambiguous or missing session-level therapy-approach labels via fixed, auditable prompts, and (b) generate standardized topic labels (concise titles, summaries, keyword sets) to be evaluated against human judgments. We evaluate models across complementary dimensions: coherence (C_v , NPMI), stability (bootstrap overlap on top words), predictive utility (therapy-approach classification from topic distributions), and coverage/redundancy of the topic space. Clinical interpretability is further assessed using word- and topic-intrusion tasks with expert raters and agreement metrics. We also derive clinically oriented analyses (topic co-occurrence networks, inter-session distances, outlier sessions for qualitative inspection) to connect topic structure with “everyday-life” psychotherapy processes at scale.

This work aims to provide a transparent benchmark for topic modelling in psychotherapy research, clarifying where LLMs add value (and where they may not) in labelling, interpretation, and reproducibility.

1.1. Database description

The present study employed the General Psychotherapy Corpus made available through Alexander Street Press (access via institutional subscription). This corpus contains approximately 1,398 transcripts of psychotherapy and medication management sessions, comprising nearly 2.0 million talk turns and 8.3 million tokens (including punctuation). Each transcript is accompanied by metadata that includes anonymized patient and therapist identifiers, session number, therapeutic approach (e.g., cognitive-behavioural, psychodynamic, humanistic/experiential, “other”), and information on the primary symptoms and subjects discussed during the session.

The metadata labels were assigned by editorial staff at Alexander Street Press and were not accompanied by inter-rater reliability statistics. Thus, while potentially informative, these labels should be treated as noisy descriptors rather than definitive classifications. In line with the procedure described by Imel, Steyvers, and Atkins (2015), all available transcripts were included in the analysis. Sessions were grouped into five broad treatment categories: (1) Psychodynamic (e.g., psychoanalysis, brief relational therapy, psychoanalytic psychotherapy), (2) Cognitive-Behavioral Therapy (including

Rational Emotive Behavior Therapy and related approaches), (3) Experiential/Humanistic (e.g., person-centered, existential therapy), (4) Drug Therapy/Medication Management, and (5) Other (e.g., Adlerian, Reality Therapy, Solution-Focused, family, and group therapy).

When transcripts contained either multiple or missing approach labels, categorization was resolved using collateral information from the metadata, such as therapist identity (when the therapist was strongly associated with a specific modality), reported use of characteristic interventions, or inspection of the raw transcript. If no collateral information was available, the first listed intervention was assigned as the treatment label; when this was not possible, the treatment label was left missing. No transcripts were excluded from the corpus under these procedures.

1.2. Significance and Methods

The present study was designed as a replication and methodological extension of the work conducted by Imel, Steyvers, and Atkins (2015). In their seminal paper, the authors demonstrated the potential of topic modelling techniques to uncover clinically relevant linguistic structures within psychotherapy sessions. Using a corpus of psychotherapy transcripts drawn primarily from the Alexander Street Press collection, supplemented with a set of Motivational Interviewing sessions from randomized controlled trials, their study provided proof-of-concept evidence that topic models can (a) identify meaningful semantic themes, (b) discriminate among different therapeutic approaches, (c) capture session-level similarities and differences through visualizations, and (d) classify sessions by treatment type with reasonable accuracy.

While groundbreaking, the Imel et al. study relied on topic modelling methods available at the time, specifically Latent Dirichlet Allocation (LDA) and related semi-supervised models, and it addressed the challenge of missing or multiple session-level labels by applying rule-based manual resolution (e.g., using collateral information from metadata, therapist identity, or assigning the first listed intervention). Topic coherence was primarily assessed qualitatively by inspecting exemplar terms and therapist statements, with limited quantitative evaluation.

The present study extends this work in several important ways. First, the dataset is restricted to the General Psychotherapy Corpus from Alexander Street Press, thereby excluding the additional Motivational Interviewing corpus used by Imel et al. (2015). This allows for a more homogeneous comparison across diverse psychotherapy modalities represented within the Alexander Street Press transcripts. Second, the issue of label assignment is addressed not through manual rules but through Large Language Models (LLMs), which categorize sessions into broad treatment classes. This

provides an opportunity to evaluate whether modern generative models can enhance the consistency and interpretability of label assignment relative to earlier, rule-based approaches

Third, beyond traditional LDA, the present analysis incorporates state-of-the-art topic modelling methods, including embedding-based techniques such as BERTopic and Combined Topic Models (CTM), as well as classical approaches (LDA, NMF). This design allows for a systematic comparison of multiple generations of topic modelling techniques applied to the same psychotherapy corpus. Fourth, the evaluation framework is expanded to include both quantitative measures (e.g., coherence metrics such as C_v and NPMI, topic stability across bootstraps, predictive accuracy for treatment classification) and human evaluation tasks (e.g., word and topic intrusion tasks, expert judgments of topic labels). Importantly, LLMs are used exclusively for topic interpretation and labelling, ensuring methodological reproducibility and testing their utility in enhancing clinical interpretability.

In summary, whereas Imel et al. (2015) established the feasibility of applying topic models to psychotherapy transcripts, the current study aims to (a) replicate their core findings using the same corpus, (b) assess whether modern topic modelling techniques improve coherence, stability, and predictive utility, and (c) directly compare rule-based versus LLM-assisted procedures for label assignment and topic interpretation. Through this replication and extension, the study advances computational psychotherapy research by evaluating how methodological innovations alter the linguistic representation of psychotherapy and clarifying the potential role of LLMs in supporting large-scale, clinically meaningful text analyses.

So, the three aims of this study are:

- 1) Replicate and extend topic modelling on the General Psychotherapy Corpus, and the question is: “Do modern topic models (BERTopic/CTM) yield higher coherence and stability than LDA/NMF on psychotherapy turns?”
- 2) Test the impact of LLM assistance on labelling and interpretability, and the questions are “Does LLM-assisted topic labelling (titles/summaries/keywords) improve human-rated interpretability compared to classical labelling workflows?” and “Can LLM-assisted therapy-approach assignment provide consistent, reproducible resolution of ambiguous/missing labels, replacing manual adjudication?”
- 3) Evaluate predictive utility and clinically oriented structure, and the questions are “Which modelling family best supports classification of macro-orientations (Psychodynamic, CBT, Humanistic/Experiential, Drug Therapy/Medication Management, Other)?” and “Do topic co-

occurrence networks and session-distance maps reveal clinically plausible communities and ‘off-cluster’ sessions warranting qualitative follow-up?”

All raw transcripts remain in the University of Utah Box repository; processing occurs via secure access (e.g., the Box API), and only derived artifacts are exported (trained models, aggregated statistics, figures). A de-identification pipeline combines spaCy NER with Microsoft Presidio, replacing sensitive entities with placeholders (e.g., [NAME], [LOC], [DATE]). Re-identification mappings are not stored unless strictly necessary and, if required, are encrypted and access-restricted.

Analyses are conducted at the level of speaker turns, systematically tagged as therapist [T] or patient [P]. All transcripts are first normalized by removing timestamps and editorial artifacts and by converting files to a uniform plain-text format (e.g., .txt). We then apply a clinically conservative NLP preprocessing pipeline including lemmatization, explicit preservation of negation, and a stopword policy designed to minimize the removal of clinically meaningful function words. Where appropriate, we additionally consider n-gram representations. Text is represented using complementary feature spaces depending on the modelling approach: bag-of-words (CountVectorizer), TF-IDF, and sentence-level transformer embeddings (Sentence-Transformers).

To ensure robust generalization and prevent information leakage, data are split into train/validation/test sets at the session level, stratified by psychotherapy approach. To further reduce leakage risk, all sessions belonging to the same case are constrained to remain within the same split.

Therapy approach labels fall into five macro-categories, as defined by Imel (2015). For sessions with ambiguous or missing labels, we implement a fixed LLM-assisted assignment protocol intended to replace ad hoc manual resolution. The protocol takes as input available session metadata (and, where permitted, rule-defined minimal excerpts) and returns a structured output consisting of a predicted label, an associated confidence estimate, and a brief rationale. Prompts and generation parameters are fixed (low-temperature) and stored in a versioned prompt log. Label reliability is assessed through targeted human spot-checking and sensitivity analyses focusing on low-confidence cases.

We compare topic modelling approaches spanning two families. Classical models include Latent Dirichlet Allocation (LDA), tuned via grid search over the number of topics (e.g., $k = 50/100/150/200$) and priors, and Non-negative Matrix Factorization (NMF) applied to TF-IDF matrices. Embedding-based models include BERTopic (transformer embeddings followed by clustering and c-TF-IDF-based topic representation) and Combined Topic Models (CTM), integrating contextual embeddings with probabilistic topic inference.

To standardize interpretability, we additionally employ LLM-based topic labelling. For each discovered topic, the LLM generates (i) a concise title (≤ 7 words), (ii) a 2–3 sentence summary, and (iii) a keyword set. Prompts and parameters are archived, and temperature is kept very low (e.g., 0–0.2) to reduce variability across runs.

Evaluation extends beyond qualitative inspection and approach discrimination. Topic quality is quantified using coherence metrics (C_v , NPMI), stability is assessed via bootstrap resampling and overlap indices (e.g., Jaccard, Rank-Biased Overlap on top- n terms), and coverage/redundancy indicators include topic assignment rates, entropy, and cluster-size distributions. Predictive utility is examined using session-level topic distributions as features to classify psychotherapy approaches with logistic regression and random forests, and we report cross-validated performance and systematic error analyses.

Clinical interpretability is further evaluated through word intrusion and topic intrusion tasks completed by 2–3 clinical experts, with inter-rater reliability estimated via Fleiss’ κ . In blinded comparisons, expert ratings are used to contrast LLM-generated versus human-generated topic labels and to assess perceived clinical usefulness. We conduct clinically oriented structural analyses, including topic co-occurrence networks (PMI-based) with Louvain community detection, and session distance analyses (e.g., Jensen–Shannon, Hellinger) visualized via heatmaps and low-dimensional embeddings (MDS/UMAP). Group differences across approaches are tested using nonparametric procedures (Kruskal–Wallis; Mann–Whitney with FDR correction), and “off-cluster” sessions are flagged for targeted qualitative inspection under minimal-excerpt rules. Reproducibility is ensured through pinned computational environments (e.g., `requirements.txt` / `environment.yml`), fixed random seeds, complete logging of hyperparameters and model versions, and versioned archiving of all LLM prompts and settings, alongside modular code organization and transparent pipeline documentation.

This study provides: a replication of the Imel (2015) topic-modelling demonstration on the Alexander Street Press psychotherapy corpus (with a homogeneity-focused design choice); a head-to-head benchmark of classical vs embedding-based topic models using modern criteria (coherence, stability, coverage, predictive utility); an explicit test of where LLMs help (label assignment, topic summarization/standardization) and where they may introduce risks (variance, bias), evaluated through human judgment and auditable prompts.

2. Extracting Personality Traits from Psychotherapy Transcripts with LDA and LLMs

This study aims to investigate whether personality-related dimensions can be inferred from psychotherapy language at a fine-grained temporal scale. I propose a hybrid framework that integrates Latent Dirichlet Allocation (LDA) and large language models (LLMs) to obtain interpretable semantic representations and trait-oriented annotations, and a neural network (ANN) trained to estimate Big Five personality scores at the level of individual talk turns. The model is trained on a large corpus of psychotherapy transcripts (Alexander Street Press) and evaluated on an independent naturalistic treatment consisting of 26 consecutive sessions. For the 26-session case, three expert raters provide turn-level personality annotations; inter-rater reliability is quantified, and model–human agreement is assessed to test out-of-sample generalization and clinical plausibility. The rationale is supported by prior evidence that personality can be partially recovered from language using topic-based models (Hu et al., 2017; Liu et al., 2016) and neural approaches to detecting psychological traits from text (El-Demerdash et al., 2021).

2.1. Database description

The present study uses the General Psychotherapy Corpus made available through Alexander Street Press (institutional subscription). The corpus includes approximately 1,398 transcripts of psychotherapy and medication management sessions, comprising nearly 2.0 million talk turns and 8.3 million tokens (including punctuation). Each transcript is accompanied by structured metadata, including anonymized patient and therapist identifiers, session number, and treatment descriptors such as therapeutic approach (e.g., cognitive-behavioural, psychodynamic, humanistic/experiential, “other”), as well as information on presenting symptoms and session topics. Metadata labels were assigned by editorial staff and are not accompanied by inter-rater reliability statistics; accordingly, these labels are treated as potentially informative but noisy descriptors rather than definitive classifications. All available transcripts are retained to preserve corpus coverage and variability. For analyses requiring approach stratification, sessions are grouped into broad treatment categories, and ambiguous or missing labels are resolved using collateral metadata and conservative decision rules without excluding transcripts. This design supports large-scale representation learning and topic-based semantic modeling, consistent with prior work linking topic structure to personality-related dimensions in large text corpora (Hu et al., 2017; Liu et al., 2016).

External validation is conducted on an independent naturalistic psychotherapy consisting of 26 consecutive sessions. The treatment is fully transcribed and segmented into speaker turns (therapist vs patient). Three expert raters provide turn-level Big Five annotations using a shared observer-rating rubric with anchored ordinal/continuous scores and explicit linguistic/interactional indicators. These ratings provide (i) an empirical estimate of inter-rater reliability and (ii) an out-of-sample benchmark for evaluating ANN predictions, aligning the validation logic with text-based psychological trait detection paradigms (El-Demerdash et al., 2021).

2.2. Significance and Methods

Automatically extracting personality-relevant information from psychotherapy transcripts can strengthen process–outcome research by enabling longitudinal measurement of individual-difference signals at a resolution that is rarely feasible with manual coding alone. While expert annotation remains the gold standard, it is costly and difficult to scale to large corpora; a validated computational framework can therefore provide a complementary measurement layer that supports hypothesis testing on within-treatment dynamics, therapist-patient interaction patterns, and stability versus fluctuation of trait-related signals across sessions. Prior work suggests that personality can be partially recovered from language, both through topic-modelling approaches that explicitly link thematic structure to trait dimensions (Hu et al., 2017; Liu et al., 2016) and through neural approaches that predict psychological traits from text representations (El-Demerdash et al., 2021). However, clinical research also requires interpretability and auditability: topic-based representations can provide human-readable semantic anchors for trait estimates, while taxonomy-derived semantic features offer an additional route to interpretable text representations and facilitate transparent error analysis (Škrlj et al., 2021; Toleu et al., 2019). In this study, Big Five traits are operationalized through a turn-level observer-rating rubric, where each dimension is scored on an anchored ordinal/continuous scale with explicit behavioural-linguistic indicators; the rubric is shared across expert raters and used to define the target construct for model evaluation, allowing the assessment to focus on model–human agreement under a clearly specified measurement framework.

The study is organized into two phases: model development and training on the Alexander Street Press psychotherapy corpus, followed by external validation on an independent naturalistic psychotherapy consisting of 26 consecutive sessions with expert human coding. All transcripts are standardized through conservative preprocessing (removal of editorial artifacts and normalization of formatting) and segmented into speaker turns, each labelled by speaker role (therapist vs. patient). To reduce leakage, data are partitioned at the session level (and, where applicable, by case), ensuring that highly similar conversational material does not appear across training and evaluation subsets. In

the development phase, LDA is fitted to the large corpus to derive latent topics and turn/session-level topic-mixture representations. These topic features serve as both an interpretable semantic layer and structured inputs for downstream prediction, consistent with topic-based approaches that connect language structure and personality-related dimensions (Hu et al., 2017; Liu et al., 2016). In parallel, a documented LLM protocol (fixed prompts, constrained outputs, low-variance settings, and a structured response schema) is used to support trait-oriented annotation workflows and targeted quality checking where necessary, while the independent 26-session dataset is kept separate for validation rather than supervision.

An ANN is then trained to estimate Big Five personality scores for each turn. Input representations may combine text embeddings with topic-mixture vectors; additionally, interpretable semantic features derived from taxonomic resources can be incorporated to support transparency and diagnostic inspection of the drivers of predictions (Škrlj et al., 2021). This design is aligned with prior work on language-based trait prediction using deep learning representations (El-Demerdash et al., 2021). For external validation, three expert raters annotate the 26-session psychotherapy using a shared turn-level observer-rating rubric: each Big Five dimension is scored with explicit anchors (e.g., 0–4 or 1–5) and defined linguistic/interactional cues, and the rubric is designed to minimize overlap between stable trait inference and transient state content. Inter-rater reliability is quantified to characterize the stability of the human criterion, and model performance is evaluated via model–human agreement by comparing ANN predictions to aggregated human scores (e.g., mean or reliability-weighted composites), using agreement-oriented indices appropriate for continuous trait estimates (e.g., ICC/correlation-based metrics and absolute-error measures), complemented by calibration checks across sessions and speakers.

3. Human-Machine Collaboration in Predicting Suicide Risk Severity from Text

In the lab of Zac Imel at the University of Utah, an ongoing line of research aim to develop a set of complementary research lines on human–machine collaboration to estimate suicide risk severity from text-based crisis counselling conversations. Building on transformer-based risk modelling in large-scale crisis-text settings (Broadbent et al., 2023) and motivated by evidence that suicidal ideation can fluctuate substantially over short time windows (Kleiman et al., 2017), these studies treat risk estimation as a dynamic, message-level process rather than a single outcome. Current directions include modelling within-chat risk trajectories, linking model-predicted risk changes to counsellor

responses and interactional markers, and evaluating interpretability and calibration as prerequisites for decision support. More broadly, this work aligns with recommendations to shift from overconfident prediction toward therapeutic assessment and risk management (Hawton et al., 2022) and emphasizes robustness and transportability when adapting large language models to specialized clinical text (Liu et al., 2019).

3.1. Database description

The project leverages a large, de-identified corpus of text-based crisis counselling conversations collected by a crisis line in the Mountain West of the United States (available 24/7 and primarily marketed to teens and young adults in Utah, while remaining accessible beyond the state and across age groups). The dataset spans 2017–2021 and includes 2,110,114 messages from 81,569 crisis chats, associated with 38,595 unique client identification numbers and 92 counsellors. Because the platform is anonymous, no demographic information is available for clients or counsellors; however, counsellors are described as licensed or license-eligible clinical social workers who have received additional training in crisis counselling. Chats are brief but highly variable in length (median 18 messages; mean 25.9, SD 25.7, range 1–593; 95% between 2–94 messages).

To support message-level modelling at scale, the corpus is accompanied by an automated suicide-risk indicator that quantifies *expressed* suicide risk in each message as the conversation unfolds. Specifically, a transformer-based model (RoBERTa) – pre-trained on large open-source text and then fine-tuned on 5,992 de-identified crisis chats from the same service – was trained to predict therapist-rated post-chat risk (low vs. high), achieving strong discrimination (reported ROC AUC = 0.904). The resulting output can be used as a probability of high expressed risk at the message level, enabling analyses of within-chat dynamics without requiring intensive manual coding for the full corpus.

In complementary work using the same broad setting, a time-bounded sample (June 2020–April 2021) was curated for detailed linguistic process analyses (e.g., readability), retaining chats that met minimum dialogic and text-length requirements (e.g., ≥ 150 words) and excluding non-dialogue artifacts and repeat texters. This illustrates how the larger crisis-text ecosystem can be systematically sub-sampled to address specific process questions while preserving the core advantages of naturalistic, high-volume crisis data.

3.2. Significance

Text-based crisis counselling offers an unusually direct window into suicide-risk communication as a rapidly changing, within-person process. Evidence from ecological momentary assessment indicates that suicidal ideation can fluctuate substantially over short time windows, which motivates analytic approaches that model *within-conversation dynamics* (e.g., escalation patterns, turning points, and stabilization phases) rather than treating risk as a single static endpoint (Kleiman et al., 2017). At the same time, contemporary clinical scholarship emphasizes that traditional “risk prediction” has inherent limits in real-world practice and that the field benefits from shifting toward therapeutic assessment, formulation, and risk management – a stance that aligns naturally with human–machine collaboration rather than automation (Hawton et al., 2022).

Within this framing, the dataset supports a program of work on human–machine collaboration in which NLP models are evaluated by their *clinical utility* and their ability to support decision-making under uncertainty. A first line of investigation concerns message-level risk trajectories, where model outputs can be used to study how expressed risk evolves across the conversational arc and how these trajectories relate to counsellor behaviours and interactional features. Prior work in text-based crisis settings suggests that transformer-based models can detect clinically meaningful risk signals from encounter content, and that error analyses may reveal clinically informative “near-miss” cases (e.g., false positives reflecting elevated – but not formally labelled-risk), which is especially relevant when designing decision-support tools (Broadbent et al., 2023).

A second line focuses on decision support and workflow integration, including the development of calibrated and interpretable signals (e.g., probabilistic summaries, uncertainty flags, and text-span attributions) that are explicitly scoped to augment counsellor judgment while reducing cognitive burden and limiting automation bias. This is consistent with the broader recommendation that high-stakes mental health applications should prioritize safe use, interpretability, and risk management over headline performance metrics alone (Hawton et al., 2022).

A third direction addresses robustness and transportability across time periods and conversational subtypes. Methodologically, this includes comparing domain-adapted transformer encoders with hybrid approaches that combine neural representations with clinically interpretable structures (e.g., topic dynamics or process markers). The RoBERTa training framework is a relevant methodological anchor for this work because it formalized design choices that materially affect downstream performance and generalization, which is central when adapting large models to specialized clinical text (Liu et al., 2019).

Overall, the dataset is valuable not only for improving prediction, but for advancing a process-oriented science of crisis intervention: identifying conversational patterns linked to escalation versus de-escalation, formalizing how clinicians and models can complement one another under uncertainty, and defining evaluation criteria that privilege safety, transparency, and actionable insight. Our ongoing collaboration with the U.S. group is intentionally preliminary at this stage, but it opens a coherent set of empirically testable questions about *when* algorithmic signals help, *when* they hinder, and which safeguards are required for responsible human–machine partnership in crisis settings.

References

- Adjerid, I., & Kelley, K. (2018). Big data in psychology: A framework for research advancement. *The American psychologist*, 73(7), 899–917. <https://doi.org/10.1037/amp0000190>
- Ægisdóttir, S., White, M. J., Spengler, P. M., Goldman, S. E., Ben-Porath, Y. S., & Handel, R. W. (2006). The meta-analysis of clinical judgment project: Fifty-six years of accumulated research on clinical versus statistical prediction. *The Counseling Psychologist*, 34(3), 341–382. <https://doi.org/10.1177/0011000005285875>
- Alonso-Sánchez, M. F., Ford, S. D., MacKinley, M., Silva, A., Limongi, R., & Palaniyappan, L. (2022). Progressive changes in descriptive discourse in First Episode Schizophrenia: a longitudinal computational semantics study. *Schizophrenia* (Heidelberg, Germany), 8(1), 36. <https://doi.org/10.1038/s41537-022-00246-8>
- Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mane, D. (2016). Concrete problems in AI safety. *arXiv preprint*. <https://arxiv.org/abs/1606.06565>
- Anderson, T., Bein, E., Pinnell, B., & Strupp, H. (1999). Linguistic analysis of affective speech in psychotherapy: A case grammar approach. *Psychotherapy research*, 9(1), 88–99.
- Angus, L., Levitt, H., & Hardtke, K. (1999). The Narrative Processes Coding System: Research Applications and Implications for Psychotherapy Practice. *Journal of clinical psychology*, 55(10), 1255–1270. [https://doi.org/10.1002/\(SICI\)1097-4679\(199910\)55:10<1255::AID-JCLP7>3.0.CO;2-F](https://doi.org/10.1002/(SICI)1097-4679(199910)55:10<1255::AID-JCLP7>3.0.CO;2-F)
- Angus, L. E., & McLeod, J. (Eds.). (2004). *The handbook of narrative and psychotherapy: Practice, theory and research*. Sage.
- Angus, L. E., & Kagan, F. (2013). Assessing client self-narrative change in emotion-focused therapy of depression: an intensive single case analysis. *Psychotherapy* (Chicago, Ill.), 50(4), 525–534. <https://doi.org/10.1037/a0033358>
- Arciero, G., & Bondolfi, G. (2012). *Sé, identità e stili di personalità*. Bollati Boringhieri.
- Arciero, G., Bondolfi, G., & Mazzola, V. (2019). *Fondamenti di psicoterapia fenomenologica: cura di sé e psicologia non razionalista*. Bollati Boringhieri.
- Atkins, D. C., Rubin, T. N., Steyvers, M., Doeden, M. A., Baucom, B. R., & Christensen, A. (2012). Topic models: a novel method for modelling couple and family text data. *Journal of Family Psychology: JFP, Journal of the Division of Family Psychology of the American Psychological Association* (Division 43), 26(5), 816–827. <https://doi.org/10.1037/a0029607>
- Atkins, D. C., Steyvers, M., Imel, Z. E., & Smyth, P. (2014). Scaling up the evaluation of psychotherapy: evaluating motivational interviewing fidelity via statistical text classification. *Implementation science: IS*, 9, 49. <https://doi.org/10.1186/1748-5908-9-49>
- Atzil-Slonim, D., Juravski, D., Bar-Kalifa, E., Gilboa-Schechtman, E., Tuval-Mashiach, R., Shapira, N., & Goldberg, Y. (2021). Using topic models to identify clients' functioning levels and alliance ruptures in psychotherapy. *Psychotherapy* (Chicago, Ill.), 58(2), 324–339. <https://doi.org/10.1037/pst0000362>

- Baes, N., Haslam, N., & Vylomova, E. (2023). Semantic shifts in mental health-related concepts. In N. Tahmasebi, S. Montariol, H. Dubossarsky, A. Kutuzov, S. Hengchen, D. Alfter, F. Periti, & P. Cassotti (Eds.), *Proceedings of the 4th Workshop on Computational Approaches to Historical Language Change* (pp. 119–128). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2023.lchange-1.13>
- Barocas, S., Hardt, M., & Narayanan, A. (2019). *Fairness and machine learning: Limitations and opportunities*. FairMLBook.org. <https://fairmlbook.org/>
- Beck, A. T. (1979). *Cognitive therapy and the emotional disorders*. Penguin.
- Beck, A. T., Steer, R. A., & Brown, G. (1996). Beck depression inventory–II. *Psychological assessment*.
- Beck, J. S. (2011). *Cognitive behaviour therapy: Basics and beyond* (2nd ed.). The Guilford Press.
- Beck, A. T. (2016). Cognitive therapy: Nature and relation to behavior therapy. *Behavior Therapy*, 47(6), 776–784.
- Bedi, G., Carrillo, F., Cecchi, G., Slezak, D. F., Sigman, M., Mota, N. B., Ribeiro, S., Javitt, D. C., & Corcoran, C. M. (2015). Automated analysis of free speech predicts psychosis onset in high-risk youths. *NPJ Schizophrenia*, 1, Article 15030. <https://doi.org/10.1038/npjrschz.2015.30>
- Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? FAccT '21: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623. <https://doi.org/10.1145/3442188.3445922>
- Ben-Zeev, D. (2017). Technology in mental health: Creating new knowledge and inventing the future of services. *Psychiatric Services*, 68(2), 107–108. <https://doi.org/10.1176/appi.ps.201600520>
- Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3, 993–1022. <https://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf>
- Blei, D. M., & McAuliffe, J. D. (2007). Supervised topic models. In *Advances in Neural Information Processing Systems* (Vol. 20).
- Blei, D. M., & Lafferty, J. D. (2009). Topic Models. In A. Srivastava, & M. Sahami (Eds.), *Text Mining: Classification, Clustering and Applications* (pp. 71-93). Cambridge: Chapman and Hall/CRC.
- Bloch-Atefi, A. (2025). Balancing ethics and opportunities: The role of AI in psychotherapy and counselling. *Psychotherapy and Counselling Journal of Australia*, 13(1). <https://doi.org/10.59158/001c.129884>
- Blodgett, S. L., Barocas, S., Daumé III, H., & Wallach, H. (2020). Language (technology) is power: A critical survey of “bias” in NLP. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5456–5476. <https://doi.org/10.18653/v1/2020.acl-main.485>

- Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5, 135–146. https://doi.org/10.1162/tacl_a_00051
- Bolasco, S., Canzonetti, A., Capo, F. M., Della Ratta-Rinaldi, F., & Singh, B. K. (2005). Understanding text mining: A pragmatic approach. In S. Sirmakessis (Ed.), *Knowledge mining* (pp. 31–50). Springer.
- Bolukbasi, T., Chang, K. W., Zou, J. Y., Saligrama, V., & Kalai, A. T. (2016). Is a man to a computer programmer as a woman is to a homemaker? Debiasing word embeddings. *Advances in Neural Information Processing Systems*, 29, 4349–4357. <https://doi.org/10.48550/arXiv.1607.06520>
- Bamler, R., & Mandt, S. (2017). Dynamic Word Embeddings. In the *proceedings of the International Conference on Machine Learning*. <https://doi.org/10.48550/arXiv.1702.08359>
- Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., ... & Liang, P. (2021). On the opportunities and risks of foundation models. *arXiv preprint*. <https://arxiv.org/abs/2108.07258>
- Bonnefon, J.-F., Shariff, A., & Rahwan, I. (2016). The social dilemma of autonomous vehicles. *Science*, 352(6293), 1573–1576. <https://doi.org/10.1126/science.aaf2654>
- Bose, I., & Mahapatra, R. K. (2001). Business data mining – A machine learning perspective. *Information & Management*, 39(3), 211–225. [https://doi.org/10.1016/S0378-7206\(01\)00091-X](https://doi.org/10.1016/S0378-7206(01)00091-X)
- Brachman, R., & Levesque, H. (2004). *Knowledge representation and reasoning*. Elsevier.
- Borckardt, J. J., Nash, M. R., Murphy, M. D., Moore, M., Shaw, D., & O'Neil, P. (2008). Clinical practice as natural laboratory for psychotherapy research: a guide to case-based time-series analysis. *The American psychologist*, 63(2), 77–95. <https://doi.org/10.1037/0003-066X.63.2.77>
- Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2), 77–101. <https://doi.org/10.1191/1478088706qp063oa>
- Broadbent, M., Medina Grespan, M., Axford, K., Zhang, X., Srikumar, V., Kiou, B., & Imel, Z. (2023). A machine learning approach to identifying suicide risk among text-based crisis counseling encounters. *Frontiers in Psychiatry*, 14, 1110527. <https://doi.org/10.3389/fpsy.2023.1110527>
- Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33, 1877-1901. <https://arxiv.org/abs/2005.14165>
- Bruner, J. (1990). *Acts of meaning*. Harvard University Press.
- Brynjolfsson, E. and McAfee, A. (2017) *The Business of Artificial Intelligence*. Harvard Business Review, 7, 3-11. <https://starlab-alliance.com/wp-content/uploads/2017/09/The-Business-of-Artificial-Intelligence.pdf>
- Buchholz, M. B., & Kächele, H. (2017). From turn-by-turn to larger chunks of talk: An exploratory study in psychotherapeutic micro-processes using conversation analysis. *Research in psychotherapy* (Milano), 20(3), 257. <https://doi.org/10.4081/ripppo.2017.257>

- Buenaño-Fernandez, D., González, M., Gil, D., & Luján-Mora, S. (2020). Text mining of open-ended questions in self-assessment of university teachers: An LDA topic modeling approach. *IEEE Access*, 8, 35318–35330. <https://doi.org/10.1109/ACCESS.2020.2974983>
- Burkhardt, H. A., Pullmann, M. D., Hull, T. D., Areán, P. A., & Cohen, T. (2022). Comparing emotion feature extraction approaches for predicting depression and anxiety. In A. Zirikly, D. Atzil-Slonim, M. Liakata, S. Bedrick, B. Desmet, M. Ireland, A. Lee, S. MacAvaney, M. Purver, R. Resnik, & A. Yates (Eds.), *Proceedings of the Eighth Workshop on Computational Linguistics and Clinical Psychology* (pp. 105–115). *Association for Computational Linguistics*. <https://doi.org/10.18653/v1/2022.clpsych-1.9>
- Caliskan, A., & Lewis, M. (2020). Social biases in word embeddings and their relation to human cognition [Preprint]. *PsyArXiv*. <https://doi.org/10.31234/osf.io/d84kg>
- Calvo, R. A., Milne, D. N., Hussain, M. S., & Christensen, H. (2017). Natural Language Processing in Mental Health Applications Using Non-Clinical Texts. *Natural Language Engineering*, 23(5), 1–37. <https://doi.org/10.1017/S1351324916000383>
- Castonguay, L. G., Eubanks, C. F., Goldfried, M. R., Muran, J. C., & Lutz, W. (2015). Research on psychotherapy integration: building on the past, looking to the future. *Psychotherapy research: journal of the Society for Psychotherapy Research*, 25(3), 365–382. <https://doi.org/10.1080/10503307.2015.1014010>
- Cautin, R. L., & Lilienfeld, S. O. (Eds.) (2015). *Encyclopedia of clinical psychology*. Chichester, West Sussex; Malden, MA: John Wiley and Sons.
- Chalkidis, I., Fergadiotis, M., Malakasiotis, P., Aletras, N., & Androutsopoulos, I. (2020). LEGAL-BERT: The Muppets straight out of law school. *arXiv preprint*. <https://arxiv.org/abs/2010.02559>
- Chambless, D. L., & Hollon, S. D. (1998). Defining empirically supported therapies. *Journal of Consulting and Clinical Psychology*, 66(1), 7–18. <https://doi.org/10.1037/0022-006X.66.1.7>
- Chang, Y.-C., Liew, D.-J., & Ku, C.-H. (2025). Personality classification from social media data using NLP. *Journal of Psycholinguistic Research*. Advance online publication. <https://doi.org/10.1080/08874417.2025.2512809>
- Chen, M., Tworek, J., Jun, H., Yuan, Q., Pinto, H. P. d. O., Kaplan, J., ... & Zaremba, W. (2021). Evaluating large language models trained on code. *arXiv preprint*. <https://arxiv.org/abs/2107.03374>
- Chen, B., Zhang, Z., Langrené, N., & Zhu, S. (2023). Unleashing the potential of prompt engineering for large language models. *arXiv preprint*. <https://doi.org/10.48550/arXiv.2310.14735>
- Chiappelli, M., Lo Coco, G., Gullo, S., Bensì, L., & Prestano, C. (2011). The Outcome Questionnaire 45.2. Italian validation of an instrument for the assessment of psychological treatments. *Epidemiology and Psychiatric Sciences*, 17(2). <https://doi.org/10.1017/S1121189X00002852>
- Chomsky, N. (1956). Three models for the description of language. *IRE Transactions on Information Theory*, 2(3), 113–124. <https://doi.org/10.1109/TIT.1956.1056813>

- Christiano, P. F., Leike, J., Brown, T. B., Martic, M., Legg, S., & Amodei, D. (2017). Deep reinforcement learning from human preferences. *arXiv preprint*. <https://doi.org/10.48550/arXiv.1706.03741>
- Cioffi, V., Mosca, L. L., Moretto, E., Ragozzino, O., Stanzione, R., Bottone, M., Maldonato, N. M., Muzii, B., & Sperandeo, R. (2022). Computational Methods in Psychotherapy: A Scoping Review. *International journal of environmental research and public health*, 19(19), 12358. <https://doi.org/10.3390/ijerph191912358>
- Cohen, A. M., & Hersh, W. R. (2005). A survey of current work in biomedical text mining. *Briefings in bioinformatics*, 6(1), 57–71. <https://doi.org/10.1093/bib/6.1.57>
- Collins, M. (2003). Head-driven statistical models for natural language parsing. *Computational Linguistics*, 29(4), 589–637. <https://doi.org/10.1162/089120103322753356>
- Cooper, M., & McLeod, J. (2010). *Pluralistic counselling and psychotherapy*. SAGE Publications Ltd.
- Coppersmith, G., Harman, C., & Dredze, M. (2014). Measuring Post-Traumatic Stress Disorder on Twitter. *Proceedings of the International AAAI Conference on Web and Social Media*, 8(1), 579-582. <https://doi.org/10.1609/icwsm.v8i1.14574>
- Corcoran, C. M., Carrillo, F., Fernández-Slezak, D., Bedi, G., Klim, C., Javitt, D. C., Bearden, C. E., & Cecchi, G. A. (2018). Prediction of psychosis across protocols and risk cohorts using automated language analysis. *World psychiatry: official journal of the World Psychiatric Association (WPA)*, 17(1), 67–75. <https://doi.org/10.1002/wps.20491>
- Corcoran, C. M., & Cecchi, G. A. (2020). Using Language Processing and Speech Analysis for the Identification of Psychosis and Other Disorders. *Biological psychiatry. Cognitive neuroscience and neuroimaging*, 5(8), 770–779. <https://doi.org/10.1016/j.bpsc.2020.06.004>
- Crevier, D. (1993). *AI: The tumultuous history of the search for artificial intelligence*. Basic Books.
- Creswell, J. W., & Clark, V. L. P. (2017). *Designing and Conducting Mixed Methods Research*. Thousand Oaks, CA: Sage Publications.
- Crits-Christoph, P., Gibbons, M. B. C., & Mukherjee, D. (2013). Psychotherapy process-outcome research. In M. J. Lambert (Ed.), *Bergin and Garfield's Handbook of Psychotherapy and Behavior Change* (6th ed., pp. 298–340). Hoboken, NJ: Wiley.
- Croskerry, P. (2009). A universal model of diagnostic reasoning. *Academic Medicine*, 84(8), 1022–1028. <https://doi.org/10.1097/ACM.0b013e3181ace703>
- Darby, W. C., & Weinstock, R. (2018). The Limits of Confidentiality: Informed Consent and Psychotherapy. *Focus (American Psychiatric Publishing)*, 16(4), 395–401. <https://doi.org/10.1176/appi.focus.20180020>
- Dartmouth College. (2022). *The Dartmouth Conference*. Retrieved from <https://www.dartmouth.edu>
- Davidson, S., & Vallée, M.-A. (Eds.). (2016). *Hermeneutics and phenomenology in Paul Ricoeur: Between text and phenomenon* (Vol. 2). Springer.

- Dawes, R. M., Faust, D., & Meehl, P. E. (1989). Clinical versus actuarial judgment. *Science*, 243(4899), 1668–1674. <https://doi.org/10.1126/science.2648573>
- Deerwester, S. C., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6), 391–407. [https://doi.org/10.1002/\(SICI\)1097-4571\(199009\)41:6<391::AID-ASII>3.0.CO;2-9](https://doi.org/10.1002/(SICI)1097-4571(199009)41:6<391::AID-ASII>3.0.CO;2-9)
- De Haan, S. (2017). *Psichiatria enattiva. La mente, il corpo e il mondo*. Raffaello Cortina Editore.
- Del Giacco, L., Salcuni, S., & Anguera, M. T. (2019). The Communicative Modes Analysis System in Psychotherapy from Mixed Methods Framework: Introducing a New Observation System for Classifying Verbal and Non-verbal Communication. *Frontiers in psychology*, 10, 782. <https://doi.org/10.3389/fpsyg.2019.00782>
- Demaria, F., Pontillo, M., Bellantoni, D., Di Vincenzo, C., & Vicari, S. (2023). Phenomenological considerations of the world of the obsessive patient. *Journal of Clinical Medicine*, 12(13), 4193. <https://doi.org/10.3390/jcm12134193>
- Demner-Fushman, D., Chapman, W. W., & McDonald, C. J. (2009). What can natural language processing do for clinical decision support? *Journal of Biomedical Informatics*, 42(5), 760–772. <https://doi.org/10.1016/j.jbi.2009.08.007>
- Demszky, D., Movshovitz-Attias, D., Ko, J., Cowen, A., Nemade, G., & Ravi, S. (2020). GoEmotions: A dataset of fine-grained emotions. In D. Jurafsky, J. Chai, N. Schluter, & J. Tetreault (Eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 4040–4054). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2020.acl-main.372>
- Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of NAACL-HLT* (pp. 4171–4186). <https://doi.org/10.18653/v1/N19-1423>
- DiMaggio, P., Nag, M., & Blei, D. M. (2013). Exploiting affinities between topic modeling and the sociological perspective on culture: Application to newspaper coverage of U.S. government arts funding. *Poetics*, 41(6), 570–606. <https://doi.org/10.1016/j.poetic.2013.08.004>
- Doshi-Velez, F., & Kim, B. (2017). Towards A Rigorous Science of Interpretable Machine Learning. arXiv preprint. <https://arxiv.org/abs/1702.08608>
- Du, K., Zhao, Y., Mao, R., & Xing, F. (2024). Natural language processing in finance: A survey. *Information Fusion*, 115, 102755. <https://doi.org/10.1016/j.inffus.2024.102755>
- Ekman, P. (1992). An argument for basic emotions. *Cognition and Emotion*, 6(3–4), 169–200. <https://doi.org/10.1080/02699939208411068>
- El-Demerdash, K., El-Khoribi, R. A., Shoman, M. A. I., & Abdou, S. (2021). Psychological human traits detection based on universal language modeling. *Egyptian Informatics Journal*, 22, 239–244. <https://doi.org/10.1016/j.eij.2020.09.001>

- Elgammal, A., Liu, B., Elhoseiny, M., & Mazzone, M. (2017). CAN: Creative adversarial networks, generating “art” by learning about styles and deviating from style norms. *arXiv preprint*. <https://arxiv.org/abs/1706.07068>
- Elhaddad, M., & Hamam, S. (2024). AI-Driven Clinical Decision Support Systems: An Ongoing Pursuit of Potential. *Cureus*, 16(4), e57728. <https://doi.org/10.7759/cureus.57728>
- Elliott, R., & Shapiro, D. A. (1992). Client and therapist as analysts of significant events. In S. G. Toukmanian & D. L. Rennie (Eds.), *Psychotherapy process research: Paradigmatic and narrative approaches* (pp. 163–186). Sage Publications, Inc.
- Elliott, R., Fischer, C. T., & Rennie, D. L. (1999). Evolving guidelines for publication of qualitative research studies in psychology and related fields. *British Journal of Clinical Psychology*, 38(3), 215–229. <https://doi.org/10.1348/014466599162782>
- Elliott R. (2002). Hermeneutic single-case efficacy design. *Psychotherapy research: journal of the Society for Psychotherapy Research*, 12(1), 1–21. <https://doi.org/10.1080/713869614>
- Elo, S., & Kyngäs, H. (2008). The qualitative content analysis process. *Journal of Advanced Nursing*, 62(1), 107–115. <https://doi.org/10.1111/j.1365-2648.2007.04569.x>
- Emmelkamp P. M. (2005). Technological innovations in clinical assessment and psychotherapy. *Psychotherapy and psychosomatics*, 74(6), 336–343. <https://doi.org/10.1159/000087780>
- Epstein, R. S. (1994). *Keeping boundaries: Maintaining safety and integrity in the psychotherapeutic process*. Washington, DC: American Psychiatric Press.
- Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639), 115–118. <https://doi.org/10.1038/nature21056>
- Eubanks, V. (2018). *Automating inequality: How high-tech tools profile, police, and punish the poor*—St. Martin's Press.
- European Parliament and Council of the European Union. (2024). Regulation (EU) 2024/1689 of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). *Official Journal of the European Union*.
- Eysenck, H. J. (1952). The effects of psychotherapy: an evaluation. *Journal of Consulting Psychology*, 16(5), 319–324. <https://doi.org/10.1037/h0063633>
- Farhud, D. D., & Zokaei, S. (2021). Ethical Issues of Artificial Intelligence in Medicine and Healthcare. *Iranian Journal of Public Health*, 50(11), i–v. <https://doi.org/10.18502/ijph.v50i11.7600>
- Feigenbaum, E. A., & Buchanan, B. G. (1993). DENDRAL and Meta-DENDRAL: Roots of knowledge systems and expert system applications. *Artificial Intelligence*, 59(1-2), 233–240. <https://stacks.stanford.edu/file/druid:pj337tr4694/pj337tr4694.pdf>
- Feyerabend, P. K. (1981). *Against Method: Outline of an Anarchistic Theory of Knowledge* (3rd ed.). Verso. (Original work published 1975)

- Feldman, R., & Sanger, J. (2007). *The text mining handbook: Advanced approaches in analyzing unstructured data*. Cambridge University Press. <https://doi.org/10.1017/CBO9780511546914>
- Fillmore, C. J. (1976). Frame semantics and the nature of language. *Annals of the New York Academy of Sciences*, 280, 1, 20–32. <https://doi.org/10.1111/j.1749-6632.1976.tb25467.x>
- Flückiger, C., Del Re, A. C., Wampold, B. E., & Horvath, A. O. (2018). The alliance in adult psychotherapy: A meta-analytic synthesis. *Psychotherapy*, 55(4), 316–340. <https://doi.org/10.1037/pst0000172>
- Floridi, L., & Cows, J. (2019). *A unified framework of five principles for AI in society*. Harvard Data Science Review, 1(1). <https://doi.org/10.1162/99608f92.8cd550d1>
- Floridi, L., & Chiriatti, M. (2020). GPT-3: Its nature, scope, limits, and consequences. *Minds and Machines*, 30, 681–694. <https://doi.org/10.1007/s11023-020-09548-1>
- Foa, E. B., Hembree, E. A., & Rothbaum, B. O. (2007). *Prolonged exposure therapy for PTSD: Emotional processing of traumatic experiences: Therapist guide*. Oxford University Press. <https://doi.org/10.1093/med:psych/9780195308501.001.0001>
- Frank, J. D., & Frank, J. B. (1991). *Persuasion and healing: A comparative study of psychotherapy* (3rd ed.). Baltimore, MD: Johns Hopkins University Press.
- Freud, S. (1963). *Introductory lectures on psychoanalysis* (J. Strachey, Trans.). Norton. (Original work published 1917)
- Fuchs, T. (2021). *Ecologia del cervello: fenomenologia e biologia della mente incarnata*. Astrolabio.
- Gadamer, H. G. (1975). *Truth and Method*. Sheed and Ward. London, UK.
- Gadamer, H.-G. (2004). *Truth and method* (J. Weinsheimer & D. G. Marshall, Trans.; 2nd rev. ed.). Continuum. (Original work published 1960)
- Gallagher, S. (2004). Hermeneutics and the Cognitive Sciences. *Journal of Consciousness Studies*, 11(10-11), 162–174.
- Gallagher, S. (2019). Phenomenology and the cognitive sciences. *The oxford handbook of phenomenological psychopathology*, 262-272.
- Garb, H. N. (1998). *Studying the clinician: Judgment research and psychological assessment*. American Psychological Association. <https://doi.org/10.1037/10299-000>
- Garb H. N. (2005). Clinical judgment and decision making. *Annual review of clinical psychology*, 1, 67–89. <https://doi.org/10.1146/annurev.clinpsy.1.102803.143810>
- Garety, P. A., & Freeman, D. (2013). The past and future of delusions research: from the inexplicable to the treatable. *The British journal of psychiatry: the journal of mental science*, 203(5), 327–333. <https://doi.org/10.1192/bjp.bp.113.126953>

- Gaut, G., Steyvers, M., Imel, Z. E., Atkins, D. C., & Smyth, P. (2017). Content Coding of Psychotherapy Transcripts Using Labeled Topic Models. *IEEE Journal of Biomedical and Health Informatics*, 21(2), 476–487. <https://doi.org/10.1109/JBHI.2015.2503985>
- Gee, J. P. (2014). *An introduction to discourse analysis: Theory and method* (4th ed.). Routledge. <https://doi.org/10.4324/9781315819679>
- Geeraerts, D. (2009). *Theories of lexical semantics*. OUP Oxford.
- Gelo, O., Braakmann, D., & Benetka, G. (2008). Quantitative and qualitative research: beyond the debate. *Integrative psychological & behavioral science*, 42(3), 266–290. <https://doi.org/10.1007/s12124-008-9078-3>
- Gelo, O. C. G., Salcuni, S., & Colli, A. (2012). Text Analysis within Quantitative and Qualitative Psychotherapy Process Research: Introduction to Special Issue. *Research in Psychotherapy: Psychopathology, Process and Outcome*, 15(2), 45–53. <https://doi.org/10.4081/ripppo.2012.144>
- Gelo, O. C., & Salvatore, S. (2016). A dynamic systems approach to psychotherapy: A meta-theoretical framework for explaining psychotherapy change processes. *Journal of Counseling Psychology*, 63(4), 379–395. <https://doi.org/10.1037/cou0000150>
- Gelo, O. C. G., Lagetto, G., Dinoi, C., Belfiore, E., Lombi, E., Blasi, S., Aria, M., & Ciavolino, E. (2020). Which Methodological Practice(s) for Psychotherapy Science? A Systematic Review and a Proposal. *Integrative psychological & behavioral science*, 54(1), 215–248. <https://doi.org/10.1007/s12124-019-09494-3>
- Giorgi, A. (1985). *Phenomenology and Psychological Research*. Pittsburgh, PA: Duquesne University Press.
- Giorgi, A. (1994). A phenomenological perspective on certain qualitative research methods. *Journal of Phenomenological Psychology*, 25(2), 190–220. <https://doi.org/10.1163/156916294X00034>
- Giorgi, A. P., & Giorgi, B. M. (2003). The descriptive phenomenological psychological method. In P. M. Camic, J. E. Rhodes, & L. Yardley (Eds.), *Qualitative research in psychology: Expanding perspectives in methodology and design* (pp. 243–273). American Psychological Association. <https://doi.org/10.1037/10595-013>
- Giorgi, A. (2009). *The descriptive phenomenological method in psychology: A modified Husserlian approach*. Duquesne University Press.
- Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... & Bengio, Y. (2014). Generative adversarial nets. *Advances in Neural Information Processing Systems*, 27, 2672–2680. <https://arxiv.org/abs/1406.2661>
- Goertzel, B., & Pennachin, C. (2007). *Artificial general intelligence*. Springer.
- Gottschalk, L. A., & Lolas, F. (1989). The Gottschalk-Gleser content analysis method of measuring the magnitude of psychological dimensions: Its application in transcultural research. *Psychopathology*, 22(suppl. 1), 83–104.

- Greenberg, L. S., & Safran, J. D. (1989). Emotion in psychotherapy. *The American psychologist*, 44(1), 19–29. <https://doi.org/10.1037//0003-066x.44.1.19>
- Greenberg L. S. (2012). Emotions, the great captains of our lives: their role in the process of change in psychotherapy. *The American psychologist*, 67(8), 697–707. <https://doi.org/10.1037/a0029858>
- Grencavage, L. M., & Norcross, J. C. (1990). Where are the commonalities among the therapeutic common factors? *Professional Psychology: Research and Practice*, 21(5), 372–378. <https://doi.org/10.1037/0735-7028.21.5.372>
- Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences of the United States of America*, 101 Suppl 1(Suppl 1), 5228–5235. <https://doi.org/10.1073/pnas.0307752101>
- Griffiths, T. L., Steyvers, M., & Tenenbaum, J. B. (2007). Topics in semantic representation. *Psychological review*, 114(2), 211–244. <https://doi.org/10.1037/0033-295X.114.2.211>
- Grimmer, J., & Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis*, 21(3), 267–297. <https://doi.org/10.1093/pan/mps028>
- Groopman, J. (2007). *How doctors think*. Houghton Mifflin Harcourt.
- Grootendorst, M. (2022). BERTopic: Neural topic modelling with a class-based TF-IDF procedure. *arXiv preprint*. <https://doi.org/10.48550/arXiv.2203.05794>
- Grove, W. M., & Meehl, P. E. (1996). Comparative efficiency of informal (subjective, impressionistic) and formal (mechanical, algorithmic) prediction procedures: The clinical–statistical controversy. *Psychology, Public Policy, and Law*, 2(2), 293–323. <https://doi.org/10.1037/1076-8971.2.2.293>
- Guba, E. G., & Lincoln, Y. S. (1994). Competing paradigms in qualitative research. In N. K. Denzin & Y. S. Lincoln (Eds.), *Handbook of qualitative research* (pp. 105–117). Sage Publications, Inc.
- Gulshan, V., Peng, L., Coram, M., Stumpe, M. C., Wu, D., Narayanaswamy, A., Venugopalan, S., ... & Webster, D. R. (2016). Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA*, 316(22), 2402–2410. <https://doi.org/10.1001/jama.2016.17216>
- Guntuku, S. C., Buffone, A., Jaidka, K., Eichstaedt, J. C., & Ungar, L. H. (2019). Understanding and Measuring psychological Stress through social media. *Proceedings of the Thirteenth International AAAI Conference on Web and Social Media (ICWSM)*, 13(1). <https://doi.org/10.1609/icwsml.v13i01.3223>
- Guo, Q., Wang, R., Guo, J., Li, B., Song, K., Tan, X., Liu, G., Bian, J., & Yang, Y. (2023). EvoPrompt: Connecting large language models with evolutionary algorithms yields powerful prompt optimizers. *arXiv preprint*. <https://doi.org/10.48550/arXiv.2309.08532>
- Haber, Y., Levkovich, I., Hadar-Shoval, D., & Elyoseph, Z. (2024). The artificial third: A broad view of the effects of introducing generative artificial intelligence on psychotherapy. *JMIR Mental Health*, 11, e54781. <https://doi.org/10.2196/54781>

- Hamilton, W. L., Leskovec, J., & Jurafsky, D. (2016). Diachronic word embeddings reveal statistical properties of semantic change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics* (pp. 1489–1501). <https://doi.org/10.18653/v1/P16-1141>
- Hayes, A. M., Feldman, G. C., Beevers, C. G., Laurenceau, J. P., Cardaciotto, L., & Lewis-Smith, J. (2007). Discontinuities and cognitive changes in an exposure-based cognitive therapy for depression. *Journal of consulting and clinical psychology*, 75(3), 409–421. <https://doi.org/10.1037/0022-006X.75.3.409>
- Hayes, A. M., & Andrews, L. A. (2020). A complex systems approach to the study of change in psychotherapy. *BMC medicine*, 18(1), 197. <https://doi.org/10.1186/s12916-020-01662-2>
- Haugeland, J. (1985). *Artificial intelligence: The very idea*. MIT Press.
- Hawton, K., Lascelles, K., Pitman, A., Gilbert, S., & Silverman, M. (2022). Assessment of suicide risk in mental health practice: Shifting from prediction to therapeutic assessment, formulation, and risk management. *The Lancet Psychiatry*, 9(11), 922–928. [https://doi.org/10.1016/S2215-0366\(22\)00232-2](https://doi.org/10.1016/S2215-0366(22)00232-2)
- He, J., Rungta, M., Koleczek, D., Sekhon, A., Wang, F. X., & Hasan, S. (2024). Does prompt formatting have any impact on LLM performance? (arXiv:2411.10541). *arXiv*.
- Heidegger, M. (1962). *Being and time* (J. Macquarrie & E. Robinson, Trans.). Harper & Row. (Original work published 1927)
- Hein, S. F., & Austin, W. J. (2001). Empirical and hermeneutic approaches to phenomenological research in psychology: a comparison. *Psychological methods*, 6(1), 3–17. <https://doi.org/10.1037/1082-989x.6.1.3>
- Hill, C. E. (2011). *Consensual qualitative research: A practical resource for investigating social science phenomena* (2nd ed.). Washington, DC: American Psychological Association.
- Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8), 1735–1780. <https://doi.org/10.1162/neco.1997.9.8.1735>
- Horvath, A. O., Del Re, A. C., Flückiger, C., & Symonds, D. (2011). *Alliance in individual psychotherapy*. *Psychotherapy*, 48(1), 9–16. <https://doi.org/10.1037/a0022186>
- Houston, N. M. (2023). Distant reading. In A. Hammond (Ed.), *Technology and literature* (pp. 361–376). Cambridge University Press. <https://doi.org/10.1017/9781108560740.023>
- Howard, J., & Ruder, S. (2018). Universal language model Fine-Tuning for Text Classification. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, 1, 328–339. <https://doi.org/10.18653/v1/P18-1031>
- Hu, Z., Liu, Y., Zhang, C., & Xu, Y. (2017, June). The analysis of the topic's personality traits using a new topic model. In *Proceedings of the 2017 2nd International Conference on Image, Vision and Computing (ICIVC)* (pp. 694–704). IEEE. <https://doi.org/10.1109/ICIVC.2017.7984720>

- Hu, K. (2023, February 2). ChatGPT sets record for fastest-growing user base—analyst note. *Reuters*. <https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/>
- Hutchins, J. (2006). Machine translation: History. In K. Brown (Ed.), *Encyclopedia of Language and Linguistics* (2nd ed., pp. 375–383). Elsevier. <https://doi.org/10.1016/B0-08-044854-2/00937-8>
- Imel, Z. E., Steyvers, M., & Atkins, D. C. (2015). Computational psychotherapy research: scaling up the evaluation of patient-provider interactions. *Psychotherapy* (Chicago, Ill.), 52(1), 19–30. <https://doi.org/10.1037/a0036841>
- Imel, Z., Tai-Seale, M., Smyth, P., & Atkins, D. (2021). Identifying Topics in Patient and Doctor Conversations Using Natural Language Processing Methods. *Patient-Centered Outcomes Research Institute (PCORI)*. <https://doi.org/10.25302/08.2021.ME.160234167>
- Janis, I. L. (1982). *Groupthink: Psychological Studies of Policy Decisions and Fiascos*. Boston: Houghton Mifflin.
- Joachims, T. (1998). Text Categorization with Support Vector Machines: Learning with many relevant features. *European Conference on Machine Learning (ECML)*, 137–142. <https://doi.org/10.1007/BFb0026683>
- Jones, K. S. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28(1), 11–21. <https://doi.org/10.1108/eb026526>
- Jones, E. E. (1985). Psychotherapy Process Q-Set manual. *Unpublished manuscript*. University of California, Berkeley.
- Jones, E. E., Cumming, J. D., & Horowitz, M. J. (1988). Another look at the nonspecific hypothesis of therapeutic effectiveness. *Journal of Consulting and Clinical Psychology*, 56(1), 48–55. <https://doi.org/10.1037/0022-006X.56.1.48>
- Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science*, 349(6245), 255–260. <https://doi.org/10.1126/science.aaa8415>
- Joshi, P., Santy, S., Budhiraja, A., Bali, K., & Choudhury, M. (2020). The state and fate of linguistic diversity and inclusion in the NLP world. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 6282–6293. <https://doi.org/10.18653/v1/2020.acl-main.560>
- Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., ... & Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873), 583–589. <https://doi.org/10.1038/s41586-021-03819-2>
- Jurafsky, D., & Martin, J. H. (2023). *Speech and Language Processing* (3rd ed.). Pearson.
- Kahneman, D., & Tversky, A. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, 185(4157), 1124–1131. <https://doi.org/10.1126/science.185.4157.1124>
- Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, 47(2), 263–291. <https://doi.org/10.2307/1914185>

- Kahneman, D., & Tversky, A. (1984). Choices, values, and frames. *American Psychologist*, 39(4), 341–350. <https://doi.org/10.1037/0003-066X.39.4.341>
- Kahneman, D. (2011). *Thinking, Fast and Slow*. Penguin Books, UK.
- Kazdin, A. E. (2008). Evidence-based treatment and practice: New opportunities to bridge clinical research and practice, enhance the knowledge base, and improve patient care. *American Psychologist*, 63(3), 146–159. <https://doi.org/10.1037/0003-066X.63.3.146>
- Kaplan, A., & Haenlein, M. (2019). Siri, Siri, in my hand: Who's the fairest in the land? On the interpretations, illustrations, and implications of artificial intelligence. *Business Horizons*, 62(1), 15–25. <https://doi.org/10.1016/j.bushor.2018.08.004>
- Katrakazas, C., Quddus, M., Chen, W.-H., & Deka, L. (2015). Real-time motion planning methods for autonomous on-road driving: State-of-the-art and future research directions. *Transportation Research Part C: Emerging Technologies*, 60, 416–442. <https://doi.org/10.1016/j.trc.2015.09.011>
- Kerz, E., Zanwar, S., Qiao, Y., & Wiechmann, D. (2023). Toward explainable AI (XAI) for mental health detection based on language behavior. *Frontiers in Psychiatry*, 14, 1219479. <https://doi.org/10.3389/fpsy.2023.1219479>
- Kherwa, P., & Bansal, P. (2019). Topic modelling: A comprehensive review. *EAI Endorsed Transactions on Scalable Information Systems*, 7(24), e2. DOI:10.4108/eai.13-7-2018.159623
- Kim, Y. (2014). Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, Doha, Qatar. Association for Computational Linguistics. <https://doi.org/10.3115/v1/D14-1181>
- Kingma, D. P., & Welling, M. (2014). Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*. <https://arxiv.org/abs/1312.6114>
- Kleiman, E. M., Turner, B. J., Fedor, S., Beale, E. E., Huffman, J. C., & Nock, M. K. (2017). Examination of real-time fluctuations in suicidal ideation and its risk factors: Results from two ecological momentary assessment studies. *Journal of Abnormal Psychology*, 126(6), 726–738. <https://doi.org/10.1037/abn0000273>
- Klein, G. A. (2017). *Sources of power: How people make decisions*. MIT press.
- Kojima, T., Gu, S., Reid, M., Matsuo, Y., & Iwasawa, Y. (2022). Large language models are zero-shot reasoners. *arXiv preprint*. <https://doi.org/10.48550/arXiv.2205.11916>
- Koskenniemi, K. (1983). Two-level morphology: A general computational model for word-form recognition and production. *Publication Series of the Department of General Linguistics*, 11, University of Helsinki.
- Krenn, M., Buffoni, L., Coutinho, B., Eppel, S., Foster, J. G., Gritsevski, A., Lee, H., Lu, Y., Moutinho, J. P., Sanjabi, N., Sonthalia, R., Tran, N. M., Valente, F., Xie, Y., Yu, R., & Kopp, M. (2022). Predicting the future of AI with AI: High-quality link prediction in an exponentially growing knowledge network. *arXiv preprint*. <https://doi.org/10.48550/arXiv.2210.00881>
- Krippendorff, K. (2018). *Content analysis: An introduction to its methodology*. Sage publications.

- Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6), 84–90. <https://doi.org/10.1145/3065386>
- Kübler, S., McDonald, R., & Nivre, J. (2009). Dependency parsing. *Synthesis lectures on human language technologies*, 1 (1), 1-127.
- Lalk, C., Steinbrenner, T., Kania, W., Popko, A., Wester, R., Schaffrath, J., Eberhardt, S., Schwartz, B., Lutz, W., & Rubel, J. (2024). Measuring Alliance and Symptom Severity in Psychotherapy Transcripts Using Bert Topic Modeling. *Administration and policy in mental health*, 51(4), 509–524. <https://doi.org/10.1007/s10488-024-01356-4>
- Lambert, M. J., & Finch, A. E. (1999). The Outcome Questionnaire. In M. E. Maruish (Ed.), *The use of psychological testing for treatment planning and outcomes assessment* (2nd ed., pp. 831–869). Lawrence Erlbaum Associates Publishers.
- Landauer, T.K. and Dumais, S.T. (1997) A Solution to Plato’s Problem: The Latent Semantic Analysis Theory of Acquisition, Induction, and Representation of Knowledge. *Psychological Review*, 104, 211-240. <http://dx.doi.org/10.1037/0033-295X.104.2.211>
- Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33(1), 159–174.
- Laska, K. M., Gurman, A. S., & Wampold, B. E. (2014). Expanding the lens of evidence-based practice in psychotherapy: A common factors perspective. *Psychotherapy*, 51(4), 467–481. <https://doi.org/10.1037/a0034332>
- LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, 521(7553), 436–444. <https://doi.org/10.1038/nature14539>
- Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2020). BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics (Oxford, England)*, 36(4), 1234–1240. <https://doi.org/10.1093/bioinformatics/btz682>
- Lenat, D. B., & Feigenbaum, E. A. (1991). On the thresholds of knowledge. *Artificial Intelligence*, 47(1-3), 185–250. <https://www.ijcai.org/Proceedings/87-2/Papers/122.pdf>
- Levitt, H. M., Pomerville, A., & Surace, F. I. (2016). A qualitative meta-analysis examining clients’ experiences of psychotherapy: A new agenda. *Psychological Bulletin*, 142(8), 801–830. <https://doi.org/10.1037/bul0000057>
- Levitt, H. M., Pomerville, A., Surace, F. I., & Grabowski, L. M. (2017). Metamethod study of qualitative psychotherapy research on clients' experiences: Review and recommendations. *Journal of Counseling Psychology*, 64(6), 626–644. <https://doi.org/10.1037/cou0000222>
- Li, Y., Song, J., & Ermon, S. (2017). InfoGAIL: Interpretable imitation learning from visual demonstrations. *Advances in Neural Information Processing Systems (NeurIPS) 2017. arXiv preprint*. <https://arxiv.org/abs/1703.08840>

- Libbrecht, M. W., & Noble, W. S. (2015). Machine learning applications in genetics and genomics. *Nature Reviews Genetics*, 16(6), 321–332. <https://doi.org/10.1038/nrg3920>
- Liccione, D. (2019). *Psicoterapia cognitive neurospicologica*. Bollati Boringhieri.
- Liccione, D., & Siciliano, L. (2025). The usefulness of combining topic modelling and statistical analysis to investigate the therapeutic process: A single case study. *Psychotherapy Research: Journal of the Society for Psychotherapy Research*, 1–21. Advance online publication. <https://doi.org/10.1080/10503307.2025.2500504>
- Liu, Y., Wang, J., & Jiang, Y. (2016). PT-LDA: A latent variable model to predict personality traits of social network users. *Neurocomputing*, 210, 155–163.
- Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint*. <https://doi.org/10.48550/arXiv.1907.11692>
- Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., & Neubig, G. (2021). Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *arXiv preprint*. <https://doi.org/10.48550/arXiv.2107.13586>
- Liu, V., & Chilton, L. B. (2023). Design Guidelines for Prompt Engineering of Text-to-Image Generative Models. *arXiv preprint*. <https://doi.org/10.48550/arXiv.2109.06977>
- Llewelyn, S. P., Elliott, R., Shapiro, D. A., Hardy, G., & Firth-Cozens, J. (1988). Client Perceptions of Significant Events in Prescriptive and Exploratory Phases of Individual Therapy. *The British Journal of Clinical Psychology*, 27(2), 105–114. <https://doi.org/10.1111/j.2044-8260.1988.tb00758.x>
- Llewelyn, S., Macdonald, J., & Aafjes-van Doorn, K. (2016). Process–outcome studies. In J. C. Norcross, G. R. VandenBos, D. K. Freedheim, & B. O. Olatunji (Eds.), *APA handbook of clinical psychology: Theory and research* (pp. 451–463). American Psychological Association. <https://doi.org/10.1037/14773-020>
- Lopez, K. A., & Willis, D. G. (2004). Descriptive versus interpretive phenomenology: Their contributions to nursing knowledge. *Qualitative Health Research*, 14(5), 726–735. <https://doi.org/10.1177/1049732304263638>
- Luborsky, L., Singer, B., & Luborsky, L. (1975). Comparative studies of psychotherapies: Is it true that “everyone has won and all must have prizes?” *Archives of General Psychiatry*, 32(8), 995–1008. <https://doi.org/10.1001/archpsyc.1975.01760260059004>
- Luborsky, L., Barber, J. P., Siqueland, L., Johnson, S., Najavits, L. M., Frank, A., & Daley, D. (1996). The Revised Helping Alliance Questionnaire (HAQ-II): Psychometric Properties. *The Journal of psychotherapy practice and research*, 5(3), 260–271. PMCID: [PMC3330423](https://pubmed.ncbi.nlm.nih.gov/3330423/)
- Luborsky, L., Diguier, L., Seligman, D. A., Rosenthal, R., Krause, E. D., Johnson, S., Halperin, G., Bishop, M., Berman, J. S., & Schweizer, E. (1999). The researcher's own therapy allegiances: A "wild card" in comparisons of treatment efficacy. *Clinical Psychology: Science and Practice*, 6(1), 95–106. <https://doi.org/10.1093/clipsy.6.1.95>

- Luckin, R., & Holmes, W. (2016). *Intelligence unleashed: An argument for AI in education*. Pearson UK.
- Lum, K., & Isaac, W. (2016). To predict and serve? *Significance*, 13(5), 14–19. DOI:10.1111/j.1740-9713.2016.00960.x
- Mairesse, F., Walker, M. A., Mehl, M. R., & Moore, R. K. (2007). Using linguistic cues for the automatic recognition of personality in conversation and text. *Journal of Artificial Intelligence Research*, 30, 457–500. <https://doi.org/10.1613/jair.2349>
- Malgaroli, M., Hull, T. D., Zech, J. M., & Althoff, T. (2023). Natural language processing for mental health interventions: a systematic review and research framework. *Translational psychiatry*, 13(1), 309. <https://doi.org/10.1038/s41398-023-02592-2>
- Manning, C. D., & Schütze, H. (1999). *Foundations of statistical natural language processing*. MIT Press.
- Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval*. Cambridge University Press. <https://nlp.stanford.edu/IR-book/>
- Marcus, G. (2020). The next decade in AI: Four steps towards robust artificial intelligence. *arXiv preprint*. <https://arxiv.org/abs/2002.06177>
- MarketsandMarkets. (2023). Generative AI market size, trends, & technology roadmap. Retrieved from <https://www.marketsandmarkets.com/Market-Reports/generative-ai-market-142870584.html>
- Mathur, S. V., Jindal, A. R., & Shrivastava, M. (2024). DaVinci at SemEval-2024 Task 9: Few-shot prompting GPT-3.5 for unconventional reasoning. *arXiv preprint*. <https://doi.org/10.48550/arXiv.2405.11559>
- MathWorks. (2021). MATLAB (Version R2021) [Computer software]. <https://www.mathworks.com>
- McCallum, A. and Nigam, K. (1998) A Comparison of Event Models for Naive Bayes Text Classification. *Proceedings in Workshop on Learning for Text Categorization*, AAAI'98, 41-48.
- McCarthy, J., Minsky, M. L., Rochester, N., & Shannon, C. E. (2006). A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence, August 31, 1955. *AI magazine*, 27(4), 12-12.
- McCarthy, K. L., Caputi, P., & Grenyer, B. F. S. (2017). Significant change events in psychodynamic psychotherapy: Is cognition or emotion more important? *Psychology and psychotherapy*, 90(3), 377–388. <https://doi.org/10.1111/papt.12116>
- McCrae, R. R., & Costa, P. T., Jr. (1999). A Five-Factor theory of personality. In L. A. Pervin & O. P. John (Eds.), *Handbook of personality: Theory and research* (2nd ed., pp. 139–153). Guilford Press.
- McLeod, J. (2011). *Qualitative research in counselling and psychotherapy* (2nd ed.). London; Thousand Oaks, CA: SAGE Publications.
- McLeod, J. (2013). *An introduction to counselling*. McGraw-Hill Education (UK).

- Meehl, P. E. (1954). *Clinical versus statistical prediction: A theoretical analysis and a review of the evidence*. University of Minnesota Press.
- Mergenthaler, E., & Stinson, C. H. (1992). Psychotherapy transcription standards. *Psychotherapy Research*, 2(2), 125–142. <https://doi.org/10.1080/10503309212331332904>
- Mergenthaler E. (1996). Emotion-abstraction patterns in verbatim protocols: a new way of describing psychotherapeutic processes. *Journal of consulting and clinical psychology*, 64(6), 1306–1315. <https://doi.org/10.1037//0022-006x.64.6.1306>
- Mergenthaler E. (2008). Resonating minds: a school-independent theoretical conception and its empirical application to psychotherapeutic processes. *Psychotherapy research: journal of the Society for Psychotherapy Research*, 18(2), 109–126. <https://doi.org/10.1080/10503300701883741>
- Meskó, B. (2023). Prompt engineering as an important emerging skill for medical professionals: Tutorial. *Journal of Medical Internet Research*, 25, e50638. <https://doi.org/10.2196/50638>
- Meystre, S. M., Savova, G. K., Kipper-Schuler, K. C., & Hurdle, J. F. (2008). Extracting information from textual documents in the electronic health record: a review of recent research. *Yearbook of medical informatics*, 128–144.
- Miao, Y., Yu, L., & Blunsom, P. (2016). Neural variational inference for text processing. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning* (pp. 1727–1736).
- Mikolov, T., Karafiát, M., Burget, L., Černocký, J., Khudanpur, S. (2010). Recurrent neural network-based language model. *Proc. Interspeech 2010*, 1045–1048, doi: 10.21437/Interspeech. 2010-343
- Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint*. <https://arxiv.org/abs/1301.3781>
- Mimno, D., Wallach, H. M., Talley, E., Leenders, M., & McCallum, A. (2011). Optimizing semantic coherence in topic models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (pp. 262–272). <https://aclanthology.org/D11-1024/>
- Miner, A. S., Fleming, S. L., Haque, A., Fries, J. A., Althoff, T., Wilfley, D. E., Agras, W. S., Milstein, A., Hancock, J., Asch, S. M., Stirman, S. W., Arnou, B. A., & Shah, N. H. (2022). A computational approach to measure the linguistic characteristics of psychotherapy timing, responsiveness, and consistency. *Npj mental health research*, 1(1), 19. <https://doi.org/10.1038/s44184-022-00020-9>
- Mitchell, T. M. (1997). *Machine learning*. McGraw-Hill.
- Mohammad, S. M., & Turney, P. D. (2012). Crowdsourcing a word–emotion association lexicon. *Computational Intelligence*, 29(3), 436–465. <https://doi.org/10.1111/j.1467-8640.2012.00460.x>
- Na, H., Hua, Y., Wang, Z., & Shen, T. (2025). A survey of large language models in psychotherapy: Current landscape and future directions. *arXiv*. <https://doi.org/10.48550/arXiv.2502.11095>
- Nathan, P. E., & Gorman, J. M. (Eds.). (2015). *A guide to treatments that work*. Oxford University.

- Newell, A., & Simon, H. A. (1976). Computer science as empirical inquiry: Symbols and search. *Communications of the ACM*, 19(3), 113–126. <https://doi.org/10.1145/360018.360022>
- Nilsson, N. J. (2010). The quest for artificial intelligence: A history of ideas and achievements. *Cambridge University Press*. <https://ai.stanford.edu/~nilsson/QAI/qai.pdf>
- Norcross, J. C., & Lambert, M. J. (2011). Psychotherapy relationships that work II. *Psychotherapy (Chicago, Ill.)*, 48(1), 4–8. <https://doi.org/10.1037/a0022180>
- Norcross, J. C., & Wampold, B. E. (2011). Evidence-Based Therapy Relationships: Research Conclusions and Clinical Practices. *Psychotherapy (Chicago, Ill.)*, 48(1), 98–102. <https://doi.org/10.1037/a0022161>
- OECD (2019). *Artificial intelligence in society*. OECD Publishing.
- Oliveira, A., Sousa, D., & Pires, A. P. (2012). Significant Events in Existential Psychotherapy: The Client's Perspective. *Existential Analysis: Journal of the Society for Existential Analysis*, 23(2).
- Orlinsky, D. E., Grawe, K., & Parks, B. K. (1994). Process and outcome in psychotherapy—Noch einmal. In A. Bergin, & J. S. Garfield (Eds.), *Handbook of psychotherapy and behaviour change* (4th edition, pp. 270–378). New York: Wiley.
- Orlinsky, D., Rønnestad, M. H., & Willutzki, U. (2004). Fifty years of psychotherapy process–outcome research: Continuity and change. In M. J. Lambert (Ed.), *Bergin and Garfield's handbook of psychotherapy and behavior change* (5th ed., pp. 307–390). Wiley NY.
- Oudin, A., Maatoug, R., Bourla, A., Ferreri, F., Bonnot, O., Millet, B., Schoeller, F., Mouchabac, S., & Adrien, V. (2023). Digital Phenotyping: Data-Driven Psychiatry to Redefine Mental Health. *Journal of medical Internet research*, 25, e44502. <https://doi.org/10.2196/44502>
- Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., & Lowe, R. (2022). Training language models to follow instructions with human feedback. *arXiv preprint*. <https://doi.org/10.48550/arXiv.2203.02155>
- Palmer, R. E. (1969). *Hermeneutics: Interpretation Theory in Schleiermacher, Dilthey, Heidegger, and Gadamer*. Northwestern University Press.
- Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up? Sentiment classification using machine learning techniques. *Proceedings of the ACL Conference*, 79–86. DOI: 10.3115/1118693.1118704
- Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1-2), 1–135. <https://doi.org/10.1561/15000000011>
- Pariser, E. (2011). *The filter bubble: What the Internet is hiding from you*. Penguin UK.
- Paul, G. L. (1967). Strategy of outcome research in psychotherapy. *Journal of Consulting Psychology*, 31(2), 109–118. <https://doi.org/10.1037/h0024436>

- Park, G., Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Kosinski, M., Stillwell, D. J., Ungar, L. H., & Seligman, M. E. (2015). Automatic personality assessment through social media language. *Journal of personality and social psychology*, 108(6), 934–952. <https://doi.org/10.1037/pspp0000020>
- Park, J. & Choo, S. (2024). Generative AI prompt engineering for educators: Practical strategies. *Journal of Special Education Technology*, 40(3). <https://doi.org/10.1177/01626434241298954>
- Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. Morgan Kaufmann. <https://dl.acm.org/doi/pdf/10.5555/534975>
- Pennebaker, J. W., Mehl, M. R., & Niederhoffer, K. G. (2003). Psychological aspects of natural language use: Our words, our selves. *Annual Review of Psychology*, 54, 547-577. <https://doi.org/10.1146/annurev.psych.54.101601.145041>
- Pennebaker, J. W., Boyd, R. L., Jordan, K., & Blackburn, K. (2015). The development and psychometric properties of LIWC2015. *University of Texas at Austin*.
- Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532-1543.
- Peräkylä, A., Antaki, C., Vehviläinen, S., & Leudar, I. (Eds.). (2008). *Conversation analysis and psychotherapy*. Cambridge University Press. <https://doi.org/10.1017/CBO9780511490002>
- Periti, F., & Tahmasebi, N. (2024). A systematic comparison of contextualized word embeddings for lexical semantic change. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)* (pp. 4262–4282). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2024.naacl-long.240>
- Plutchik, R. (1980). A general psychoevolutionary theory of emotion. In R. Plutchik & H. Kellerman (Eds.), *Theories of Emotion* (pp. 3–33). Academic Press. <https://doi.org/10.1016/B978-0-12-558701-3.50007-7>
- Polkinghorne, D. E. (1988). *Narrative knowing and the human sciences*. State University of NYC.
- Pryzant, R., Iter, D., Li, J., Lee, Y. T., Zhu, C., & Zeng, M. (2023). Automatic prompt optimization with “gradient descent” and beam search. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 7957–7968). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2023.emnlp-main.494>
- Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving Language Understanding by Generative Pre-Training. *OpenAI Technical Report*.
- Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI Blog*. <https://openai.com/blog/better-language-models>
- Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. (2021). Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning* (pp. 8748–8763). PMLR. <https://doi.org/10.48550/arXiv.2103.00020>

- R Core Team. (2023). R: A language and environment for statistical computing [Software]. R Foundation for Statistical Computing. <https://www.R-project.org/>.
- Rennie, D. L. (1994). Clients' deference in psychotherapy. *Journal of Counseling Psychology*, 41(4), 427–437. <https://doi.org/10.1037/0022-0167.41.4.427>
- Reuter, A., Thielmann, A., Weisser, C., Safken, B., & Kneib, T. (2025). Probabilistic Topic Modeling With Transformer Representations. *IEEE transactions on neural networks and learning systems*, 36(8), 14551–14565. <https://doi.org/10.1109/TNNLS.2025.3538262>
- Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1144. <https://doi.org/10.1145/2939672.2939778>
- Ricci, F., Rokach, L., & Shapira, B. (2015). Recommender systems: Introduction and challenges. In *Recommender systems handbook* (pp. 1–34). Springer. https://doi.org/10.1007/978-1-4899-7637-6_1
- Ricoeur, P. (1976). *Interpretation theory: Discourse and the surplus of meaning*. Texas Christian University Press.
- Ricoeur, P. (1981). *Hermeneutics and the human sciences*. Cambridge University Press.
- Ricoeur, P. (1986). *Time and narrative* (Vol. 3). University of Chicago Press.
- Ricoeur, P. (1992). *Oneself as another*. University of Chicago press.
- Rogers, C. R. (1942). *Counseling and psychotherapy; newer concepts in practice*. Houghton Mifflin.
- Rosenzweig, S. (1936). Some implicit common factors in diverse methods of psychotherapy. *American Journal of Orthopsychiatry*, 6(3), 412–415. <https://doi.org/10.1111/j.1939-0025.1936.tb05248.x>
- Rosen-Zvi, M., Chemudugunta, C., Griffiths, T., Smyth, P., & Steyvers, M. (2010). Learning author-topic models from text corpora. *ACM Transactions on Information Systems*, 28(1), Article 4, 1–38. <https://doi.org/10.1145/1658377.1658381>
- Roth, A., & Fonagy, P. (2005). *What works for whom: A critical review of psychotherapy research* (2nd ed.). Guilford Publications.
- Royce, C. S., Hayes, M. M., & Schwartzstein, R. M. (2019). Teaching critical thinking: A case for instruction in cognitive biases to reduce diagnostic errors and improve patient safety. *Academic Medicine*, 94(2), 187–194. <https://doi.org/10.1097/ACM.0000000000002518>
- Rubin, T. N., Chambers, A., Smyth, P., & Steyvers, M. (2012). Statistical topic models for multi-label document classification. *Machine Learning*, 88(1-2), 157–208. <https://doi.org/10.1007/s10994-011-5272-5>
- Rumelhart, D. E., Hinton, G. E., & McClelland, J. L. (1986). A general framework for parallel distributed processing. *Parallel distributed processing: Explorations in the microstructure of cognition*, 1(45-76), 26.

- Rumshisky, A., Ghassemi, M., Naumann, T., Szolovits, P., Castro, V. M., McCoy, T. H., & Perlis, R. H. (2016). Predicting early psychiatric readmission with natural language processing of narrative discharge summaries. *Translational psychiatry*, 6(10), e921. <https://doi.org/10.1038/tp.2015.182>
- Rush, A. J., Beck, A. T., Kovacs, M., & Hollon, S. D. (1977). Comparative efficacy of cognitive therapy and pharmacotherapy in the treatment of depressed outpatients. *Cognitive Therapy and Research*, 1(1), 17–37. <https://doi.org/10.1007/BF01173502>
- Russell, R. L., & Orlinsky, D. E. (1996). Psychotherapy research in historical perspective. Implications for mental health care policy. *Archives of general psychiatry*, 53(8), 708–715. <https://doi.org/10.1001/archpsyc.1996.01830080060010>
- Russell, S. J., & Norvig, P. (2021). *Artificial Intelligence: A Modern Approach*, Global Edition 4e.
- Sacks, H., Schegloff, E. A., & Jefferson, G. (1974). A simplest systematics for the organization of turn-taking for conversation. *Language*, 50(4), 696–735. <https://doi.org/10.1353/lan.1974.0010>
- Safran, J. D., & Muran, J. C. (2000). *Negotiating the therapeutic alliance: A relational treatment guide*. Guilford Press.
- Sahoo, P., Singh, A. K., Saha, S., Jain, V., Mondal, S., & Chadha, A. (2025). A systematic survey of prompt engineering in large language models: Techniques and applications. *arXiv preprint*. <https://doi.org/10.48550/arXiv.2402.07927>
- Saldanha, G., & O'Brien, S. (2014). *Research methodologies in translation studies*. Routledge.
- Salemi, M. H., Foroozandeh, E., & Ashkzari, M. K. (2025). Applications, Challenges, and Future Perspectives of Artificial Intelligence in Psychopharmacology, Psychological Disorders and Physiological Psychology: A Comprehensive Review. *Journal of pharmacy & bioallied sciences*, 17(Suppl 1), S229–S233. https://doi.org/10.4103/jpbs.jpbs_1765_24
- Salton, G., Wong, A., & Yang, C. S. (1975). A vector space model for automatic indexing. *Communications of the ACM*, 18(11), 613–620. <https://doi.org/10.1145/361219.361220>
- Salvatore, S., Gennaro, A., Auletta, A. F., Tonti, M., & Nitti, M. (2012). Automated method of content analysis: a device for psychotherapy process research. *Psychotherapy Research: Journal of the Society for Psychotherapy Research*, 22(3), 256–273. <https://doi.org/10.1080/10503307.2011.647930>
- Salvatore, S., Gelo, O. C., Gennaro, A., Metrangolo, R., Terrone, G., Pace, V., Venuleo, C., Venezia, A., & Ciavolino, E. (2017). An automated method of content analysis for psychotherapy research: A further validation. *Psychotherapy research: journal of the Society for Psychotherapy Research*, 27(1), 38–50. <https://doi.org/10.1080/10503307.2015.1072282>
- Schiepek, G., Gelo, O., Viol, K., Kratzer, L., Orsucci, F., de Felice, G., Stöger-Schmidinger, B., Sammet, I., Aichhorn, W., & Schöller, H. (2020). Complex individual pathways or standard tracks? A data-based discussion on the trajectories of change in psychotherapy. *Counselling and Psychotherapy Research*, 20(4), 689–702. <https://doi.org/10.1002/capr.12300>
- Schillinger, D., Balyan, R., Crossley, S. A., McNamara, D. S., Liu, J. Y., & Karter, A. J. (2021). Employing computational linguistics techniques to identify limited patient health literacy: Findings

from the ECLIPPSE study. *Health services research*, 56(1), 132–144. <https://doi.org/10.1111/1475-6773.13560>

Schulhoff, S., Ilie, M., Balepur, N., Kahadze, K., Liu, A., Si, C., Li, Y., Gupta, A., Han, H., Schulhoff, S., Dulepet, P. S., Vidyadhara, S., Ki, D., Agrawal, S., Pham, C., Kroiz, G., Li, F., Tao, H., Srivastava, A., ... Resnik, P. (2024). The prompt report: A systematic survey of prompting techniques (arXiv:2406.06608). *arXiv*.

Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Ramones, S. M., Agrawal, M., Shah, A., Kosinski, M., Stillwell, D., Seligman, M. E., & Ungar, L. H. (2013). Personality, gender, and age in the language of social media: the open-vocabulary approach. *PloS one*, 8(9), e73791. <https://doi.org/10.1371/journal.pone.0073791>

Sennrich, R., Haddow, B., & Birch, A. (2016). Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics* (pp. 1715–1725). <https://doi.org/10.18653/v1/P16-1162>

Shah, C. (2025). From prompt engineering to prompt science with humans in the loop. *Communications of the ACM*, 68(6), 54–61. <https://doi.org/10.1145/3709599>

Shedler J. (2010). The efficacy of psychodynamic psychotherapy. *The American psychologist*, 65(2), 98–109. <https://doi.org/10.1037/a0018378>

Shortliffe, E. H. (1976). Computer-based medical consultations: MYCIN. *Journal of Clinical Engineering*, 388(1). <https://doi.org/10.1097/00004669-197610000-00011>

Shortliffe, E. H., & Sepúlveda, M. J. (2018). Clinical decision support in the era of artificial intelligence. *JAMA*, 320(21), 2199–2200. <https://doi.org/10.1001/jama.2018.17163>

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., Dieleman, S., Grewe, D., Nham, J., Kalchbrenner, N., Sutskever, I., Lillicrap, T., Leach, M., Kavukcuoglu, K., Graepel, T., & Hassabis, D. (2016). Mastering the game of Go with deep neural networks and tree search. *Nature*, 529(7587), 484–489. <https://doi.org/10.1038/nature16961>

Silverman, D. (2021). *Doing qualitative research* (6th ed.). SAGE Publications Ltd.

Simon, G. E., & Ludman, E. J. (2010). Predictors of early dropout from psychotherapy for depression in community practice. *Psychiatric Services*, 61(7), 684–689. <https://doi.org/10.1176/ps.2010.61.7.684>

Škrlj, B., Martinc, M., Kralj, J., Lavrač, N., & Pollak, S. (2021). tax2vec: Constructing interpretable features from taxonomies for short text classification. *Computer Speech & Language*, 65, 101104. <https://doi.org/10.1016/j.csl.2020.101104>

Smink, W. A. C., Fox, J. P., Tjong Kim Sang, E., Sools, A. M., Westerhof, G. J., & Veldkamp, B. P. (2019). Understanding Therapeutic Change Process Research Through Multilevel Modeling and Text Mining. *Frontiers in psychology*, 10, 1186. <https://doi.org/10.3389/fpsyg.2019.01186>

Smith, M. L., & Glass, G. V. (1977). Meta-analysis of psychotherapy outcome studies. *American Psychologist*, 32(9), 752–760. <https://doi.org/10.1037/0003-066X.32.9.752>

- Smith, M. L., Glass, G. V., & Miller, T. I. (1980). *The benefits of psychotherapy*. Baltimore, MD: John Hopkins University Press.
- Smith, G. T., Atkinson, E. A., Davis, H. A., Riley, E. N., & Oltmanns, J. R. (2020). The General Factor of Psychopathology. *Annual review of clinical psychology*, 16, 75–98. <https://doi.org/10.1146/annurev-clinpsy-071119-115848>
- Spearman C. (2010). The proof and measurement of association between two things. *International journal of epidemiology*, 39(5), 1137–1150. <https://doi.org/10.1093/ije/dyq191>
- Sperandeo, R., Messina, G., Iennaco, D., Sessa, F., Russo, V., Polito, R., Monda, V., Monda, M., Messina, A., Mosca, L. L., Mosca, L., Dell'Orco, S., Moretto, E., Gigante, E., Chiacchio, A., Scognamiglio, C., Carotenuto, M., & Maldonato, N. M. (2020). What Does Personality Mean in the Context of Mental Health? A Topic Modeling Approach Based on Abstracts Published in Pubmed Over the Last 5 Years. *Frontiers in psychiatry*, 10, 938. <https://doi.org/10.3389/fpsy.2019.00938>
- Srivastava, A., & Sutton, C. (2017). Autoencoding variational inference for topic models. *arXiv preprint*. <https://doi.org/10.48550/arXiv.1703.01488>
- Stanghellini G. (2010). A hermeneutic framework for psychopathology. *Psychopathology*, 43(5), 319–326. <https://doi.org/10.1159/000319401>
- Stanghellini, G. (2016). *An experiential approach to psychopathology: What is it like to suffer from mental disorders?* Springer. <https://doi.org/10.1007/978-3-319-29945-7>
- Steinbach, M., Karypis, G., and Kumar, V. (2000). A Comparison of Document Clustering Techniques. *KDD Workshop on Text Mining*, 400, 525-526.
- Steyvers, M., & Griffiths, T. (2007). Probabilistic topic models. In *Handbook of latent semantic analysis* (pp. 439-460). Psychology Press.
- Stiles, W. B., & Sultan, F. E. (1979). Verbal response mode used by clients in psychotherapy. *Journal of Consulting and Clinical Psychology*, 47(3), 611–613. <https://doi.org/10.1037/0022-006X.47.3.611>
- Stiles, W. B. (1980). Measurement of the impact of psychotherapy sessions. *Journal of Consulting and Clinical Psychology*, 48(2), 176–185. <https://doi.org/10.1037/0022-006X.48.2.176>
- Stiles, W. B. (1992). *Describing talk: A taxonomy of verbal response modes*. Newbury Park, Sage.
- Stone, P., Brooks, R., Brynjolfsson, E., Calo, R., Etzioni, O., Hager, G., ... & Teller, A. (2022). Artificial Intelligence and Life in 2030: The One-Hundred-Year Study on Artificial Intelligence. *arXiv preprint*. <https://doi.org/10.48550/arXiv.2211.06318>
- Straw, I. (2021). Ethical implications of emotion mining in medicine. *Health Policy and Technology*, 10(1), 191–195. <https://doi.org/10.1016/j.hlpt.2020.11.006>
- Strubell, E., Ganesh, A., & McCallum, A. (2019). Energy and policy considerations for deep learning in NLP. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 3645–3650. <https://doi.org/10.18653/v1/P19-1355>

- Strupp, H. H. (1986). Psychotherapy: Research, practice, and public policy: How to avoid dead ends. *American Psychologist*, 41(2), 120–130. <https://doi.org/10.1037/0003-066X.41.2.120>
- Strupp, H. H., & Howard, K. I. (1992). A brief history of psychotherapy research. In D. K. Freedheim, H. J. Freudenberger, J. W. Kessler, S. B. Messer, D. R. Peterson, H. H. Strupp, & P. L. Wachtel (Eds.), *History of psychotherapy: A century of change* (pp. 309–334). APA.
- Su, S.-Y., Huang, C.-W., & Chen, Y.-N. (2019). Dual supervised learning for natural language understanding and generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 5472–5477). Association for Computational Linguistics. <https://doi.org/10.18653/v1/P19-1545>
- Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction* (2nd ed.). MIT Press.
- Ding, Z., Liu, W., Zhang, G., & Wang, H. (2018). Supervisor Narcissism and Time Theft: Investigating the Mediating Roles of Emotional Exhaustion and the Moderating Roles of Attachment Style. *Frontiers in psychology*, 9, 2215. <https://doi.org/10.3389/fpsyg.2018.02215>
- Tanana, M., Hallgren, K. A., Imel, Z. E., Atkins, D. C., & Srikumar, V. (2016). A Comparison of Natural Language Processing Methods for Automated Coding of Motivational Interviewing. *Journal of substance abuse treatment*, 65, 43–50. <https://doi.org/10.1016/j.jsat.2016.01.006>
- Tanana, M. J., Soma, C. S., Kuo, P. B., Bertagnolli, N. M., Dembe, A., Pace, B. T., Srikumar, V., Atkins, D. C., & Imel, Z. E. (2021). How do you feel? Using natural language processing to automatically rate emotion in psychotherapy. *Behavior research methods*, 53(5), 2069–2082. <https://doi.org/10.3758/s13428-020-01531-z>
- Tausczik, Y. R., & Pennebaker, J. W. (2010). The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. *Journal of Language and Social Psychology*, 29(1), 24–54. <https://doi.org/10.1177/0261927X09351676>
- Toleu, A., Mussabayev, N., & Tolegen, G. (2019). KeyVector: Unsupervised keyphrase extraction using weighted topic via semantic relatedness. *Computación y Sistemas*, 23(3), 861–869. <https://doi.org/10.13053/CyS-23-3-3264>
- Topol, E. J. (2019). High-performance medicine: The convergence of human and artificial intelligence. *Nature Medicine*, 25(1), 44–56. <https://doi.org/10.1038/s41591-018-0300-7>
- Traini, M. (2008). *The Language of Psychotherapy: Linguistic and Hermeneutic Approaches to Texts*. Franco Angeli.
- Truijens, F., Zühlke-van Hulzen, L., & Vanheule, S. (2019). To manualize, or not to manualize: Is that still the question? A systematic review of empirical evidence for manual superiority in psychological treatment. *Journal of Clinical Psychology*, 75(3), 329–343. <https://doi.org/10.1002/jclp.22712>
- Tsakalidis, A., Atzil-Slonim, D., Polakovski, A., Shapira, N., Tuval-Mashiach, R., & Liakata, M. (2021). Automatic identification of ruptures in transcribed psychotherapy sessions. In *Proceedings of the Seventh Workshop on Computational Linguistics and Clinical Psychology: Improving Access (CLPsych 2021)* (pp. 55–69). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2021.clpsych-1.15>

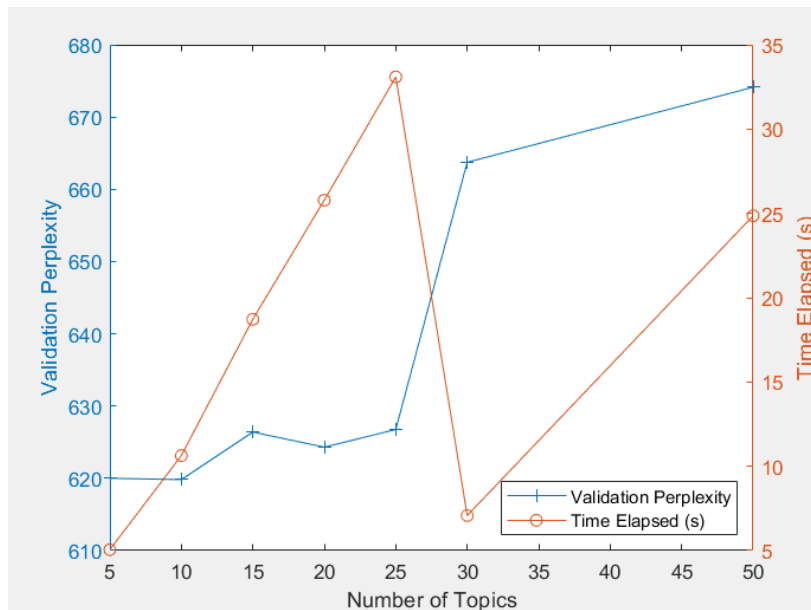
- Tumasjan, A., Sprenger, T. O., Sandner, P. G., & Welpe, I. M. (2010). Predicting elections with Twitter: What 140 characters reveal about political sentiment. *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media (ICWSM)*, 4(1). <https://doi.org/10.1609/icwsml.v4i1.14009>
- Turing, A. M. (1950). Computing machinery and intelligence. *Mind*, 59(236), 433-460. <https://doi.org/10.1093/mind/LIX.236.433>
- Turney, P. D. (2002). Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 417–424. <https://aclanthology.org/P02-1053/>
- Udvardi A. (2019). The role of linguistics in improving the evidence base of healthcare communication. *Patient education and counselling*, 102(2), 388–393. <https://doi.org/10.1016/j.pec.2018.09.012>
- Van Staden, C. W., & Fulford, K. W. M. (2004). Changes in semantic uses of first-person pronouns as possible linguistic markers of recovery in psychotherapy. *Australasian Psychiatry*, 38(4), 226-232. <https://doi.org/10.1080/j.1440-1614.2004.01339.x>
- Vanin, A., Bolshev, V., & Panfilova, A. (2025). Psychotherapist remarks' ML classifier: insights from LLM and topic modeling application. *Frontiers in psychiatry*, 16, 1608163. <https://doi.org/10.3389/fpsy.2025.1608163>
- Van Rossum, G., & Drake, F. L. (2009). *Python 3 Reference Manual*. Scotts Valley, CA: CreateSpace.
- Vapnik, V. N. (1998). *Statistical learning theory*. Wiley.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30. <https://arxiv.org/abs/1706.03762>
- Wallach, H. M., Murray, I., Salakhutdinov, R., & Mimno, D. (2009). Evaluation methods for topic models. In *Proceedings of the 26th Annual International Conference on Machine Learning* (pp. 1105–1112). <https://doi.org/10.1145/1553374.1553515>
- Wampold, B. E., & Imel, Z. E. (2015). *The great psychotherapy debate: The evidence for what makes psychotherapy work* (2nd ed.). Routledge/Taylor & Francis Group.
- Wang, D., & Barabási, A. L. (2021). *The science of science*. Cambridge University Press.
- Wang, J., Shi, E., Yu, S., Wu, Z., Hu, H., Ma, C., Dai, H., Yang, Q., Kang, Y., Wu, J., Yue, C., Zhang, H., Liu, Y., Pan, Y., Liu, Z., Sun, L., Li, X., Ge, B., Jiang, X., ... Zhang, S. (2025). Prompt engineering for healthcare: Methodologies and applications. *Meta-Radiology*, 100190. <https://doi.org/10.1016/j.metrad.2025.100190>
- Watson J. C. (2025). Psychotherapy process research: Identifying productive in-session processes to enhance treatment outcomes and therapist responsiveness. *Psychotherapy research: journal of the Society for Psychotherapy Research*, 35(1), 4–16. <https://doi.org/10.1080/10503307.2023.2252160>

- Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E. H., Le, Q. V., & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. *arXiv preprint*. <https://doi.org/10.48550/arXiv.2201.11903>
- Westen, D., & Weinberger, J. (2004). When clinical description becomes statistical prediction. *American Psychologist*, 59(7), 595–613. <https://doi.org/10.1037/0003-066X.59.7.595>
- White, M., & Epston, D. (1990). *Narrative means to therapeutic ends*. WW Norton & Company.
- Wickham, H., & Grolemund, G. (2017). *R for data science* (Vol. 2). Sebastopol: O'Reilly.
- Wilson, T. D., & Dunn, E. W. (2004). Self-knowledge: Its limits, value, and potential for improvement. *Annual Review of Psychology*, 55, 493–518. <https://doi.org/10.1146/annurev.psych.55.090902.141954>
- Wood, S. N. (2017). *Generalized additive models: An introduction with R* (2nd ed.). CRC Press.
- Woolf, B. P. (2008). *Building intelligent interactive tutors: Student-centered strategies for revolutionizing e-learning*. Elsevier & Morgan Kaufmann.
- Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., ... & Dean, J. (2016). Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint*. <https://doi.org/10.48550/arXiv.1609.08144>
- Xiao, B., Imel, Z. E., Georgiou, P., Atkins, D. C., & Narayanan, S. S. (2016). Computational Analysis and Simulation of Empathic Behaviors: A Survey of Empathy Modeling with Behavioral Signal Processing Framework. *Current psychiatry reports*, 18(5), 49. <https://doi.org/10.1007/s11920-016-0682-5>
- Zhang, C., Lucas, J., Hinton, G., & Ba, J. (2021). Lookahead optimizer: k steps forward, 1 step back. *Neural Information Processing Systems*. <https://arxiv.org/abs/1907.08610>
- Zhang, Z., & Wang, J. (2024). Can AI replace psychotherapists? Exploring the future of mental health care. *Frontiers in Psychiatry*, 15, 1444382. <https://doi.org/10.3389/fpsy.2024.1444382>
- Zhou, Y., Muresanu, A. I., Han, Z., Paster, K., Pitis, S., Chan, H., & Ba, J. (2022). Large language models are human-level prompt engineers *arXiv preprint*. <https://doi.org/10.48550/arXiv.2211.01910>
- Zhu, K., Zhao, Q., Chen, H., Wang, J., & Xie, X. (2023). PromptBench: A unified library for evaluation of large language models. *arXiv preprint*. <https://doi.org/10.48550/arXiv.2312.07910>
- Zuur, A. F., Ieno, E. N., Walker, N., Saveliev, A. A., & Smith, G. M. (2009). *Mixed effects models and extensions in ecology with R*. Springer.

Appendices

Appendix 1

Perplexity Function. The validation perplexity (blue line) and time elapsed (orange line) are functions of the number of topics. Lower perplexity values indicate better algorithm performance. The perplexity is lowest between 5 and 15 topics. The processing time in this range is also reduced, highlighting the algorithm's efficiency. Beyond this range, perplexity increases, and processing time becomes more unstable.



Appendix 2

For each Session, the initial Number (No.) of Conversational Turn (CT) and the Final Number are presented. This is useful to interpret the graphic representation of Gamms that show the probability of a topic occurring during the different sessions, i.e., that are represented on the X-axis as the total number of conversational turns throughout the entire psychotherapy process.

Session	Initial No. of CT	Final No. of CT
1	2	279
2	280	462
3	463	652
4	653	868
5	869	1004
6	1005	1175
7	1176	1324
8	1325	1464
9	1465	1631
10	1632	1750
11	1751	1869
12	1870	2046
13	2047	2183
14	2184	2255

15	2256	2350
16	2351	2498
17	2499	2692
18	2693	2937
19	2938	2093
20	3094	3311
21	3312	3495
22	3496	3751
23	3752	3968
24	3969	4195
25	4196	4390
26	4391	4538

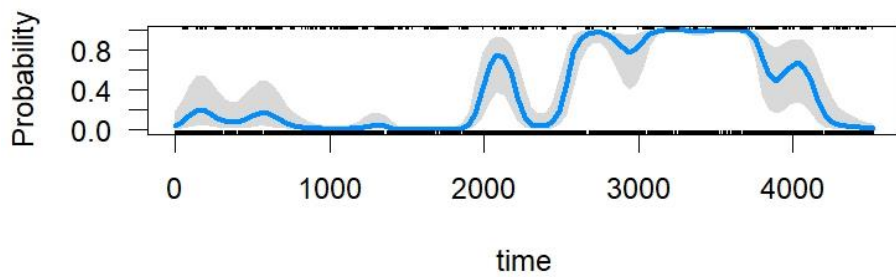
Appendix 3

Cohen's K index for each pair of evaluators. The agreement is measured against the topic assignment work performed by each evaluator for the entire text corpus, specifically 26 psychotherapeutic sessions. Cohen's k is a statistical index reflecting the degree of agreement between two raters who rate the same sample on a nominal scale with the same number of categories. Scores range from 0 (disagreement) to 1 (perfect agreement).

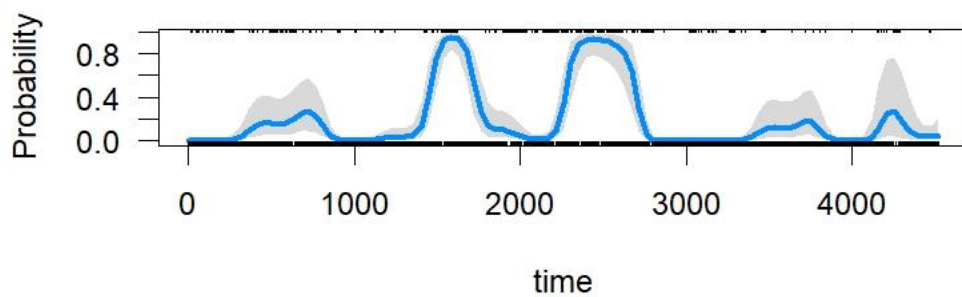
	Misura di accordo	Valore
r1*r2	Kappa	0,491
r1*r3	Kappa	0,582
r2*r3	Kappa	0,567

Appendix 4

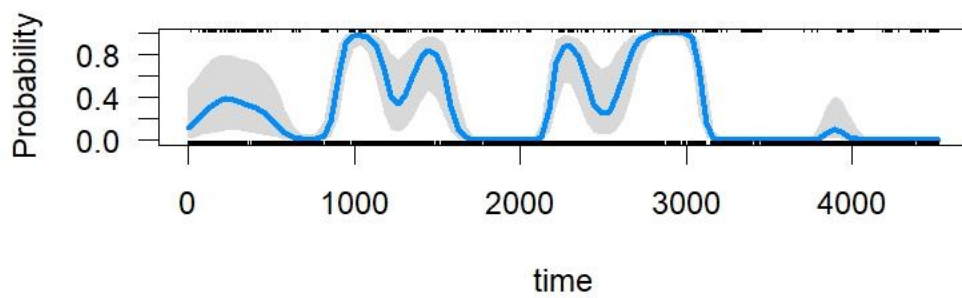
Graphical representation of the probability of occurrence of each topic listed in Table 3 along the entire treatment pathway (represented by the total number of conversational shifts, see Appendix 2). On the x-axis, time is defined by the sequence of conversational turns, and the y-axis displays the probability (ranging from 0 to 1) of each topic being addressed in the therapeutic sessions. The shaded grey area indicates the standard error of the estimated probability, providing a visual representation of the variability in the trend estimates.



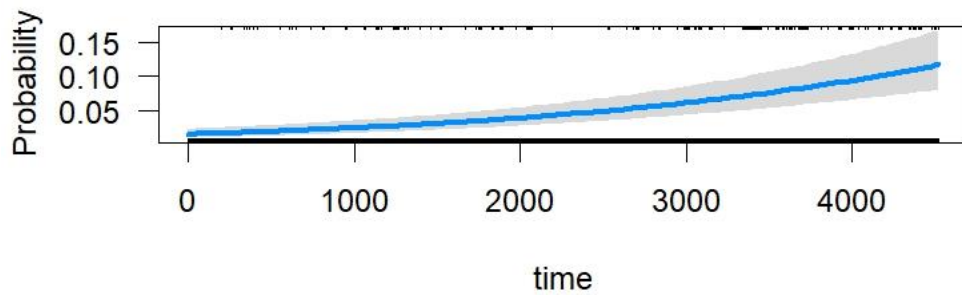
4.1. This Gamm illustrates the trajectory of Affective Relation (AR) during the present psychotherapy.



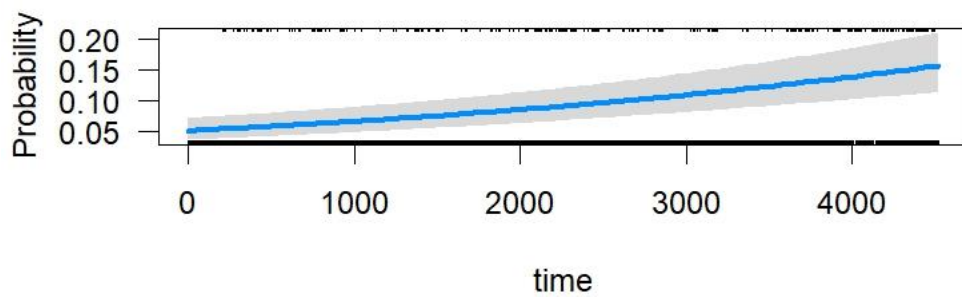
4.2. This Gamm illustrates the trajectory of Family (FAM) during the present psychotherapy.



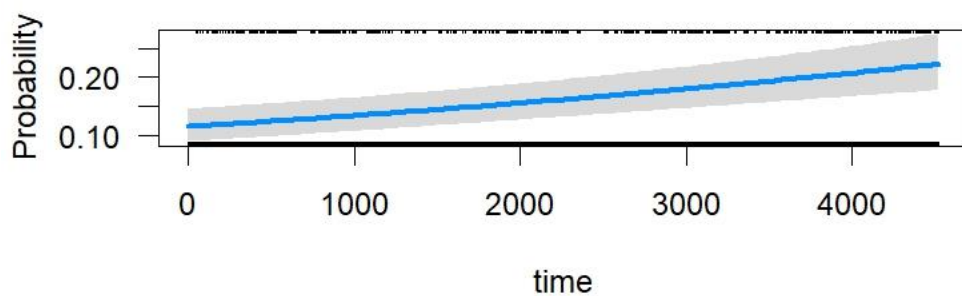
4.3. This Gamm illustrates the trajectory of Occupation/Study (OS) during the present psychotherapy.



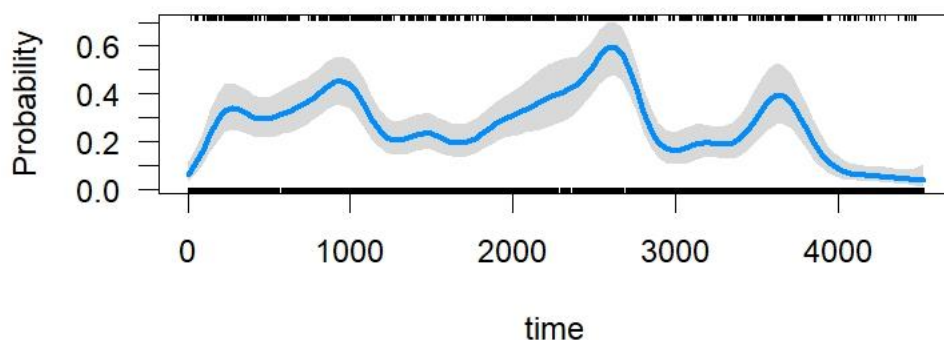
4.4. This Gamm illustrates the trajectory of Projectuality (PRJ) during the present psychotherapy.



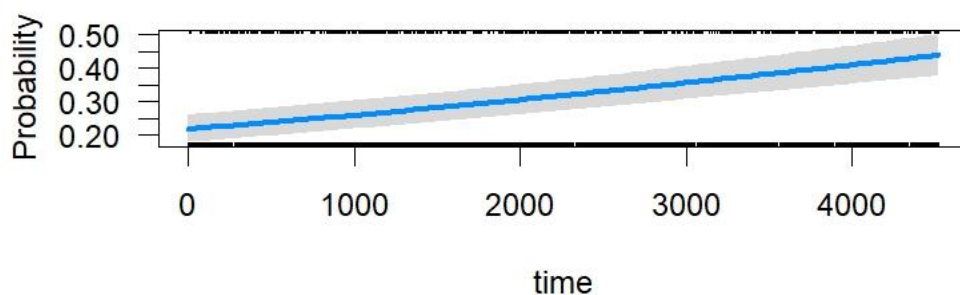
4.5. This Gamm illustrates the trajectory of Social Relations (SR) during the present psychotherapy.



4.6. This Gamm illustrates the trajectory of Ways of Being (WB) during the present psychotherapy.



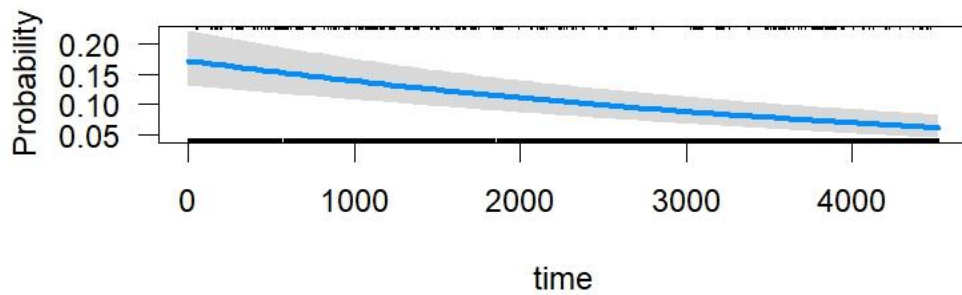
4.7. This Gamm illustrates the trajectory of Suffering/Pathology/Symptoms (SPS) during the present psychotherapy.



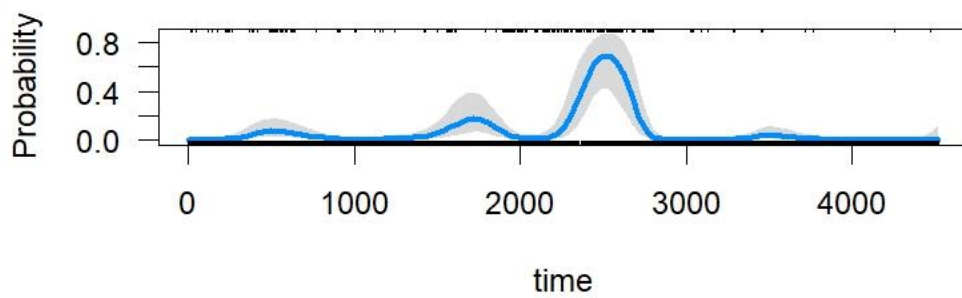
4.8. This Gamm illustrates the trajectory of Therapeutic Refiguration and Insight (TRI) during the present psychotherapy.

Appendix 5

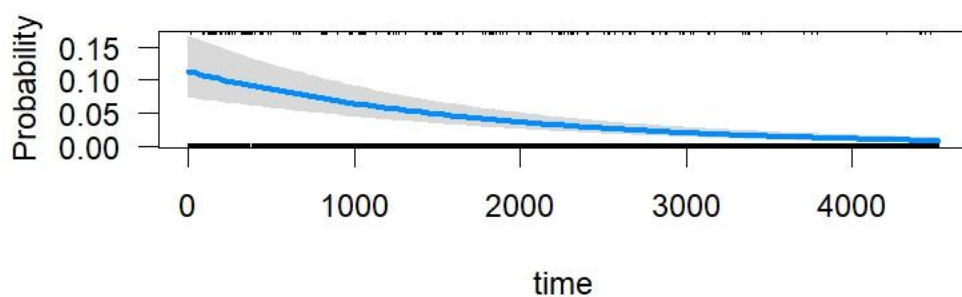
Graphical representation of the probability of co-occurrence of some significant topics along the entire treatment pathway of the case study examined. The x-axis represents time, measured by the progression of conversational turns during the therapeutic sessions, while the y-axis indicates the probability (ranging from 0 to 1) of the co-occurrence of these topics. The shaded grey area represents the standard error of the estimated probability.



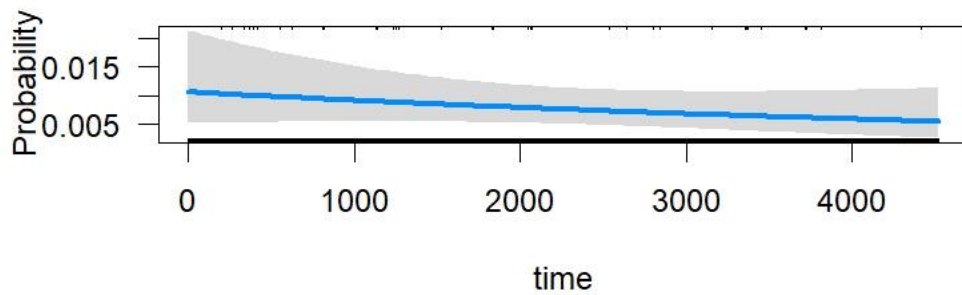
5.1. SPS_AR. This Gamm illustrates the co-occurrence trajectory of Suffering/Pathology/Symptoms (SPS) and Affective Relation (AR) during the present psychotherapy.



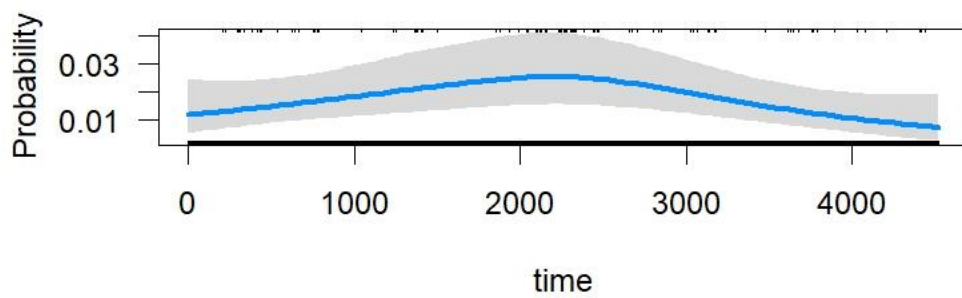
5.2. SPS_FAM. This Gamm illustrates the co-occurrence trajectory of Suffering/Pathology/Symptoms (SPS) and Family (FAM) during the present psychotherapy.



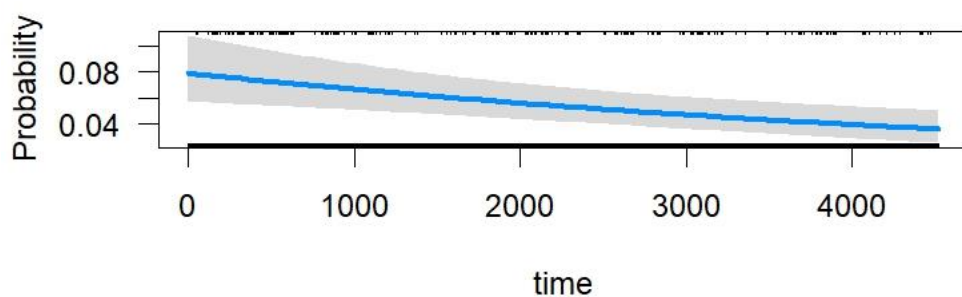
5.3. SPS_OS. This Gamm illustrates the co-occurrence trajectory of Suffering/Pathology/Symptoms (SPS) and Occupation/Study (OS) during the present psychotherapy.



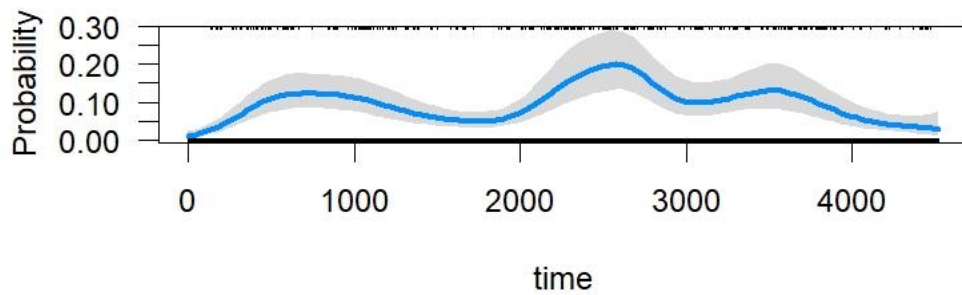
5.4. SPS_PRJ. This Gamm illustrates the co-occurrence trajectory of Suffering/Pathology/Symptoms (SPS) and Projectuality (PRJ) during the present psychotherapy.



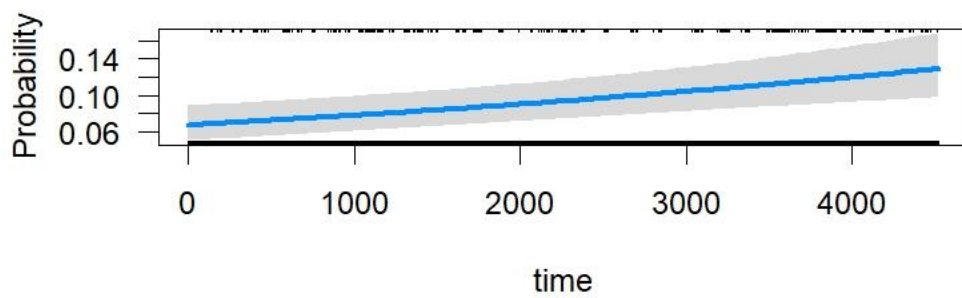
5.5. SPS_SR. This Gamm illustrates the co-occurrence trajectory of Suffering/Pathology/Symptoms (SPS) and Social Relations (SR) during the present psychotherapy.



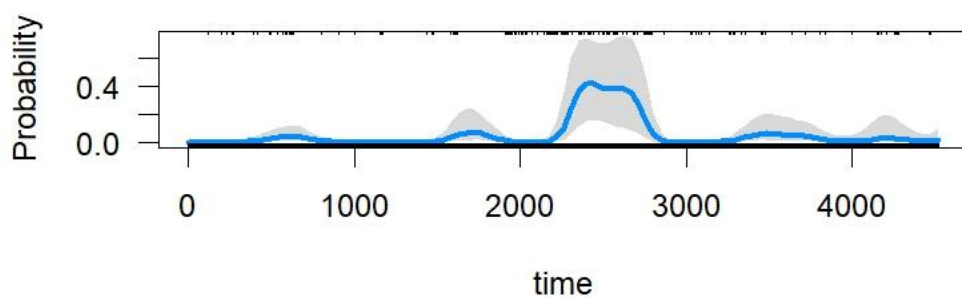
5.6. SPS_WB. This Gamm illustrates the co-occurrence trajectory of Suffering/Pathology/Symptoms (SPS) and Ways of Being (WB) during the present psychotherapy.



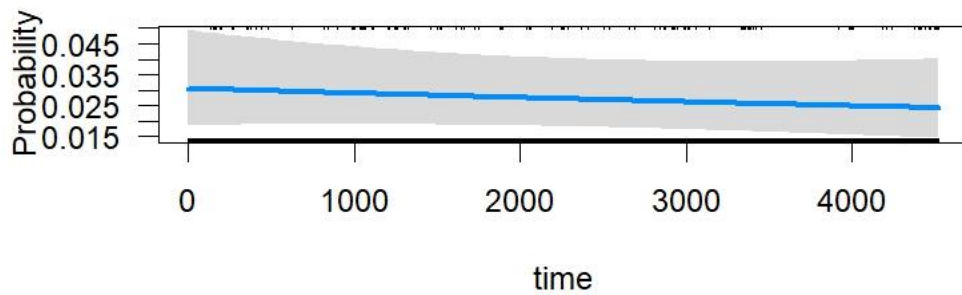
5.7. SPS_TRI. This Gamm illustrates the co-occurrence trajectory of Suffering/Pathology/Symptoms (SPS) and Therapeutic Refiguration and Insight (TRI) during the present psychotherapy.



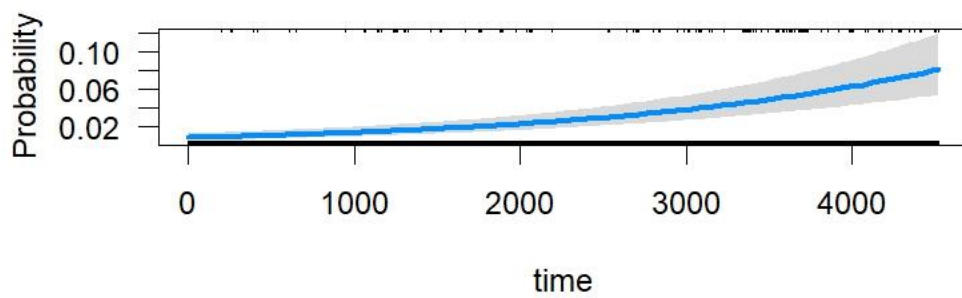
5.8. TRI_AR. This Gamm illustrates the co-occurrence trajectory of Therapeutic Refiguration and Insight (TRI) and Affective Relation (AR) during the present psychotherapy.



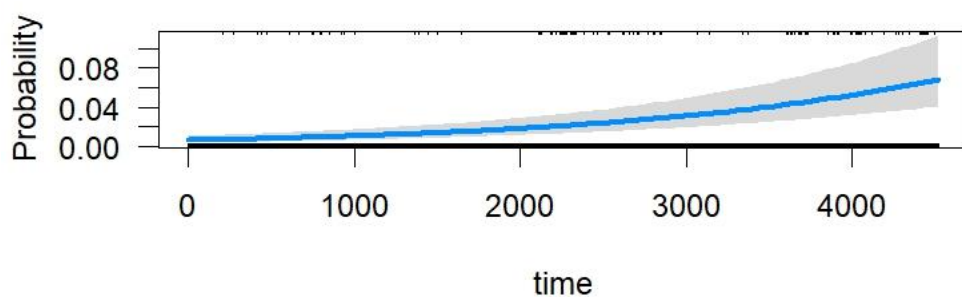
5.9. TRI_FAM. This Gamm illustrates the co-occurrence trajectory of Therapeutic Refiguration and Insight (TRI) and Family (FAM) during the present psychotherapy.



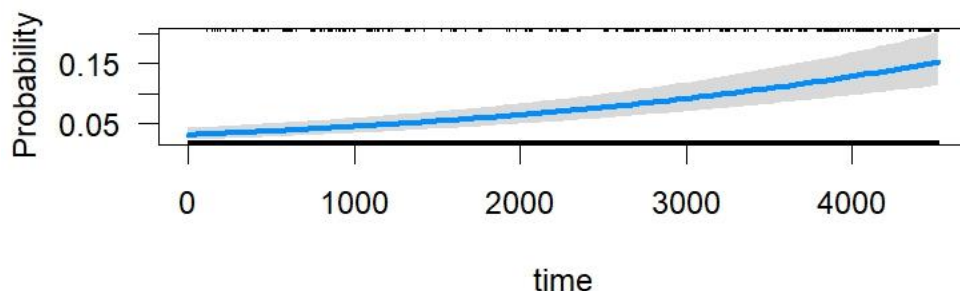
5.10. TRI_OS. This Gamm illustrates the co-occurrence trajectory of Therapeutic Refiguration and Insight (TRI) and Occupation/Study (OS) during the present psychotherapy.



5.11. TRI_PRJ. This Gamm illustrates the co-occurrence trajectory of Therapeutic Refiguration and Insight (TRI) and Projectuality (PRJ) during the present psychotherapy.



5.12. TRI_SR. This Gamm illustrates the co-occurrence trajectory of Therapeutic Refiguration and Insight (TRI) and Social Relations (SR) during the present psychotherapy.



5.13. TRI_WB. This Gamm illustrates the co-occurrence trajectory of Therapeutic Refiguration and Insight (TRI) and Ways of Being (WB) during the present psychotherapy.

Appendix 6

Several excerpts of the Python code, displayed in Visual Studio, are presented here across multiple figures to illustrate how topics are extracted from a completed psychotherapy consisting of 28 sessions divided into four equal blocks. These figures illustrate representative code snippets corresponding to different steps of the analysis. The code first outlines the preprocessing stages necessary to clean the textual dataset, followed by the execution of each model – namely, LDA, BERTopic, and Top2Vec. Each model is subsequently presented with different visualizations, which are reported in the Results section.

preprocessing

obtain processed conversation turns starting from conversational turns.

```
import tqdm
import nltk
nltk.download('stopwords')
from nltk.corpus import stopwords
from nltk.stem.snowball import SnowballStemmer
stemmer = SnowballStemmer("english")
sw = set(stopwords.words('english'))

def preprocess_text(text: str) -> str:
    tokens = nltk.word_tokenize(text)
    stemmed_tokens = [stemmer.stem(token) for token in tokens] # I'm ignoring it
    stemmed_tokens = tokens # override stemming!
    no_sw_tokens = [token for token in stemmed_tokens if token.lower() not in sw]
    no_sw_tokens = [token for token in no_sw_tokens if len(token) > 2]
    processed_text = " ".join(no_sw_tokens)
    processed_text = remove_additional_stuff(processed_text)
    return processed_text
```

6.1. Preprocessing phase. First, we define the stop words and other relevant functions. Uploaded the corpus of data. Then, we preprocess the data by defining and extracting sessions and conversational turns.

embed each group of sessions into a document-term matrix

A document-term matrix is a matrix that has one row per each document and one column for each word in its vocabulary. Each cell contains the number of occurrences of the word w in document d .

```
import numpy as np
group_to_docterm = {}
for group, sessions in p_group_to_ses.items():
    print("-"*30)
    num_docs = len(sessions)
    vocab = group_to_vocab[group]
    num_terms = len(vocab)
    docterm = np.zeros((num_docs, num_terms))
    sessions_tokens = group_to_sessions_tokens[group]
    print(f"the group {group} has {num_docs} sessions.")
    for doc_index in range(num_docs):
        session_tokens = sessions_tokens[doc_index]
        for term_index in range(num_terms):
            term = group_to_vocab[group][term_index]
            count = sum([token==term for token in session_tokens])
            docterm[doc_index, term_index] = count
    print(f"a hint of the docterm matrix for group {group} is:")
    docterm = docterm.astype(int)
    print(docterm[0:2, 0:10])
    group_to_docterm[group] = docterm
```

6.2. Document-Term Matrix. We define a DTM to proceed with the other operationalization.

```
import numpy as np
import lda
import lda.datasets
X = lda.datasets.load_reuters()
vocab = lda.datasets.load_reuters_vocab()
titles = lda.datasets.load_reuters_titles()

# initialize model for each group
group_to_model = {}
for group in group_to_docterm.keys():
    group_to_model[group] = lda.LDA(n_topics=5, n_iter=8000, random_state=1)

group_to_ordered_topics = {}
for group, docterm in group_to_docterm.items():
    # print("-----")
    # print(f"group: {group}")
    X = docterm
    titles = tuple([f"document {i} in group {group}"])
    vocab = group_to_vocab[group]
    model = group_to_model[group]
    model.fit(X) # model.fit_transform(X) is also available

    topic_word = model.topic_word_ # model.components_ also works
    n_top_words = 8
    index_to_topic_words = {}
    for i, topic_dist in enumerate(topic_word):
        topic_words = np.array(vocab)[np.argsort(topic_dist)[:-(n_top_words+1):-1]]
        index_to_topic_words[i] = topic_words
        # print('Topic {}: {}'.format(i, ' '.join(topic_words))) DO NOT CANCEL

    group_to_ordered_topics[group] = index_to_topic_words
```

6.3. Try out LDA. Part of the starting operations to use LDA on our dataset was reported here.

```

from top2vec import Top2Vec
top2vec_group_to_model = {}
for group, pcts in p_group_to_cts.items():
    model = Top2Vec(pcts)
    top2vec_group_to_model[group] = model

```

6.4. Try out Top2Vec. The first passage to extract topics with Top2Vec was reported here.

```

from sklearn.datasets import fetch_20newsgroups
from bertopic import BERTopic
from wordcloud import WordCloud
import matplotlib.pyplot as plt

def create_wordcloud(topic_model, topic):
    text = {word: value for word, value in topic_model.get_topic(topic)}
    wc = WordCloud(background_color="white", max_words=1000)
    wc.generate_from_frequencies(text)
    plt.imshow(wc, interpolation="bilinear")
    plt.axis("off")
    plt.show()

```

from IPython.core.interactiveshell import InteractiveShell ...

instantiate, fit, and transform the documents

```

from bertopic import BERTopic

BERTopic_group_to_topic_model = {}
for group, processed_conversational_turns in p_group_to_cts.items():
    topic_model = BERTopic(language="english",
                           calculate_probabilities=True,
                           verbose=True,
                           nr_topics=8)
    topics, probs = topic_model.fit_transform(processed_conversational_turns)
    BERTopic_group_to_topic_model[group] = topic_model

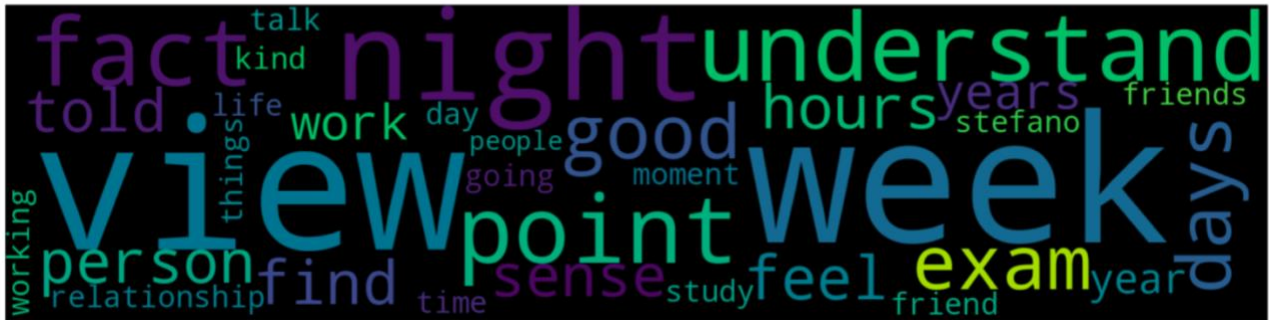
```

6.5. Try out BERTopic. Part of the starting operations to use LDA on our dataset was reported here.

Appendix 7

All the word clouds associated with each topic extracted from four different blocks, in which the 28-session psychotherapy was equally divided, are shown below. This word cloud was generated using the Top2vec model.

Topic 0



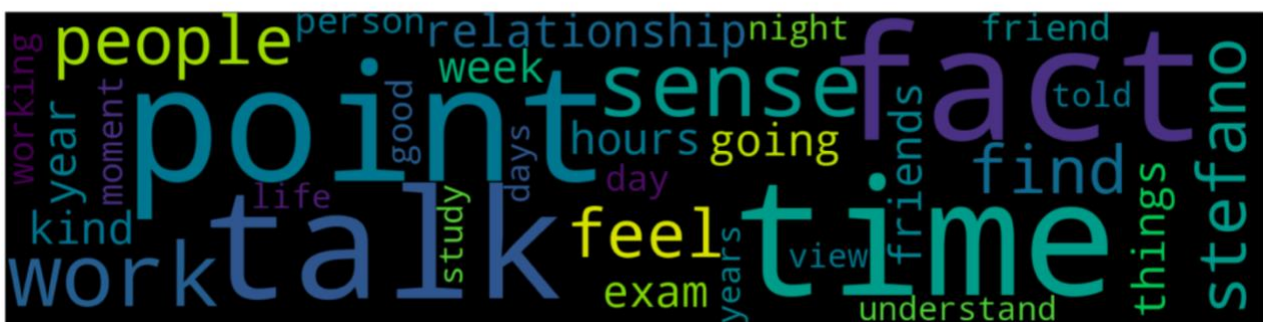
7.1. This is the word cloud of Topic 0 extracted from the first block of the psychotherapy.

Topic 1



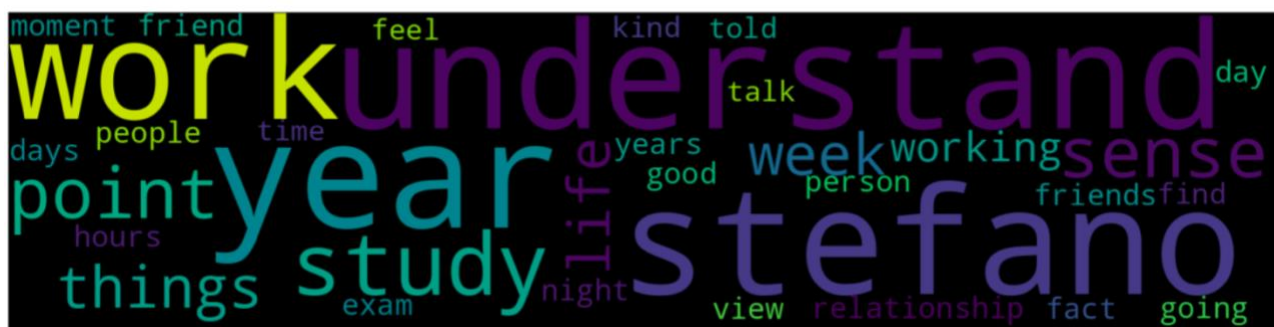
7.2. This is the word cloud of Topic 1 extracted from the first block of the psychotherapy.

Topic 2



7.3. This is the word cloud of Topic 2 extracted from the first block of the psychotherapy.

Topic 3



7.4. This is the word cloud of Topic 3 extracted from the first block of the psychotherapy.

Topic 4



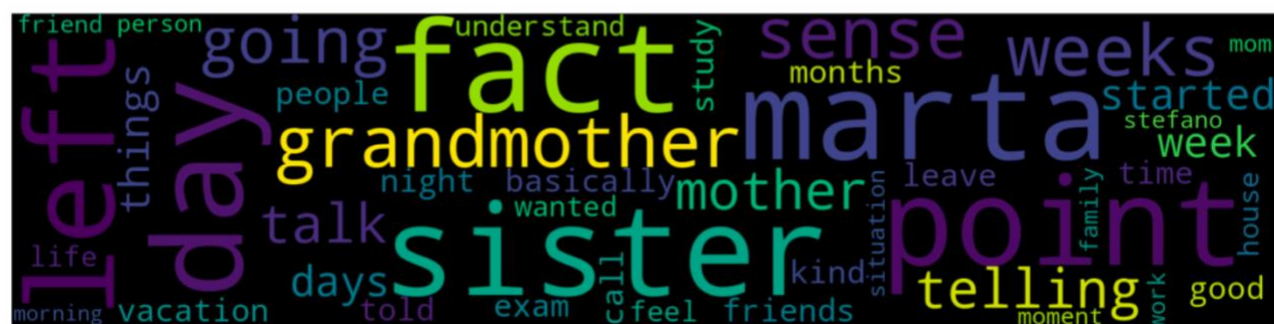
7.5. This is the word cloud of Topic 4 extracted from the first block of the psychotherapy.

Topic 0



7.6. This is the word cloud of Topic 0 extracted from the second block of the psychotherapy.

Topic 1



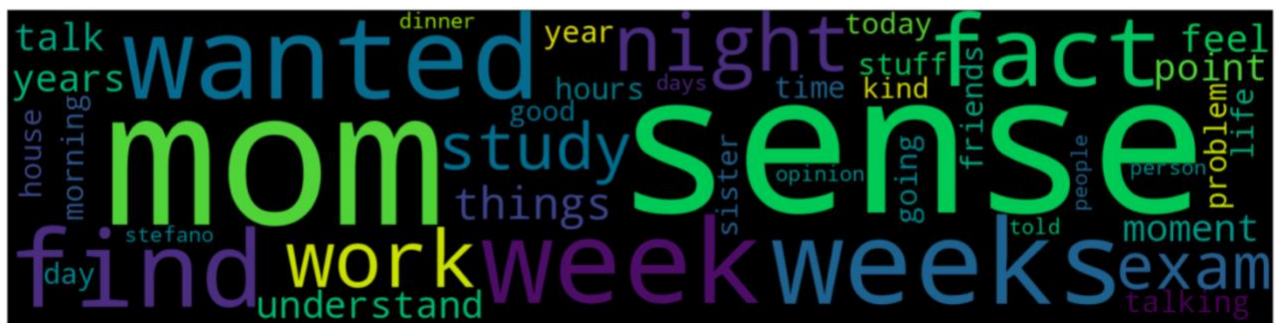
7.7. This is the word cloud of Topic 1 extracted from the second block of the psychotherapy.

Topic 2



7.8. This is the word cloud of Topic 2 extracted from the second block of the psychotherapy.

Topic 0



7.9. This is the word cloud of Topic 0 extracted from the third block of the psychotherapy.

Topic 1



7.10. This is the word cloud of Topic 1 extracted from the third block of the psychotherapy.

Topic 0



7.11. This is the word cloud of Topic 0 extracted from the fourth block of the psychotherapy.

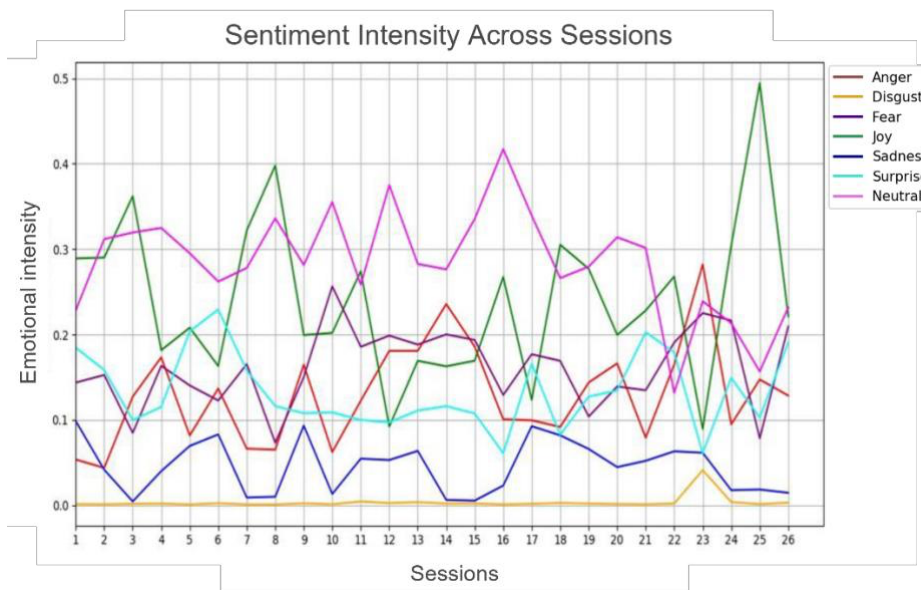
Topic 1



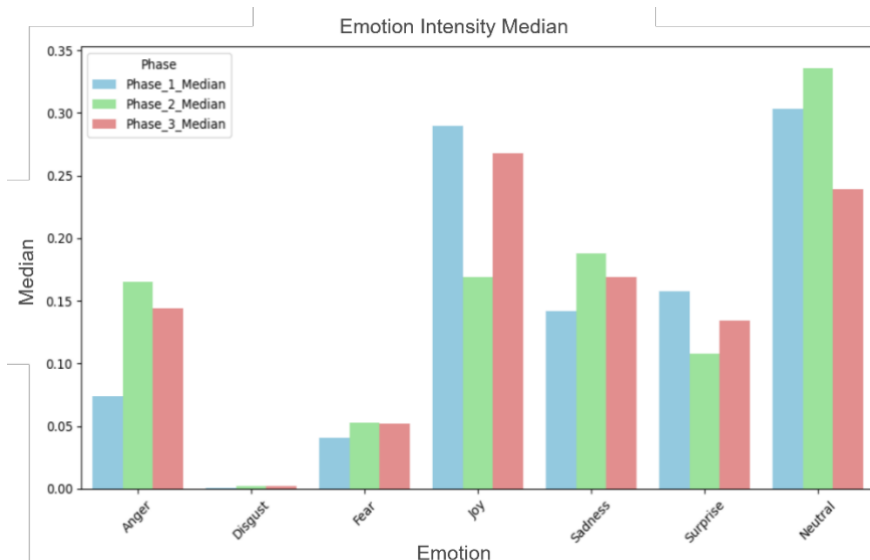
7.12. This is the word cloud of Topic 1 extracted from the fourth block of the psychotherapy.

Appendix 8

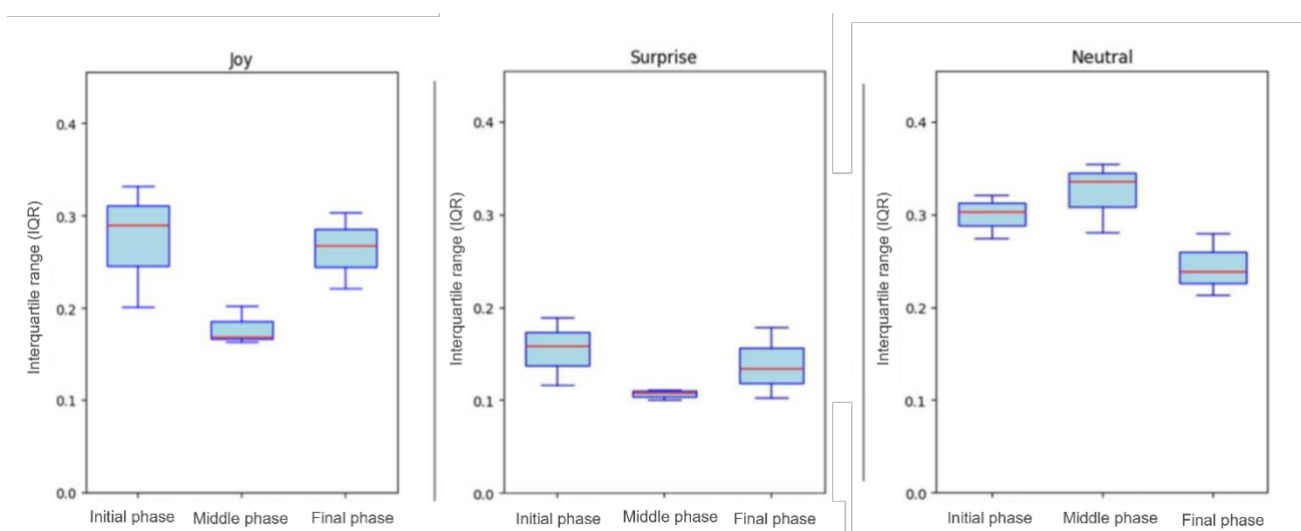
Supplementary visualizations of session-level emotion indices derived from psychotherapy transcripts (26 sessions) using a RoBERTa model fine-tuned on GoEmotions, with emotion labels aggregated into Ekman's six basic categories (anger, disgust, fear, joy, sadness, surprise) plus neutral.



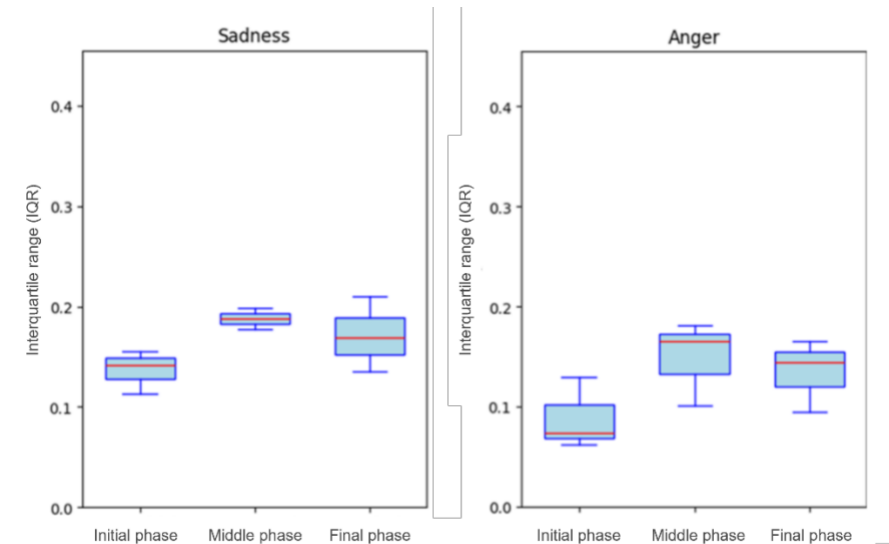
8.1. Emotion trajectories across sessions. Session-level trends for anger, disgust, fear, joy, sadness, surprise, and neutral across the 26 sessions. X-axis: session number (1–26). Y-axis: aggregated session-level emotion scores (0–1).



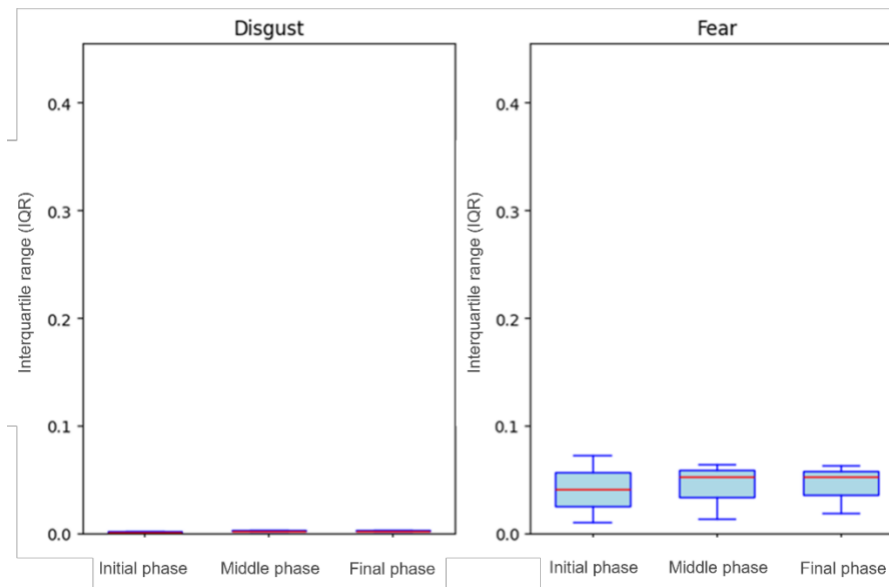
8.2. Median emotion scores across three treatment phases. Grouped bar chart showing the median session-level scores for each emotion category across the three phases of treatment (Phase 1, Phase 2, Phase 3). X-axis: emotion categories. Y-axis: median emotion score (aggregated at the session level). For each emotion, the three bars represent the median value computed within each phase, allowing comparison of phase-related shifts in emotional intensity across the therapeutic process.



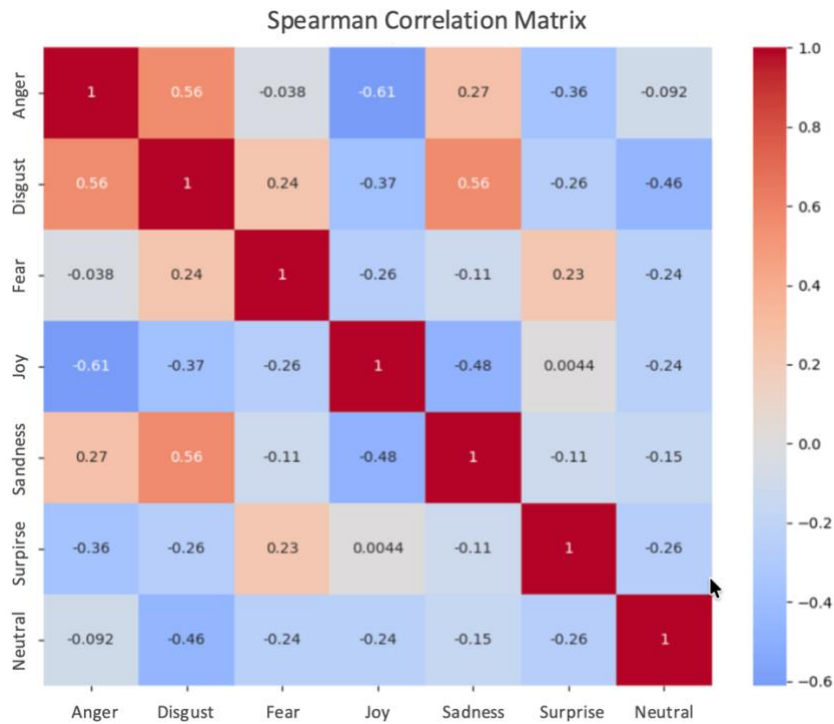
8.3. Phase-specific distributions of positive and neutral emotions. Three-panel boxplots showing the distribution of session-level emotion scores for Joy, Surprise, and Neutral across the three phases of treatment. X-axis: treatment phase (initial, middle, final). Y-axis: session-level emotion score (aggregated per session). In each panel, the central line indicates the median, boxes represent the interquartile range (IQR), and whiskers indicate values beyond the IQR, as specified by the plotting convention. This figure highlights phase-related differences in the dispersion and central tendency of these emotions across the therapeutic process.



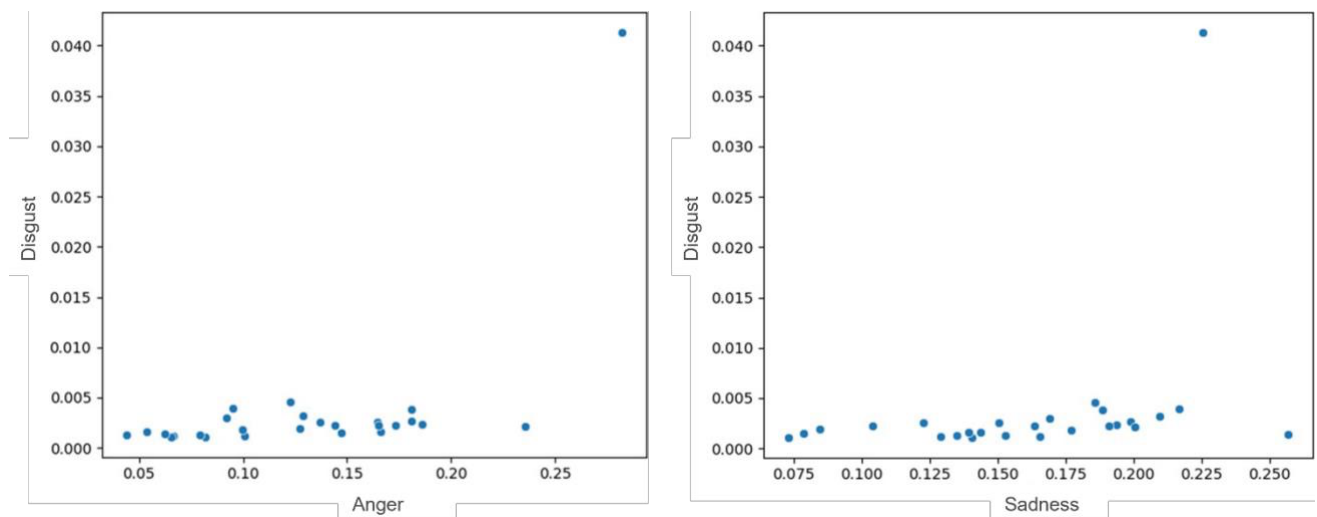
8.4. Phase-specific distributions of sadness and anger. Two-panel boxplots showing the distribution of session-level emotion scores for Sadness and Anger across the three phases of treatment. X-axis: treatment phase (initial, middle, final). Y-axis: session-level emotion score (aggregated per session). In each panel, the central line indicates the median, boxes represent the interquartile range (IQR), and whiskers indicate values beyond the IQR, as specified by the plotting convention. The figure allows comparison of phase-related shifts in the central tendency and dispersion of these negative emotions across the therapeutic process.



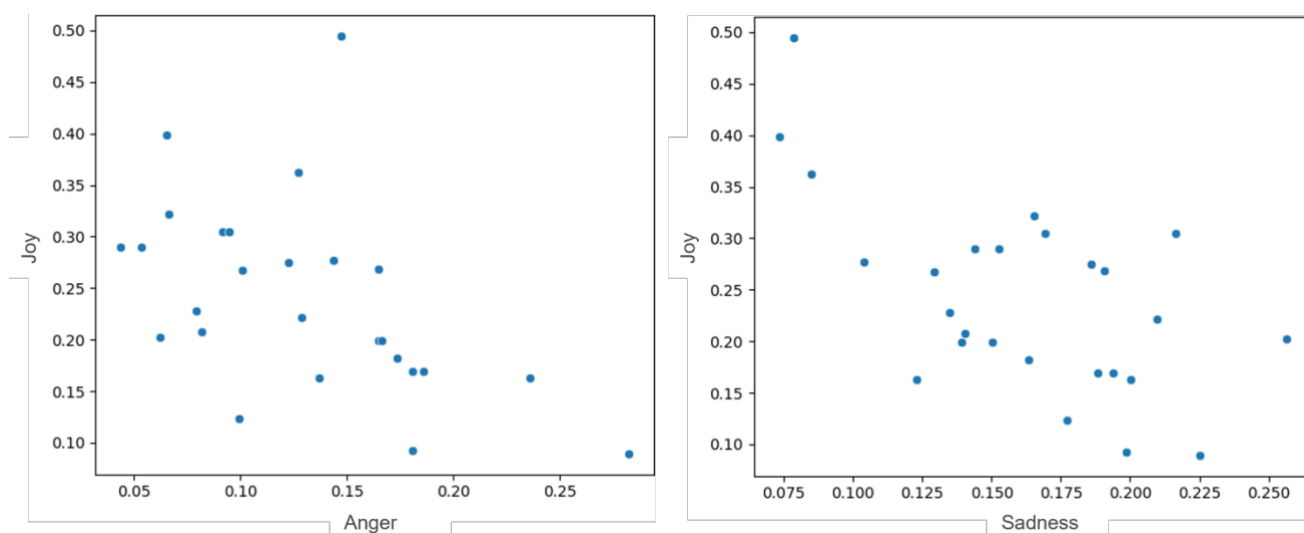
8.5. Phase-specific distributions of disgust and fear. Two-panel boxplots showing the distribution of session-level emotion scores for Disgust and Fear across the three phases of treatment. X-axis: treatment phase (initial, middle, final). Y-axis: session-level emotion score (aggregated per session). In each panel, the central line indicates the median, boxes represent the interquartile range (IQR), and whiskers indicate values beyond the IQR, as specified by the plotting convention. This figure shows consistently low disgust levels across phases and a relatively stable distribution of fear throughout treatment.



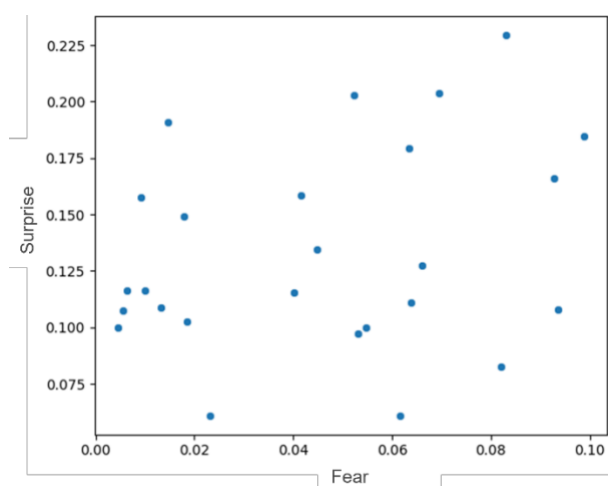
8.6. Spearman correlation matrix among emotion scores across sessions. Heatmap summarizing pairwise Spearman's ρ correlations among the seven session-level emotion scores (anger, disgust, fear, joy, sadness, surprise, neutral) across the therapeutic process. X-axis and Y-axis: emotion categories. Each cell reports the corresponding correlation coefficient, while the colour scale indicates the direction and magnitude of association (warm colours = positive; cool colours = negative). This matrix provides an overview of the co-variation structure among negative emotions and their inverse relationships with joy.



8.7. Associations of disgust with anger and sadness across sessions. Two-panel scatterplots showing session-level relationships involving disgust. Left panel: anger (X-axis) vs. disgust (Y-axis). Right panel: sadness (X-axis) vs. disgust (Y-axis). Each point represents one session, with emotion values aggregated at the session level. The plots visualize the positive co-variation between disgust and anger and between disgust and sadness observed in the correlational analyses.



8.8. Associations of joy with anger and sadness across sessions. Two-panel scatterplots showing session-level relationships involving joy. Left panel: anger (X-axis) vs. joy (Y-axis). Right panel: sadness (X-axis) vs. joy (Y-axis). Each point represents one session, with emotion values aggregated at the session level. The plots visualize the inverse relationships between joy and anger and between joy and sadness observed in the correlational analyses.



8.9. Association between fear and surprise across sessions. Scatterplot showing the session-level relationship between fear and surprise scores across the therapeutic process. X-axis: fear (session-level aggregated score, 0–1). Y-axis: surprise (session-level aggregated score, 0–1). Each point represents one session. The plot visualizes the positive co-variation between fear and surprise observed in the correlational analyses.

Appendix 9

In this appendix, a selection of partial prompts (i.e., excerpts) of the AI-generated outputs is reported. The prompts presented here are intentionally partial and provided for documentation and transparency purposes. They are periodically reviewed and continuously refined to improve clarity, robustness, and overall performance of the AI-assisted analysis workflow.

```

1  italian:
2      speaker_role_detection:
3          prompt: |
4              Adesso leggi il seguente trascritto diarizzato e fa quanto segue:
5              1. assegna a ciascun turno conversazionale anzichè Speaker A e Speaker B il label "paziente" e "terapeuta". Devi capire tu
6              2. correggi le parole e le frasi che presentano errori grammaticali o che non hanno un senso compiuto, in modo da ottenere
7
8              Tieni conto delle seguenti indicazioni:

```

9.1. Partial excerpt of the prompt used to generate the verbatim transcription output for an individual uploaded therapeutic session (via the AssemblyAI API). The prompt is routinely revised and updated to optimise transcription quality and downstream analytical outputs.

```

italian:
  summary: |
    Obiettivo:
    Dato il trascritto di un colloquio terapeutico (o un riassunto parziale), fornisci un riassunto conciso dei punti principali.

    Raccomandazioni e indicazioni:
    1. Il numero di persone coinvolte nel colloquio {num_speakers}
    2. Il paziente si chiama {patient_name}
    3. Il colloquio si è svolto in data {date}
    4. Riferisciti al terapeuta col termine "terapeuta"
    5. Concentrati sui temi principali, gli stati d'animo e le interazioni chiave tra paziente e terapeuta.
    6. Mantieni un tono neutro e professionale, evitando un linguaggio troppo forte e giudizi personali.
    7. Non utilizzare formattazione HTML in questo riassunto.
    8. Per le liste puntate fornisci un massimo di 5 elementi ed evita di ripetere più volte concetti molto simili.
    9. Includi dettagli rilevanti come età, occupazione o situazione familiare del paziente, se menzionati.
    10. Evidenzia eventuali progressi o intuizioni significative emerse durante il colloquio.
    11. Riporta eventuali tecniche terapeutiche o esercizi suggeriti dal terapeuta.

    Di seguito il testo da riassumere:

  summary_with_history: |
    Obiettivo:
    Dato il trascritto di un colloquio terapeutico (o un riassunto parziale) e un riassunto dei colloqui precedenti, fornisci un riassunto conciso dei punti principali.

    Raccomandazioni e indicazioni:
    1. Il numero di persone coinvolte nel colloquio {num_speakers}
    2. Il paziente si chiama {patient_name}
    3. Il colloquio si è svolto in data {date}
    4. Riferisciti al terapeuta col termine "terapeuta"
    5. Concentrati sui temi principali, gli stati d'animo e le interazioni chiave tra paziente e terapeuta.
    6. Mantieni un tono neutro e professionale, evitando un linguaggio troppo forte e giudizi personali.
    7. Non utilizzare formattazione HTML in questo riassunto.

```

9.2. Partial excerpt of the prompt used to generate the Session Summary output for an individual uploaded therapeutic session. The prompt is routinely revised and updated to optimise the quality and consistency of the generated summaries.

```

italian:
  prompt: |
    Basandoti sul seguente riassunto di un colloquio terapeutico, genera una wordcloud.
    Identifica almeno 20 delle parole più importanti che comunicano i punti principali del colloquio.
    Evita inoltre di inserire parole troppo generali come 'terapia', 'colloquio', 'paziente', 'terapeuta', ecc.
    Per ogni parola, assegna un punteggio compreso tra 0.0 e 1.0, dove i punteggi più alti indicano una maggiore importanza.
    Fornisci l'output come una lista JSON di dizionari, dove ogni dizionario ha le chiavi 'word' e 'score'.
    Esempio di formato: [{"word": "terapia", "score": 0.9}, {"word": "ansia", "score": 0.8}]

```

9.3. Partial excerpt of the prompt used to generate the Word Cloud output for an individual uploaded therapeutic session (i.e., the extraction and scoring of the most salient terms). The prompt is routinely revised and updated to optimise the quality and interpretability of the generated outputs.

```
italian:
system_instruction:
- Sei un assistente esperto nell'analisi di trascrizioni di sedute terapeutiche.
- Il tuo compito è categorizzare ogni turno conversazionale secondo le categorie fornite.
- Devi sempre assegnare esattamente 3 etichette per ogni turno conversazionale.
- Usa le etichette solo quando sei sicuro di esse. Se non sei sicuro che un tema sia rilevante, non includerlo. Invece, usa 'X'.
- Se un turno è breve o non significativo, usa tre etichette 'X'.
- Usa l'etichetta 'RCT' ogni volta che il terapeuta parla della terapia stessa o offre riformulazioni terapeutiche.
- È fondamentale utilizzare l'etichetta 'SES' per temi legati alla sessualità.
- Non devi mai inventare nuove categorie o usare etichette non presenti nella lista fornita.
- L'output deve essere in formato JSON secondo lo schema fornito.

prompt: |
Ti verrà fornito il trascritto di un colloquio tra un terapeuta e un paziente.
Dovrai fornirmi, per ciascuna riga (turno conversazionale), i temi a esso più correlati.

L'insieme dei temi disponibili è il seguente:

SPS: (Sofferenza, Patologia, o Sintomi) Questa categoria raccoglie tutte le informazioni che riguardano il malessere emotivo, p
FAM: (Famiglia) Questa categoria riguarda tutte le informazioni e i significati relativi alla famiglia del paziente, comprese le
```

9.4. Partial excerpt of the prompt used to generate the Nuclear Topics output for an individual uploaded therapeutic session, including the classification of each conversational turn into predefined macro-topic categories and the production of structured (JSON) topic annotations. The prompt is routinely revised and updated to optimise the quality, consistency, and interpretability of the generated outputs.

```
prompt: |
Compito: Analizza il seguente riassunto di un colloquio terapeutico e identifica i 4 argomenti principali con 4 espressioni

Formato di output richiesto (JSON):
[
  {"name": "Nome Argomento 1", "words": ["espressione1", "espressione2", "espressione3", "espressione4"]},
  {"name": "Nome Argomento 2", "words": ["espressione1", "espressione2", "espressione3", "espressione4"]},
  {"name": "Nome Argomento 3", "words": ["espressione1", "espressione2", "espressione3", "espressione4"]},
  {"name": "Nome Argomento 4", "words": ["espressione1", "espressione2", "espressione3", "espressione4"]}
]

Istruzioni dettagliate:
1. Identifica un massimo di 4 argomenti principali.
2. Ogni argomento deve avere al massimo 4 espressioni associate. L'importante è la qualità e non la quantità.
3. Le espressioni devono essere brevi ma significative per l'argomento.
4. Ordina le espressioni per rilevanza semantica, non per frequenza nel testo.
5. Usa nomi concisi e descrittivi per gli argomenti.
6. Non utilizzare parole del nome dell'argomento nelle espressioni associate.
7. Evita di inserire parole troppo generali come 'terapia', 'colloquio', 'paziente', 'terapeuta', ecc.
8. Fornisci solo il JSON richiesto, senza testo aggiuntivo.
```

9.5. Partial excerpt of the prompt used to generate the AI-Derived Topics output for an individual uploaded therapeutic session, including the identification of up to four session topics and their associated key expressions in a structured (JSON) format. The prompt is routinely revised and updated to optimise the quality, consistency, and clinical interpretability of the generated outputs.

```

prompt: |
Analizza attentamente il seguente testo contenente la concatenazione di riassunti terapeutici di un paziente. Crea un documento

a. Informazioni generali
b. Sintomi/sofferenza
c. Storia clinica
d. Progressi e piano terapeutico
e. Eventi e date rilevanti
f. Aree di Interesse Diagnostico

Linee guida importanti:
1. La sezione "Informazioni Generali" include:
  - Dettagli demografici: età, occupazione, provenienza geografica
  - Ambiente familiare e sociale
  - Contesto culturale e socio-economico
  - Fattori esterni che influenzano lo stile di vita del paziente
  Nota: Questa sezione fornisce un quadro generale della vita del paziente, non direttamente collegato a tematiche emotive o psichiche.
2. Nella sezione "Eventi e date rilevanti", riporta le date di eventi significativi se presenti, ma non riportare mai le date in assenza di informazioni.
3. Per sezioni senza informazioni, includile comunque specificando chiaramente l'assenza di dati.
4. Per le liste puntate fornisci un massimo di 5 elementi ed evita di ripetere più volte concetti molto simili.
5. Nella sezione "Aree di Interesse Diagnostico", riporta solo informazioni presenti nel testo. In assenza di dati, indica che non sono presenti.

```

9.6. Partial excerpt of the prompt used to generate the Dynamic Clinical Sheet overall output, which aggregates and organises information extracted from multiple uploaded sessions into predefined clinical sections (e.g., general information, symptoms/distress, clinical history, progress/treatment plan, relevant events/dates, and diagnostic areas of interest). The prompt is routinely revised and updated to optimise the quality, completeness, and clinical interpretability of the generated outputs.

```

summary: |
Data la seguente concatenazione di riassunti delle sedute terapeutiche del paziente, genera un riassunto completo e discorsivo.

1. Il riassunto deve essere in prosa, senza capitoli o sezioni formali.
2. Usa paragrafi semplici, separati da una linea vuota, per organizzare il contenuto.
3. Non utilizzare titoli o sottotitoli.
4. Copri tutti gli aspetti importanti della storia del paziente in modo fluido e coerente.
5. Non inventare informazioni o dettagli che non sono presenti nel testo originale.
6. evita di utilizzare un linguaggio troppo forte, devi evitare di dare giudizi personali.
7. Per le liste puntate fornisci un massimo di 5 elementi ed evita di ripetere più volte concetti molto simili.
8. Concludi con una breve valutazione complessiva e eventuali raccomandazioni.
9. Il testo deve essere dettagliato ma scorrevole, senza formattazione speciale o elenchi.

Genera il riassunto basandoti su queste linee guida e sul seguente contenuto delle sedute:

html: |
Dato il seguente riassunto testuale del paziente, arricchiscilo con tag HTML per migliorarne la leggibilità e la struttura.

1. Mantieni il flusso del testo originale, aggiungendo solo formattazione HTML.
2. Usa il tag <strong> per evidenziare nomi, date, diagnosi o altre informazioni chiave.
3. Crea elenchi puntati (<ul>) o numerati (<ol>) quando appropriato, ad esempio per liste di sintomi o trattamenti.
4. Utilizza paragrafi <p> per separare le sezioni logiche del testo.
5. Non aggiungere titoli o sottotitoli con tag come <h1>, <h2>, ecc.
6. Inserisci tag <br> per creare interruzioni di riga dove necessario per migliorare la leggibilità.
7. Assicurati che l'HTML sia valido e ben formattato.

Esempio di come potrebbe apparire una parte del testo formattato:

```

9.7. Partial excerpt of the prompt used to generate the Overall Narrative Summary overall output, producing a discursive, continuously updated synthesis of the patient's therapeutic trajectory from the cumulative content of multiple uploaded sessions (including an optional HTML-formatted version for improved readability). The prompt is routinely revised and updated to optimise the quality, coherence, and clinical interpretability of the generated outputs.