Current Psychology (2026) 45:1110
https://doi.org/10.1007/s12144-026-09685-3

Using LLMs as a clinician support for detecting defense mechanisms from psychotherapy transcripts: a proof-of-concept study

Lorenzo Antichi[1] · **Elena Sajno**[1,2] · **Chiara Rossi**[3] · **Fabio Frisone**[1,4] · **Francesca De Salve**[5] · **Osmano Oasi**[5] · **Giuseppe Riva**[1,6]

Received: 13 November 2025 / Accepted: 18 June 2026

Abstract

This proof-of-concept study evaluated the feasibility of using large language models (LLMs) to support clinicians in identifying defense mechanisms in psychodynamic psychotherapy transcripts, comparing LLM assessments with those of psychodynamic therapists in both solo and supervision group settings. The first two sessions of a psychoanalytic case were analyzed. The transcribed sessions were segmented into excerpts and coded by two psychodynamic clinicians and two LLMs (ChatGPT and Gemini) guided with a structured prompt. Coding followed an adapted version of DMRS comprising 12 defenses organized into three maturity levels. For each excerpt, defense mechanisms were identified, and an overall defensive functioning index (ODF) was calculated. Inter-rater agreement was estimated using the Intraclass Correlation Coefficient (ICC), considering human-human, LLM-LLM, human-LLM, and integrated versions (clinical supervision vs. LLM integration) pairs. LLMs tended to assign a higher number of defenses, including more immature ones, producing profiles with lower average ODFs than clinicians, who instead provided more selective and, overall, more mature ratings. However, some of the defenses reported only by LLMs were subsequently recognized by supervised clinicians as clinically relevant and worthy of inclusion. Clinicians showed good to excellent inter-rater reliability, while ChatGPT and Gemini showed low inter-rater reliability, suggesting that the two models are not interchangeable. Agreement between clinicians and LLMs was low across the sessions. Therefore, the combined use of LLMs among clinicians is technically feasible and could potentially enhance understanding of patients' defensive functioning, especially as a support for case formulation and supervision. Currently, LLMs cannot replace clinical judgment or the therapeutic relationship, but they might be considered as a support tool for clinical practice under human professional guidance.

Keywords Psychodynamic psychotherapy · Defense mechanisms · Large language models · ChatGPT · Gemini

Introduction

Defense mechanisms are mostly unconscious psychological processes that help individuals manage internal conflicts or stressful situations by regulating affects, impulses, and self-representations (APA, 2013; Cramer, 1998). Originally conceptualized within psychoanalytic theory (Freud, 1937, 1946), defenses have since been reconceptualized across different perspectives, including relational and attachment-based models (e.g., Bowlby, 1980; Fonagy et al., 1991; Mitchell, 1988). Despite these theoretical differences, most contemporary approaches conceptualize defenses as mental operations that protect the individual from excessive psychological distress while maintaining self-coherence (Perry & Cooper, 1986). Notably, defense mechanisms are not pathognomonic for any single diagnostic condition,

✉ Lorenzo Antichi
lorenzo.antichi@hotmail.it

1 Humane Technology Lab, Catholic University of Sacred Heart, Via Fra Agostino Gemelli 1, Milan 20123, Italy

2 Department of Computer Science, Università degli Studi di Milano, Milan, Italy

3 Department of Human Sciences, Guglielmo Marconi University, Rome, Italy

4 Departments of Economics, Psychology, Communication, Education, and Motor Sciences, Niccolò Cusano University, Rome, Italy

5 Department of Psychology, Catholic University of Sacred Heart, Milan, Italy

6 Applied Technology for Neuro-Psychology Laboratory, IRCCS Istituto Auxologico Italiano, Milan, Italy

Published online: 25 June 2026

as the same mechanisms can occur across a wide range of conditions and at varying levels of severity (Lingiardi & Madeddu, 2023). Clinically, what renders a defense problematic is its rigidity and pervasiveness rather than the defense itself; indeed, the same defense may be observed across different conditions and developmental levels (PDM-2; Lingiardi & McWilliams, 2017).

Therapists have numerous advantages in assessing defense mechanisms during therapy, including identifying areas of conflict requiring intervention, providing a diagnostic framework for patients, monitoring therapy outcomes, anticipating future resistance, and tailoring interventions (Di Giuseppe et al., 2021; Perry & Bond, 2012). For example, patients with neurotic-level functioning may benefit from interpretations that increase insight into defensive patterns, whereas patients with personality disorders may experience such interpretations as intrusive or threatening (Vaillant, 1998). Defense mechanisms are also frequently discussed in clinical supervision, since therapists can better understand patients' psychological functioning and reflect on therapeutic strategies and countertransference reactions (Andersson, 2008).

However, detecting defenses in psychotherapy transcripts is often complex and time-consuming, requiring specialized training and careful interpretation of subtle linguistic and affective cues. Recent advances in natural language processing (NLP) and large language models (LLMs) raise the possibility that automated tools may assist clinicians in identifying defensive processes within psychotherapy material. Therefore, the present proof-of-concept study explores whether LLMs can support clinicians in detecting defense mechanisms in psychodynamic psychotherapy transcripts.

Defense mechanisms measurement

There are many methodological approaches to assessing defense mechanisms during psychotherapy sessions. A first option is self-report instruments, which are easy to use and require no specific training, such as the Defense Style Questionnaire (DSQ) (Andrews et al., 1993), the Defense Mechanisms Rating Scale (DMRS-SR-30) (Di Giuseppe et al., 2020), and the Defense Mechanisms Inventory (DMI) (Gleser & Ihilevich, 1969). However, self-reports may be subject to social desirability bias and may fail to distinguish between symptoms, defenses, or coping (Cramer, 1998; Paulhus, 1991). Furthermore, the defense is mostly unconscious, while self-reporters might evaluate its conscious derivative rather than the mechanism itself (APA, 2013; Andrews et al., 1993; San Martini et al., 2004).

Another possibility is to rely on projective tests, such as the Rorschach defensive indices (Schafer, 1954) or the Defense Mechanism Manual (Cramer, 1991, based on the Thematic Apperception Test, Murray, 1943). The principal advantage is reduced evaluator influence, as they must follow a manual to code the defenses. Nevertheless, projective tests require time and cost and do not guarantee that the defenses identified in a single-session administration correspond to those the patient usually uses in real life (e.g., Crasovan, 2014).

Finally, approaches based on clinical observation consist of analyzing a recorded and transcribed clinical interview. Examples include the Defense Mechanism Rating Scale (DMRS; Perry, 1990) or the Structured Interview of Personality Organization (STIPO; Stern et al., 2010). Observational methods that analyze psychotherapy transcripts have been widely used in defense research. For example, the DMRS has been applied to recorded psychotherapy sessions and clinical interviews to examine defensive functioning, its changes over the course of treatment, and its association with therapeutic outcomes (e.g., Perry & Bond, 2012; Babl et al., 2019). Nevertheless, observation methods require specialized training, lengthy coding, and resources to locate the material. One solution could be Q-sort versions, such as the DMRS-Q-sort (Di Giuseppe et al., 2014), in which the rater orders a list of items from least to most representative following a forced distribution. The DMRS-Q-sort administration takes only a short time and requires no specific training. However, the Q-sort approach requires the clinician to follow the scale ranks and provide an overall summary judgment of the session, rather than evaluating each defense mechanism at specific moments or excerpts of the interview. This kind of finer-grained, excerpt-level analysis is instead the typical focus of case discussion in clinical supervision.

The use of artificial intelligence for detecting defenses

Recently, automated methods have been developed to assist clinicians in assessment, such as classifying therapist utterances or identifying disorders using machine learning algorithms (Ewbank et al., 2021; Hawa et al., 2020). Specifically for defense mechanisms, Tasca et al. (2023) used the RoBERTa algorithm to distinguish four defenses from the DMRS (i.e., repression, reaction formation, suppression, and intellectualization) and to indicate the presence or absence of all others in transcripts of the Adult Attachment Interview (AAI; George et al., 1985) previously coded by human therapists. The models were

capable of distinguishing defenses, but not enough to replace human raters, as a quarter to a fifth of the defenses were undetected. Increasing the threshold for true positives increased the rate of truly present defenses identified, but also the false-positive rate (e.g., 91% true positives and 60% false positives for repression, misclassifying over half of the defenses).

Although the study by Tasca et al. (2023) offers promising evidence for the use of automated methods to support defense assessment, machine-learning-based tools may require technical expertise that is not always accessible to clinicians, unless a graphical user interface (GUI) is used to facilitate interaction. The advent LLMs increases the capabilities of NLP systems (such as RoBERTa) in two ways: first, models such as ChatGPT-3 have already proven to be equally or more performant than more traditional NLP systems (including BERT++) (Brown et al., 2020); second, by enabling the shift from programming to prompting, they make automated text analysis accessible to non-expert users, which previously needed to be implemented through code pipelines, and increase the flexibility that can be obtained from texts (Liu et al., 2023). However, defense detection based on platforms that leverage LLMs' potential and enable simplified user interaction, such as ChatGPT or Gemini, has never been tested.

In recent years, various field applications of ChatGPT and Gemini have emerged for clinical practice. Case studies in psychotherapy show that ChatGPT can serve as a complementary tool for therapists, providing a "second look" at clinical material. At the same time, some patients use it between sessions or during the therapist's vacation (Raile, 2024). On the evaluation side, a cross-sectional study pre-trained ChatGPT-4 on validated scales and found strong agreement between scores generated by the structured interview with ChatGPT and those from standard questionnaires, suggesting its potential as a complementary assessment tool for loneliness and online social support (Gu et al., 2025). In training and case conceptualization, a feasibility trial found that ChatGPT can help trainees practice case formulation skills, save time, and improve the organization of clinical reasoning (Hsieh et al., 2024). The literature on Gemini is still emerging, but shows that Gemini can be a training aid for assessing clinical knowledge (Neubauer et al., 2025), for producing reasonable differential diagnosis lists and readable informational texts, with results often close to, but not superior to, those of physicians (Hirosawa et al., 2023; Fattah et al., 2025; Wang et al., 2025). Based on this promising evidence, we wonder whether ChatGPT and Gemini could be used to support individual clinicians or as comparison tools in supervision to detect defenses in psychodynamic psychotherapy transcripts.

Aim

To the best of our knowledge, no studies have yet examined whether general-purpose LLM interfaces such as ChatGPT and Gemini can be used through structured prompting to support clinicians in identifying defense mechanisms within psychotherapy session transcripts. Accordingly, this proof-of-concept study aimed to evaluate the feasibility of a prompting pipeline using psychodynamic psychotherapy material from a single case. Specifically, we examined whether: (1) LLMs could generate explicit defense labels from transcript segments and provide structured rationales for choices; (2) LLM outputs could be compared with clinician ratings by examining their inter-rater agreement in solo or in a supervision setting.

Tasca et al. (2023) previously applied a supervised machine-learning classifier to detect defenses in AAI transcripts, demonstrating the potential of automated approaches. However, the present study focuses on general-purpose LLM interfaces as prompting-based tools that can be used directly in everyday clinical and supervisory workflows without programming expertise, thereby shifting the emphasis from algorithm development to clinician usability.

Given the exploratory nature of the study, we did not formulate confirmatory hypotheses regarding accuracy or generalizability. Instead, the aims would be considered supported if LLMs consistently produced identifiable defense labels, if their outputs allowed comparison with clinician judgments within the workflow, and if the generated material could be integrated into supervision discussions without replacing clinical decision-making. Our idea is that LLMs can provide insights to enhance clinical reasoning in solo or supervised settings by considering defense mechanisms that human raters may not have considered, while stressing that LLMs cannot replace clinical judgment or the therapeutic relationship (e.g., Wang et al., 2025).

Method

Study design

This was a proof-of-concept, single-case feasibility study using transcripts of psychodynamic psychotherapy. We compared defense identification across two general-purpose LLM interfaces (ChatGPT and Gemini) and psychodynamic clinicians (individual ratings and supervision-based consensus). LLMs were neither trained nor fine-tuned; instead, they were guided by a standardized prompting protocol to generate structured defense annotations. Human raters and LLMs were blinded to each other's outputs. Quantitative comparisons focused on inter-rater agreement.

Case study

The single case was about Annie, a woman in her early thirties at the time of the beginning of psychoanalytic treatment (1982) in the United States, which lasted approximately three and a half years for a total of 324 sessions. She was a married housewife with two children, engaged in a relational network limited almost exclusively to the nuclear and extended family (e.g., parents, in-laws, husband). She experienced conflict and dissatisfaction in her most significant relationships, particularly with her husband, whom she perceived as devaluing and blaming her. The patient entered therapy for agoraphobia with significant psychosomatic manifestations (e.g., nausea, vomiting, diarrhea), which interfered with daily functioning and social life. She described herself as inadequate in her roles as wife, mother, and daughter, and she experienced recurrent guilt and self-depreciation. The transcripts of the first two psychotherapy sessions were used for this study. The material is available for research purposes through the Psychoanalytic Research Consortium (for further information, please visit https://psychoanalyticresearch.org/case-studies/).

Instruments

We simplified and adapted a version of the DMRS (see Lingiardi & Madeddu, 2023, for the Italian version; see Perry, 1990, for the original version), which is considered the observational gold standard for assessing defense mechanisms in psychodynamic and psychoanalytic psychotherapy. DMRS is a coding system for identifying defense mechanisms in clinical session transcripts, describing their psychological functions, and quantifying their distribution along a hierarchy of defensive maturity. The DMRS manual describes 30 individual defense mechanisms. It organizes defenses into a hierarchy of seven levels of increasing adaptability, from the most mature to the least mature: (7) highly adaptive defenses (also known as mature defenses), such as self-observation, humor, altruism, and sublimation; (6) obsessive defenses, such as intellectualization, isolation of affect, and undoing; (5) neurotic defenses, including reaction formation, displacement, and repression; (4) minor image-distorting defenses, including devaluation and idealization; (3) disavowal defenses, such as rationalization, denial, and projection; (2) major image-distorting defenses, characterized by splitting and projective identification; and (1) action defenses, such as impulsive acting out or passive-aggression.

Quantitatively, the DMRS allows for three main types of scoring: (a) the proportional frequency of each specific defense mechanism (i.e., how often a given defense appears relative to the total defenses observed in the session); (b) the proportional frequency of defense levels (e.g., percentage of mature defenses vs. percentage of immature defenses); and (c) Overall Defensive Functioning (ODF), a continuous summary index representing the average level of a person's defensive maturity. Higher ODF values reflect more adaptive defensive functioning, while lower values reflect more immature functioning. Moreover, DMRS can be used to qualitatively code defenses with a confidence index indicating the diagnostic certainty within the extract (i.e., 0 = weak/contradictory evidence; 1 = probable but not unequivocal; 2 = clear and specific) following specific criteria (e.g., frequency of reported acting-out episodes, presence of unjustified persecutory attributions for projection, cold cognitive description of emotionally charged events for isolation of affect, etc.).

For this proof of concept, the entire DMRS system was partially scaled down to make it usable for the LLM task. Therefore, we used only 12 defenses grouped into three clinically significant and well-described macro-clusters in the literature (Vaillant, 1998): mature defenses (e.g., humor, sublimation, anticipation, affiliation), neurotic defenses (e.g., displacement, repression, isolation of affect), and immature defenses (e.g., denial, projection, acting out, splitting, and rationalization). This methodological compromise had two purposes: (1) we evaluated that the 12 defenses selected are those that an LLM can most reliably recognize by reading only the session transcript, focusing on clear verbal markers, without the need to access to the therapist's emotional tone, the here-and-now relational exchange, or the countertransference, as other defenses require (e.g., projective identification, identification with the aggressor, more subtle forms of passive-aggression, or psychotic distortions); (2) we aimed to shorten and reduce the content of the input to be given to the LLM. The second point is critical because LLMs do not process text as full sentences, but as tokens (small text units, such as words or word fragments). Thus, prompt formulation (the way a prompt is written) affects how many tokens are needed and, thus, the total input length the model must process. This matters because many LLMs use self-attention, a mechanism that evaluates how each token relates to every other token in the sequence, so computational and memory costs increase rapidly as input length grows (following an approximate quadratic trend) (Tay et al., 2023). Additionally, LLMs pay less attention to information in the middle of long inputs, and prompt compression methods, which make prompts denser with relevant information, can maintain or even improve performance on shorter texts (Jiang et al., 2023; Liu et al., 2024).

Consequently, while in the original DMRS the score of a defense is obtained by multiplying its frequency by a weight corresponding to one of the seven maturity levels (for example: level 7 defenses receive a weight of 7, level 4 defenses a

weight of 4), in our simplified model – in which we reduced the levels from seven to three clusters – the frequency of defenses was multiplied by 1 for the immature cluster, 2 for the neurotic cluster and 3 for the mature cluster[1]. Therefore, the defense count (absolute frequency of a defense in a session), the sum of every defense count (F), the defense score (i.e., defense count x corresponding weight), the sum of the scores between the defenses (W), and ODF (W divided by F) were calculated for every rater. The ODF of this study is not directly comparable to standard DMRS ODF scores. We recognized that this adaptation represents a methodological trade-off: while it enables the implementation of a structured LLM-based workflow, it reduces comparability with prior DMRS studies that relied on the original seven-level ODF scoring.

Procedure

Material and segmentation of transcripts

The study task was to analyze excerpts and identify defense mechanisms present only in the patient's interventions. The therapist's lines were retained where necessary to understand the context, but the coding unit was always the patient's speech. Therefore, the first two transcribed sessions of the psychoanalytic treatment of the clinical case of "Annie" were segmented into short excerpts, each corresponding to a clinically coherent sequence of the patient's speech. Segmentation has been chosen to improve LLMs' performance by mitigating context-window and processing limitations and ensuring more stable, reliable analyses. Analyzing the entire session in a single submission proved suboptimal due to the LLMs' context-length and computational constraints. When very long transcripts were processed simultaneously, performance and response stability decreased, resulting in lower analytical coherence and output quality.

LLMs prompting

The LLMs were prompted only with the session transcripts and the standardized instructions and did not receive any human coding, summaries, or supervisory notes. The prompt here proposed is an example of few-shot prompting, because alongside the instruction, the prompt also includes a few demonstrations or examples (the "shot") of the final results we want to obtain (Brown et al., 2020). It was created through an iterative process of calibration and refinement applied to both the content and the structure (e.g., Madaan et al., 2023). The quantity and quality of information provided to the LLM had to be balanced with two main objectives: 1) all text was considered (e.g., all defense mechanisms throughout the content text), and 2) the response was appropriate to the context without hallucinations. The final prompt was divided into steps (See Supplementary File A for the entire prompt).

In the first phase (*familiarization*), the LLM was instructed to serve as an expert evaluator in psychodynamic psychotherapy using the DMRS. The prompt initially included a brief, theory-informed definition of defense mechanisms drawn from the psychodynamic literature[2]. The operational framework of the task was grounded in the DMRS manual and defense literature. Accordingly, the prompt provided cues to facilitate defense detection, including: (a) a list of possible conversational "triggers" to monitor, such as sudden shifts in tone or theme, affective inconsistencies, contradictions, lack of affect where expected, and sudden persecutory attributions. Triggers were derived from Perry's (2014) framework for identifying defensive "anomalies" in affect, behavior, and discourse; (b) Operational definitions and brief clinical examples of 12 target defense mechanisms; (c) key diagnostic distinctions (e.g., sublimation vs. displacement; denial vs. repression vs. isolation of affect; genuine humor vs. devaluing sarcasm) to reduce classification errors. The authors summarized the information on defense definitions and confidence score criteria through consensus and adapted it into prompting instructions.

In the second phase (*coding)*, the LLM received the qualitative criteria for assigning the confidence score (0, 1, or 2) to each mechanism. The transcript was presented as numbered excerpts. For each extract, the model must detect whether none, one, or more defenses were present. For each defense identified, LLMs provided a structured output row in table format (CSV) with the following fixed columns: Session_ID, Extract_ID, Rater_ID, Defense (i.e., label), Confidence (i.e., diagnostic certainty), Quote (i.e., patient's textual quote justifying the assignment), and Explanation (i.e., brief clinical justification related to the quote). LLMs were restricted to identifying the predefined set of 12 defenses and were not permitted to propose additional mechanisms beyond this list. The LLM should not have produced additional text beyond the required CSV line(s), thereby ensuring standardized, directly analyzable output. If no defense was detected in an excerpt, the output was "no defense found".

In the third phase (uploading), we submitted the therapy session transcripts to the LLMs. Specifically, for ChatGPT-5 (thinking mode), we used the platform's file upload feature to upload a document containing the session text. In contrast, for Gemini 2.5 Pro, we copied and pasted the

[1] The weights' values were chosen heuristically to preserve ordinal maturity while keeping computation simple.

[2] Defenses were defined as a mostly unconscious mental operation aimed at protecting the individual from intolerable affects or thoughts and preserving the integrity of the self (Cramer, 1998; APA, 2013).

transcript directly into the prompt field. In both cases, each therapy session was divided into two consecutive parts prior to submission. In total, four transcript segments were analyzed across the two sessions (two segments per session).

Each therapy session was processed in a separate interaction, and the LLMs were re-initialized between sessions to prevent any cross-session learning effects and to ensure that outputs were generated independently. The prompt explicitly instructed the models not to consider the content of previous conversations. Within each session, however, the two segments were analyzed in the same conversation thread. This allowed the model to retain contextual information from the first segment when processing the second, thereby maintaining internal coherence within the single session while avoiding memory carryover across sessions.

Human raters coding

By design, the human raters were untrained: as recommended by Perry and Henry (2004), they read the DMRS-abstracted research coding descriptions—the operational definitions of the 12 adapted defenses and the hierarchical logic of the three clusters, the same materials provided to the LLMs (see Supplementary File A)—but were otherwise untrained or naive to identifying defenses according to the system, without DMRS certification or prior calibration exercises. This places the human raters at a level of formal DMRS training equivalent to that of the LLMs. However, the two groups differed in clinical background: the human raters were four clinicians, psychodynamically oriented, licensed psychotherapists with concurrent postdoctoral positions, each with ≥5 years of clinical and research experience. Subsequently, clinicians (LA, FF, CR, FDS) conducted coding exercises on session transcripts under supervision by O.O., including case discussions and clarification of defensive ambiguities, before focusing on the material analyzed in the study. Contrary to Perry and Henry's (2004) recommendation, raters were not trained to meet a predefined reliability criterion prior to coding (while inter-rater agreement was evaluated for Annie's transcripts), reflecting the realistic conditions of a clinician using an LLM-based support tool in their practice or supervision. Indeed, we preferred to preserve an ecological and "non-optimized" situation, since we aimed to integrate human and AI coding into a regular clinical context (where not everyone is a certified DMRS rater).

The four clinicians formed two pairs, LA with FF and CR with FDS, following a two-step procedure. First, a different pair of clinicians worked together (LA and FF for Session 1; CR and FDS for Session 2). Within each pair, the two raters independently coded all extracts without consulting each other or LLMs' outputs and recorded the time required to complete the task. For each defense identified, every human rater provided a structured output row in table format (CSV) with the following fixed columns: Session_ID, Extract_ID, Rater_ID, Defense, Confidence, Quote, and Explanation.

Second, each pair of raters compared their versions of the same session with a supervisor (OO) to produce a single, consensual, integrated rating (LAFF for the first session and CRFDS for the second). Specifically, agreements and disagreements within the pair were discussed in supervision, leading to a new list of defense mechanisms, with some defenses retained and others discarded by clinicians through consensual reasoning. In parallel, ChatGPT and Gemini annotations for each session were merged (creating GPTGEM evaluations for the first and second sessions) via a purely mechanical union, without consensus discussion or adjudication. Then, for each extract, any defense flagged by at least one LLM was included, with duplicates collapsed (i.e., a defense flagged by both LLMs was counted once). Therefore, the integrated ratings across human pairs (LAFF in the first session and CRFDS in the second) were compared with the LLM-merged evaluation (GPTGEM), again in a supervised context. Considering the defense mechanisms identified only by the LLMs, during supervision, each pair of human raters chose which mechanisms to include and which to exclude based on consensual clinical reasoning and the DMRS definitions of defenses. This step was taken to determine whether LLMs could serve as a stimulus for clinical reasoning during supervision, thereby suggesting a defense that clinicians must either accept or decline.

Data analysis

The analysis was conducted at the session level, using, for each rater, the count of each of the 12 DMRS defenses. For each session, a 12×k matrix was constructed (rows = defenses; columns = raters), including all 12 defenses for all raters, with absent defenses coded as zero counts. This choice can inflate between-rater differences and widen confidence intervals, but it preserves comparability across raters. Reliability was estimated using the Intraclass Correlation Coefficient (ICC). In particular, ICC (2, 1) assumes a two-way random-effects model, in which raters are considered a sample from a larger population. ICC (2, 1) assesses absolute agreement and penalizes systematic differences between raters, returning the expected reliability of a single rater's measurement. ICC (2, k) uses the same random model but estimates the average reliability of k raters, which is typically higher because averaging attenuates idiosyncratic variability among individuals. The guidelines of Koo and Li (2016) were used to interpret the ICC: values < 0.50 indicate poor reliability, 0.50–0.75 moderate, 0.75–0.90 good, and >0.90 excellent. If the 95% CI includes zero or falls below 0.50, poor reliability cannot be ruled out. Inter-rater

agreement was calculated for: (a) all raters jointly in the session (k = 4), (b) human × human pairs, (c) human × LLM pairs, (d) LLM × LLM pairs, and (e) integrated human rater evaluations (group supervision) vs. integrated LLMs (LAFF vs. GPTGEM for the first session; CRFDS vs. GPTGEM for the second one).

The entire analytic workflow (i.e., data management and the computation of ICC) was implemented in the R statistical software environment (RStudio), using the *dplyr*, *tidyr*, *purrr*, and *tibble* packages for data handling and the *irr* package for the reliability analyses.

Results

The LLMs were able to match each defense in each extract with a confidence score, report the quote, and provide an explanation. On average, obtaining structured output from the LLMs (i.e., pasting the prompt in chat, submitting the transcript, and extracting CSV-formatted responses) required approximately 15 min of operator time, compared to approximately 3.5 h of manual coding by human raters. Therefore, both LLMs consistently generated explicit defense labels and accompanying rationales when guided by the standardized prompt, supporting the feasibility of the prompting-based pipeline. Examples of the structured defense-coding output produced by an LLM and a human rater are provided in Appendix 1.

Descriptively, LLMs (especially Gemini) tended to have higher defense frequencies than a human. Similarly, LLM-integrated assessments (GPTGEM) showed higher frequency scores than human-integrated assessments/supervision (LAFF/CRFDS) in both sessions (first session: F + 52%, W + 12%; second session: F + 32%, W + 22%). Conversely, humans (and human supervision) showed a higher ODF than LLMs (and their integration), both in the first session (ODF_{LAFF} = 2.48 vs. ODF_{GPTGEM} = 2.08) and in the second (ODF_{CRFDS} = 2.11 vs. ODF_{GPTGEM} = 1.94). Overall, LLMs counted more defenses and a lower average ODF; humans counted fewer, yielding a higher ODF score. This configuration remained stable from the first to the second session, albeit with variations in amplitude. Table 1 shows the frequency of defense mechanisms identified in each session by each rater (clinicians, LLMs, and supervision/integrations) and ODF indexes.

Notably, the LAFF supervision identified 10 defenses shared with GPTGEM. Additionally, 11 mechanisms were detected exclusively by humans. The integrated version of the LLMs proposed 21 defenses that humans did not recognize (7 were later validated under supervision, while 14 were rejected). Moreover, the CRFDS integrated evaluation shared 21 defenses with GPTGEM. Also, 16 mechanisms

Table 1 Defense counting and overall defense functioning scores

Session	Rater	Sublimation	Humor	Anticipation	Affiliation	Displacement	Isolation	Rationalization	Repression	Projection	Acting out	Splitting	Denial	F	W	ODF
S1	LA	0	6	0	4	1	0	2	1	0	0	0	1	15	37	2.47
S1	FF	0	3	2	3	0	0	1	1	0	0	0	2	12	29	2.42
S1	ChatGPT	0	1	3	1	0	1	1	5	0	1	0	1	14	30	2.14
S1	Gemini	0	2	1	0	0	2	6	4	1	1	2	1	20	32	1.60
S1	LAFF	0	6	2	6	1	0	2	2	0	0	0	2	21	52	2.48
S1	GPTGEM	0	3	4	1	0	3	7	7	1	2	2	2	32	58	1.81
S2	CR	0	8	4	2	1	3	7	0	1	1	2	1	30	62	2.07
S2	FDS	0	7	3	2	1	2	6	1	0	1	2	0	25	53	2.12
S2	ChatGPT	0	6	1	1	0	2	2	5	2	0	1	0	20	43	2.15
S2	Gemini	1	2	5	2	2	5	6	2	1	1	6	1	34	63	1.85
S2	CRFDS	0	9	5	3	2	4	7	1	1	1	3	1	37	78	2.11
S2	GPTGEM	1	6	6	3	2	6	7	6	3	1	7	1	49	95	1.94

The table reports, for each Session and Rater, the occurrence counts of the 12 DMRS defense mechanisms over the entire session. The LAFF and CRFDS rows indicate the integrated ratings of the two human raters (joint supervision), while GPTGEM indicates the union of the two LLM assessments (ChatGPT and Gemini) per session: for each extract, a defense common to both is counted only once; a defense encountered by only one rater is counted once. Any absences are recorded as 0. The columns from Sublimation to Denial are frequencies. F = number of detected defenses; W = sum of every product between the defense's frequency and its corresponding weight (in this adapted scoring system, mature defenses were weighted 3, neurotic defenses 2, and immature defenses 1); ODF = the average level of a person's defensive maturity (i.e., W divided by F)

Table 2 Distribution of defense mechanisms identified by supervision and LLM integration across sessions

Session	Supervision	Supervision-only	In common	GPTGEM-only (accepted: rejected)	Total
S1	LAFF	11	10	21 (7: 14)	42
S2	CRFDS	16	21	28 (14: 14)	65

The table shows the number of defense mechanisms identified by raters in different conditions. For each session, "Supervision-only" indicates defenses identified only by the integrated human supervision (LAFF/CRFDS), "In common" indicates defenses identified by both supervision and GPTGEM, and "GPTGEM-only" indicates defenses initially flagged only by GPTGEM. The numbers in parentheses under GPTGEM-only indicate how many of those GPTGEM-only defenses were later judged clinically valid ("accepted") or not ("rejected") during human supervision. "Total" is the number of unique defenses considered in the session, derived from the union of all distinct defenses discussed in that session

Table 3 Mean pairwise ICCs by Rater

Rater	Number of ratings	Mean ICC(2,1)	Mean ICC(2,k)
FDS	3	0.671	0.786
CR	3	0.638	0.758
FF	3	0.346	0.440
ChatGPT	6	0.305	0.425
LA	3	0.283	0.357
Gemini	6	0.257	0.364

The table reports, for each rater, the mean of all two-rater pairwise ICC estimates involving that rater (human×human, human×LLM, LLM×LLM), shown for both ICC(2,1) and ICC(2,k). The number of ratings indicates the number of pairwise comparisons contributing to each rater's mean; LLMs appear in both sessions and therefore contribute to more pairwise comparisons

were found only by the human pair, whereas 28 were identified solely by the LLM pair (14 accepted and 14 rejected). Overall, across both sessions, 31 defenses were detected by both humans and LLMs, 27 only by the human pair, and 49 only by the LLMs. Of these, 21 were accepted (should have been included in the report), and 28 were rejected by human supervision. Hence, LLMs' output can be useful for detecting defenses in supervision, stimulating clinical reasoning, and considering overlooked mechanisms. Table 2 compares defenses identified only by human supervision, only by LLMs, and the number of defenses proposed exclusively by LLMs that were accepted or rejected by a clinician during supervision.

The reliability of the DMRS scores varied markedly between the two sessions: considering all raters together, the first session showed a lower overall agreement (ICC (2, 1) = 0.253, 95% CI [−0.010, 0.614]; ICC (2, k) = 0.576, 95% CI [−0.039, 0.864]), while in the second session the reliability was significantly higher (ICC (2, 1) = 0.527, 95% CI [0.247, 0.798]; ICC (2, k) = 0.817, 95% CI [0.568, 0.940]). As summarized in Table 3 (and detailed in Appendix 2), mean pairwise ICCs were generally higher for human raters than for both LLMs; ChatGPT and Gemini showed broadly similar performance, and overall reliability was higher in Session 2 than in Session 1. Indeed, the human-by-human ICC was high, both in the first (LA vs. FF: ICC (2, 1) = 0.720, 95% CI [0.283, 0.910]; ICC (2, k) = 0.837, 95% CI [0.443, 0.953]) and in the second session (CR vs. FDS: ICC (2, 1) = 0.952, 95% CI [0.808, 0.987]; ICC (2, k) = 0.975, 95% CI [0.893, 0.993]), confirming the good reproducibility of the coding. The LLM×LLM comparison showed poor inter-rater reliability in the first session (ChatGPT vs. Gemini: ICC (2, 1) = 0.401, 95% CI [−0.180, 0.779]; ICC (2, k) = 0.573, 95% CI [−0.433, 0.876]) and in the second (ICC (2, 1) = 0.048, 95% CI [−0.445, 0.565]; ICC (2, k) = 0.092, 95% CI [−1.580, 0.721]), suggesting that the two LLMs are not interchangeable (see Appendix 2).

The human-by-LLM pairs in the first session showed weak or highly variable agreement, indicating discrepancies in level and profile across defenses. However, several pairs reached higher agreement in the second session (see Appendix 2). The same pattern was found in the integrated assessments (LAFF/CRFDS supervision vs. GPTGEM integration), with low agreement in the first session (LAFF vs. GPTGEM: ICC (2, 1) = 0.140, 95% CI [−0.430, 0.640]; ICC (2, k) = 0.246, 95% CI [−1.490, 0.780]) and good agreement in the second (CRFDS vs. GPTGEM: ICC (2, 1) = 0.649, 95% CI [0.184, 0.882]; ICC (2, k) = 0.787, 95% CI [0.299, 0.938]).

Discussion

The present proof-of-concept study evaluated whether general-purpose LLM interfaces could be used via structured prompting to support the identification of defense mechanisms in psychotherapy transcripts. Overall, the findings suggest that the article's aims were satisfied: (1) the proposed pipeline was feasible, since the models consistently generated defense labels with accompanying rationales; (2) LMMs' outputs could be compared with clinician ratings through inter-rater agreement metrics, and the workflow could be implemented in supervision contexts without requiring programming expertise to stimulate clinical reasoning, considering overlooked defense mechanisms.

Specifically, the LLMs and human raters produced outputs that were identical in structure (e.g., rows and column names), confirming the prompt's validity. From an operator's perspective, obtaining the structured output from the LLMs took roughly 15 min (prompting, submission, extraction), compared to approximately 3.5 h of manual coding by human raters. Additionally, the LLMs were able to assign a confidence score to each defense, indicating that these scores can be used both quantitatively and qualitatively.

However, because the confidence scores should have been assigned for the entire session in accordance with DMRS guidelines, the scores were not used in the data analysis. Although the scores are not interpretable, this result indicates that the LLM can codify the defense and estimate its confidence. As LLMs gain more computing power in the future, it may be possible to encode per-session rather than per-extract, without sacrificing precision for power.

Overall, LLMs identified more defenses and underestimated ODF than human raters did. These results suggest two complementary phenomena: LLMs tend to be more "extensive," that is, they detect more defenses, and clinicians tend to be more "selective," organizing that material along a more mature defensive profile on average. This is consistent with previous studies on LLMs applied to clinical assessment, which show that LLMs tend to apply a lower threshold for flagging potential clinical phenomena, resulting in more false positives relative to the human standard (Tasca et al., 2023). Moreover, human therapists tend to contextualize and grade severity, while LLMs assign labels more quickly and are less sensitive to the relational context and the function of the behavior (Scholich et al., 2025; Acevedo et al., 2025). One possible explanation for the discrepancy involves two separate, potentially independent mechanisms. First, LLMs may apply a lower threshold for identifying a potential defense, being more strongly cued by literal, manifest behavior (i.e., particularly linguistically salient or behaviorally explicit expressions). In contrast, human raters appear more attuned to the idiosyncratic, context-dependent meaning of speech (e.g., recognizing whether a patient's generalizing statement is simply a way of summarizing what happened or instead serves to maintain distance from feelings about that event). This phenomenon could explain why LLMs identified more defenses and also produced more false positives. Second, since immature defenses tend to have clearer behavioral referents, they may be easier for LLMs to detect from text alone, whereas neurotic and mature defenses depend more on tone, register, and relational nuance. Given the limited dataset and the study's proof-of-concept nature, these interpretations are preliminary and should be examined in larger, more diverse samples before drawing firm conclusions.

Furthermore, human pairs showed a high ICC in the first session and an excellent ICC in the second. Indeed, human raters generally showed higher mean pairwise reliability than both LLMs, although one clinician performed closer to the model-level estimates. This result confirms that, even without formal DMRS certification and prior quantitative calibration of reliability, clinicians can produce fairly convergent ratings when raters share a psychodynamic language and discuss the operational definitions of defenses (Perry & Henry, 2004). Consequently, if clinicians strongly agree with one another, discrepancies with LLMs are due not only to the material's ambiguity but also to the models' limitations. Conversely, the ChatGPT vs. Gemini comparison shows poor reliability across both sessions. This indicates that the two models did not perform the same interpretative work by applying a manual; rather, they generated different readings of the defenses. The finding that two LLMs are not interchangeable limits the ability to reliably automate psychometric tools and treat them as a single, stable instrument. An alternative explanation is that the prompt was not specific, leading to low agreement among the LLMs. This finding raises the question of how LLM reliability and validity could be improved in future applications of this kind.

When comparing assessments between a human rater and an LLM, integrated assessments (i.e., clinical supervision and merged LLM evaluation) showed the same pattern: low reliability in the first session and fair-to-good reliability in the second. Since the agreement on the defensive profile depends on the clinical material, it is plausible that the first session, in which Annie and the therapist met for the first time, featured a more chaotic, fragmented discourse with fewer defenses. This is consistent with the idea that the assessment of defense mechanisms depends on the patient's state at that moment in treatment and the degree of organization of their discourse (Perry, 1990/2014). In addition, the higher agreement observed in the second session may partly reflect the greater number and variability of defenses identified, which can facilitate higher ICC estimates. Indeed, the first session was an intake, with many brief, factual answers to queries that were often devoid of defenses, likely leading to fewer defenses than in the second session. Notably, this difference is unlikely to reflect learning effects across sessions, as each human rater pair coded only one session, and the LLMs were re-initialized and prompted separately for each session.

Regarding this re-initialization design, it is worth noting that both LLMs were deliberately reset between sessions and explicitly instructed not to draw on the content of prior conversations, mirroring the blinding of human raters to each other's prior ratings. This design choice ensures that the higher reliability observed in Session 2 cannot be attributed to the LLMs retaining patterns or calibrations from Session 1. At the same time, this represents a departure from how a clinician would typically work. Indeed, unlike a clinician who builds a longitudinal formulation of a patient over time by drawing on memory from prior sessions, the LLMs' pipeline did not allow access to cross-session information that might, in principle, improve consistency or contextual understanding. Future work could explore designs in which a summary of prior session outputs is reintroduced into the prompt for subsequent sessions to test whether this

improves consistency across sessions or instead introduces anchoring bias toward earlier ratings.

Finally, LLMs seemed to stimulate clinical reasoning. Indeed, human group supervision recognized that a portion of the defenses reported only by LLMs (21 out of 49) were clinically relevant and should have been included in the case formulation. Therefore, LLMs could serve as an exploration or "second look" to help the therapist focus on defensive passages that may have been overlooked, in line with the idea that generative models can support and offer insights that need to be critically evaluated (Riva et al., 2025; Wang et al., 2025). LLMs cannot be a diagnostic substitute; rather, they are a tool for broadening the clinician's attention, especially in the early stages of therapy, where the quality and function of defenses are still emerging and can easily be underestimated in the here-and-now of the session.

Strengths and limitations

This study has several notable strengths. First, it is a proof-of-concept study conducted on real clinical material, in which a pipeline (transcription, clinical coding, prompting, LLM coding, and comparison) has been applied to psychodynamic psychotherapy sessions. Furthermore, the prompt was designed to be usable by therapists without advanced technical skills.

However, the study also has several limitations. First, the results are not generalizable, as they are based on a single case, only two initial therapy sessions, and an altered version of the DMRS (i.e., limiting to 12 defenses instead of 30, modifying the ODF coding method, setting confidence scores at the extract level, and summarizing the instructions for defense definitions and confidence criteria). Nevertheless, this study aimed to evaluate the workflow feasibility rather than establish generalizable performance across patients or to compare results with prior studies that relied on the original DMRS version. Moreover, this study did not aim to automate the DMRS but used the tool only as a "guideline" for coding; a simplified version of DMRS was the option that best balanced LLM accuracy and power, as currently available. Improvements in LLM intelligence and computational power over time may alter this balance.

Second, the human raters were psychodynamically trained clinicians but not formally certified raters of the full DMRS, and, contrary to the recommendations of Perry and Henry (2004), they were not trained to meet a predefined reliability criterion prior to coding. Their preparation relied primarily on manual-based familiarization and supervision exercises rather than on structured competency-based DMRS training, which places them closer to non-certified or "naive" DMRS raters. This choice, which deviates from standard DMRS procedures, is due to the study's proof-of-concept nature. Indeed, our primary objective was to simulate a realistic context in which a therapist or supervisor uses LLM support without being a certified DMRS evaluator (rather than to obtain generalizable estimates of human agreement or to verify that the LLMs were as reliable as the standardized DMRS). Accordingly, the level of agreement observed in this study should be interpreted as an exploratory result rather than as evidence of performance under formal DMRS training conditions.

Third, the two LLMs (ChatGPT and Gemini) were not interchangeable, indicating that AI is not yet a single, standardizable tool; using different LLMs produces incomparable results. Fourth, the session contents were provided to the LLMs after segmentation to improve their performance, which may have influenced the analysis, as it is not certain that defenses correspond precisely to segments, thereby partially modifying clinical practice. Moreover, the location and number of segment boundaries could, in principle, influence which defenses are identified; future studies should systematically vary segmentation strategies (e.g., clinically determined units versus fixed-length or mechanically determined segments, and segmentation performed by different raters) to test this potential confound. Fifth, important details about the patient, such as her history, were missing, hindering the interpretation of her defense mechanisms. Furthermore, relying exclusively on the transcript not only obscures prosodic elements but also denies access to the nonverbal dimensions of communication (e.g., facial expressions, gestures, and posture) that video recordings could have captured. These nonverbal cues are integral to the interpersonal process and provide essential information for identifying and interpreting underlying defense mechanisms.

Conclusion

The purpose of this study was to evaluate the feasibility of using LLMs (ChatGPT and Gemini) to identify defense mechanisms in transcripts of actual psychodynamic psychotherapy sessions and compare their assessments with those of human psychodynamic clinicians. We did not seek to demonstrate the definitive accuracy of LLMs or to propose them as a substitute for clinical judgment, but rather to test a practical pipeline that therapists could adopt without specialized training in structured observational tools such as the DMRS. In summary, the LLMs produced the expected outputs much more quickly than human raters, but clinicians showed higher inter-rater reliability, highlighting that ChatGPT and Gemini were not yet interchangeable. The integration between clinicians and LLMs was promising, as

the LLMs reported more defenses, some of which were later confirmed in supervision as clinically relevant, suggesting that the models may be more useful for supporting case formulation and supervision.

Three main developments can be outlined. First, rigorous studies are needed to control for intervening variables and verify whether the study's results are replicable. For instance, larger samples (i.e., more patients, more treatment phases, different therapeutic orientations) are mandatory. Moreover, the effects of LLMs on clinical ratings need to be tested using experimental designs. For example, future research could adopt a controlled experimental design comparing independent human coding with coding conducted after reviewing LLM-generated defense suggestions to examine whether LLM support influences inter-rater agreement or the number and types of defenses identified. Future studies should also compare LLM outputs with fully DMRS-trained and certified raters to evaluate performance under standardized training conditions.

Second, future research should implement structured calibration procedures between clinicians and LLMs to determine whether achieving a minimum level of agreement improves reliability. In observational research with DMRS, human raters are typically required to reach a predefined minimum level of inter-rater agreement before proceeding with formal coding. Similarly, LLMs' outputs could be calibrated against human raters until a minimum agreement threshold is reached (e.g., ICC $\geq$ 0.30) before their use in clinical support or empirical testing. We believe that more refined prompting, iterative calibration procedures, or supervised alignment between clinicians and LLM outputs may improve reliability, reduce false positives, and better align with clinical judgment. Future studies could experimentally evaluate whether structured feedback loops or prompt-tuning phases improve agreement with clinically trained raters.

Third, as the complexity of LLM-based analyses of defense mechanisms may increase in the future given ongoing technological advancements, it may become feasible to process entire session transcripts without segmentation and to expand the full DMRS taxonomy beyond the 12 selected defenses. This expansion could help clarify whether the current, more constrained model tends to inflate defense counts under restricted label sets (e.g., restricting the coding to 12 defenses may lead raters or LLMs to assign a broader category like rationalization, because more specific options, such as intellectualization, were not included among the available labels).

In conclusion, the combined use of LLMs by clinicians is practically feasible and could improve the detection of defense mechanisms in psychotherapy. Currently, LLMs do not replace the therapist's judgment or human supervision; instead, they can serve as focused tools for case formulation, clinical training, and expanding clinicians' perspectives. Framed as AI for clinical reasoning support, LLMs help generate hypotheses and rationales and gauge uncertainty, enhancing decision-making while keeping the human decisively at the center.

Appendix 1

Table 4 Example rows of structured defense coding produced by ChatGPT and a human rater (FDS)

Session_ID	Extract_ID	Rater_ID	Defense	Confidence	Quote	Explanation
S2	E003	chatgpt	Humor	2	(laughs) I think if I lay down I'll never get up again.	A joking exaggeration about lying down helps her keep talking about an anxiety-provoking topic.
S2	E003	FDS	Humor	1	(laughs) I think if I lay down I'll never get up again.	Releases tension related to the fear of losing control while relaxing on the couch.
S2	E005	chatgpt	Rationalization	1	Well, that could be the Catholic upbringing too.	Offers a socially acceptable moral explanation for reluctance to lie down, deflecting more personal motives.
S2	E005	FDS	Rationalization	1	Well, that could be the Catholic upbringing too.	Attributes her difficulty in letting go while lying down to religious upbringing or to the jealousy she fears her husband may develop, thereby deflecting from more personal emotional motives and potentially interfering with the therapeutic process.

The table reports verbatim examples of structured coding rows generated by ChatGPT and by a human rater (FDS) for the same transcript excerpts. Examples are provided for illustrative purposes only and do not represent the defense-coding dataset. The authors originally wrote human explanations in Italian and translated them into English.

Appendix 2

Table 5 Detailed pairwise ICC(2,1) and ICC(2,k) estimates

Session	Raters	Pair	Comparisons	ICC (2,1)	95% CI (2,1)	ICC (2, k)	95% CI (2, k)
S1	Integrated (Humans vs. LLMs)	LAFF vs. GPTGEM	2	0.140	[−0.430, 0.640]	0.246	[−1.490, 0.780]
S2	Integrated (Humans vs. LLMs)	CRFDS vs. GPTGEM	2	0.649	[0.184, 0.882]	0.787	[0.299, 0.938]
S1	All Raters	LA, FF, ChatGPT, Gemini	4	0.253	[−0.010, 0.614]	0.576	[−0.039, 0.864]
S2	All Raters	CR, FDS, ChatGPT, Gemini	4	0.527	[0.247, 0.798]	0.817	[0.568, 0.940]
S1	Human×Human	LA vs. FF	2	0.720	[0.283, 0.910]	0.837	[0.443, 0.953]
S2	Human×Human	CR vs. FDS	2	0.952	[0.808, 0.987]	0.975	[0.893, 0.993]
S1	Human×LLM	LA vs. ChatGPT	2	0.017	[−0.621, 0.585]	0.034	[−3.270, 0.738]
S1	Human×LLM	FF vs. ChatGPT	2	0.319	[−0.329, 0.747]	0.483	[−0.976, 0.855]
S1	Human×LLM	LA vs. Gemini	2	0.112	[−0.513, 0.636]	0.202	[−2.090, 0.777]
S1	Human×LLM	FF vs. Gemini	2	0.000	[−0.542, 0.550]	0.000	[−2.370, 0.709]
S2	Human×LLM	CR vs. ChatGPT	2	0.466	[−0.079, 0.806]	0.636	[−0.172, 0.893]
S2	Human×LLM	FDS vs. ChatGPT	2	0.579	[0.039, 0.857]	0.733	[0.079, 0.923]
S2	Human×LLM	CR vs. Gemini	2	0.497	[−0.099, 0.826]	0.664	[−0.216, 0.904]
S2	Human×LLM	FDS vs. Gemini	2	0.482	[−0.061, 0.814]	0.651	[−0.130, 0.897]
S1	LLM×LLM	ChatGPT vs. Gemini	2	0.401	[−0.180, 0.779]	0.573	[−0.433, 0.876]
S2	LLM×LLM	ChatGPT vs. Gemini	2	0.048	[−0.445, 0.565]	0.092	[−1.580, 0.721]

The table reports the inter-rater reliability calculated on the per-session counts for the 12 DMRS mechanisms. ICC (2, 1) and ICC (2, k) are estimated using the two-way random, absolute-agreement model for the single rater and the mean of k raters, respectively; the 95% confidence intervals (CIs) are shown in parentheses. Negative lower bounds in CI can occur with ICC in small-sample settings and indicate that poor reliability cannot be ruled out. The comparison column indicates the number of raters involved (2 for pairs, 4 for the overall comparison). The "Human × Human," "Human × LLM," and "LLM × LLM" categories distinguish the type of comparison; "All Raters" includes both humans and both LLMs in the session. The "Integrated (Human vs. LLM)" rows compare the integrated ratings obtained from joint supervision: LAFF/CRFDS vs. GPTGEM.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s12144-026-09685-3.

Acknowledgements The authors wish to thank the *Psychoanalytic Research Consortium* for making the anonymized clinical materials from the "Annie" case available for research purposes.

Funding Open access funding provided by Università Cattolica del Sacro Cuore within the CRUI-CARE Agreement.

Data availability The datasets used and/or analysed during the current study are available from the corresponding author on reasonable request.

Declarations

Ethics Statement This study utilizes a single, fully anonymized psychotherapy transcript ("Annie") that is already publicly accessible online. The human ratings of the transcript were performed by research team members, who are also co-authors of this article, acting in their professional capacity as clinical experts. No external human participants were involved, and no personal or sensitive data about the raters were collected or analyzed. Therefore, no formal ethics committee approval was required.

Informed consent The psychotherapy material analyzed in this study ("Annie") is a publicly available, fully anonymized clinical case that was originally collected and published elsewhere under the responsibility of the original authors and their ethical procedures, including informed consent or equivalent safeguards for publishing anonymized material. In this study, no new data were collected from patients, and the only individuals involved as "human raters" were members of the research team, who are co-authors of this article, acting in their professional capacity. No personal or sensitive data about the raters was collected or analyzed, and no additional informed consent was required.

Conflict of interest The authors reported no potential conflicts of interest.

References

Acevedo, S., Aneja, E., Opler, D. J., Valera, P., & Jarmon, E. (2025). Evaluating the efficacy of ChatGPT-3.5 versus human-delivered text-based cognitive-behavioral therapy: A comparative pilot study. *American Journal of Psychotherapy*. https://doi.org/10.1176/appi.psychotherapy.20240070

American Psychiatric Association, DSM-5 Task Force. (2013). *Diagnostic and statistical manual of mental disorders: DSM-5™* (5th ed.). American Psychiatric Publishing, Inc. https://doi.org/10.1176/appi.books.9780890425596

Andersson, L. (2008). Psychodynamic supervision in a group setting: Benefits and limitations. *Psychotherapy in Australia*, *14*(2), 36–43.

Andrews, G., Singh, M., & Bond, M. (1993). The defense style questionnaire. *The Journal of Nervous and Mental Disease, 181*(4), 246–256. https://doi.org/10.1097/00005053-199304000-00006

Babl, A., Grosse Holtforth, M., Perry, J. C., Schneider, N., Dommann, E., Heer, S., Stähli, A., Aeschbacher, N., Eggel, M., Eggenberg, J., Sonntag, M., Berger, T., & Caspar, F. (2019). Comparison and change of defense mechanisms over the course of psychotherapy in patients with depression or anxiety disorder: Evidence from a randomized controlled trial. *Journal of Affective Disorders, 252*, 212–220. https://doi.org/10.1016/j.jad.2019.04.021

Bowlby, J. (1980). *Attachment and loss.* Basic Books.

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.

Cramer, P. (1991). *The development of defense mechanisms: Theory, research, and assessment.* Springer-Verlag Publishing. https://doi.org/10.1007/978-1-4613-9025-1

Cramer, P. (1998). Coping and defense mechanisms: What's the difference? *Journal of Personality, 66*(6), 919–946. https://doi.org/10.1111/1467-6494.00037

Crasovan, D. I. (2014). Limits of methods used in analysing the psychological defense: Psychological defense mechanisms and coping mechanisms. *Journal of Educational Sciences and Psychology*, *4*(2), 76–84.

Di Giuseppe, M., Perry, J. C., Lucchesi, M., Michelini, M., Vitiello, S., Piantanida, A., Fabiani, M., Maffei, S., & Conversano, C. (2020). Preliminary reliability and validity of the DMRS-SR-30, a novel self-report measure based on the defense mechanisms rating scales. *Frontiers in Psychiatry, 11*, Article 870. https://doi.org/10.3389/fpsyt.2020.00870

Di Giuseppe, M., Perry, J. C., Petraglia, J., Janzen, J., & Lingiardi, V. (2014). *Defense mechanisms rating scales–Q-sort version (DMRS-Q)* [Database record]. APA PsycTests. https://doi.org/10.1037/t69996-000

Di Giuseppe, M., Perry, J. C., Prout, T. A., & Conversano, C. (2021). Editorial: Recent empirical research and methodologies in defense mechanisms: Defenses as fundamental contributors to adaptation. *Frontiers in Psychology, 12*, Article Article 802602. https://doi.org/10.3389/fpsyg.2021.802602

Ewbank, M. P., Cummins, R., Tablan, V., Catarino, A., Buchholz, S., & Blackwell, A. D. (2021). Understanding the relationship between patient language and outcomes in internet-enabled cognitive behavioural therapy: A deep learning approach to automatic coding of session transcripts. *Psychotherapy Research*, *31*(3), 326–338. https://doi.org/10.1080/10503307.2020.1788740

Fattah, F. H., Salih, A. M., Salih, A. M., Asaad, S. K., Ghafour, A. K., Bapir, R., Abdalla, B. A., Othman, S., Ahmed, S. M., Hasan, S. J., Mahmood, Y. M., & Kakamad, F. H. (2025). Comparative analysis of ChatGPT and Gemini (Bard) in medical inquiry: A scoping review. *Frontiers in Digital Health, 7*, Article Article 1482712. https://doi.org/10.3389/fdgth.2025.1482712

Fonagy, P., Steele, M., Steele, H., Moran, G. S., & Higgitt, A. C. (1991). The capacity for understanding mental states: The reflective self in parent and child and its significance for security of attachment. *Infant Mental Health Journal*, *12*(3), 201–218. https://doi.org/10.1002/1097-0355(199123)12:3%3C201::AID-IMHJ2280120307%3E3.0.CO;2-7

Freud, A. (1946). *The ego and the mechanisms of defence.* International Universities Press.

Freud, S. (1937). Konstruktionen in der Analyse [Constructions in analysis]. *Internationale Zeitschrift für Psychoanalyse*, *23*, 459–469.

George, C., Main, M., & Kaplan, N. (1985). *Adult Attachment Interview (AAI)* [Database record]. APA PsycTests. https://doi.org/10.1037/t02879-000

Gleser, G. C., & Ihilevich, D. (1969). An objective instrument for measuring defense mechanisms. *Journal of Consulting and Clinical Psychology, 33*(1), 51–60. https://doi.org/10.1037/h0027381

Gu, J., Liu, J., Zeng, L., Yu, Y., Qiu, Y., Yue, Y., Tong, M., Yang, F., & Zhang, X. (2025). Comparing ChatGPT and validated questionnaires in assessing loneliness and online social support among college students: A cross-sectional study. *Scientific Reports, 15*, Article Article 20621. https://doi.org/10.1038/s41598-025-06358-2

Hawa, S., Akella, S., Kaushik, S., Joshi, V., & Kalbande, D. (2020). Analysis of therapy transcripts using natural language processing. *International Journal of Engineering and Advanced Technology, 9*(6), 489–493. https://doi.org/10.35940/ijeat.F1598.089620

Hirosawa, T., Mizuta, K., Harada, Y., & Shimizu, T. (2023). Comparative evaluation of diagnostic accuracy between Google Bard and physicians. *The American Journal of Medicine*, *136*(11), 1119–1123. https://doi.org/10.1016/j.amjmed.2023.08.003

Hsieh, L.-H., Liao, W.-C., & Liu, E.-Y. (2024). Feasibility assessment of using ChatGPT for training case conceptualization skills in psychological counseling. *Computers in Human Behavior: Artificial Humans, 2*(2), Article 100083. https://doi.org/10.1016/j.chbah.2024.100083

Jiang, H., Wu, Q., Lin, C. Y., Yang, Y., & Qiu, L. (2023). LLMLingua: Compressing prompts for accelerated inference of large language models. In H. Bouamor, J. Pino, & K. Bali (Eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (pp. 13358–13376). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.emnlp-main.825

Koo, T. K., & Li, M. Y. (2016). A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. *Journal of Chiropractic Medicine*, *15*(2), 155–163. https://doi.org/10.1016/j.jcm.2016.02.012

Lingiardi, V. (2017). Psychodynamic Diagnostic Manual (PDM-2). In N. McWilliams (Ed.), Guilford Press.

Lingiardi, V., & Madeddu, F. (2023). *I meccanismi di difesa: Teoria, valutazione, clinica*. Raffaello Cortina Editore.

Liu, N. F., Lin, K., Hewitt, J., Paranjape, A., Bevilacqua, M., Petroni, F., & Liang, P. (2024). Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, *12*, 157–173. https://doi.org/10.1162/tacl_a_00638

Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., & Neubig, G. (2023). Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, *55*(9), 1–35.

Madaan, A., Tandon, N., Gupta, P., Hallinan, S., Gao, L., Wiegreffe, S., & Clark, P. (2023). Self-refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems*, *36*, 46534–46594.

Mitchell, S. A. (1988). *Relational concepts in psychoanalysis: An integration.* Harvard University Press.

Murray, H. A. (1943). *Thematic apperception test.* Harvard University Press.

Neubauer, J. C., Kaiser, A., Lettermann, L., Volkert, T., & Häge, A. (2025). Performance of large language models ChatGPT and Gemini in child and adolescent psychiatry knowledge assessment. *PLoS One, 20*(9), Article e0332917. https://doi.org/10.1371/journal.pone.0332917

Paulhus, D. L. (1991). Measurement and control of response bias. In J. P. Robinson, P. R. Shaver, & L. S. Wrightsman (Eds.), *Measures of personality and social psychological attitudes* (pp. 17–59). Academic Press.

Perry, J. C. (1990). *Defense Mechanism Rating Scales (DMRS)* (5th Ed.). Cambridge.

Perry, J. C. (2014). Anomalies and specific functions in the clinical identification of defense mechanisms. *Journal of Clinical Psychology*, *70*(5), 406–418. https://doi.org/10.1002/jclp.22085

Perry, J. C., & Bond, M. (2012). Change in defense mechanisms during long-term dynamic psychotherapy and five-year outcome. *The American Journal of Psychiatry*, *169*(9), 916–925. https://doi.org/10.1176/appi.ajp.2012.11091403

Perry, J. C., & Cooper, S. H. (1986). A preliminary report on defenses and conflicts associated with borderline personality disorder. *Journal of the American Psychoanalytic Association*, *34*(4), 863–893. https://doi.org/10.1177/000306518603400405

Perry, J. C., & Henry, M. (2004). Studying defense mechanisms in psychotherapy using the Defense Mechanism Rating Scales. In U. Hentschel, G. Smith, J. G. Draguns, & W. Ehlers (Eds.), *Defense mechanisms: Theoretical, research and clinical perspectives* (pp. 165–192). Elsevier Science Ltd. https://doi.org/10.1016/S0166-4115(04)80034-7

Raile, P. (2024). The usefulness of ChatGPT for psychotherapists and patients. *Humanities and Social Sciences Communications*, *11*, 47. https://doi.org/10.1057/s41599-023-02567-0

Riva, G., Rossi, C., Frisone, F., & Oasi, O. (2025). Intelligenza Artificiale in Psicologia: Potenziare la Cura, Preservare l'Umanità. *Psicoterapia Cognitiva e Comportamentale*, 31(2).

San Martini, P., Roma, P., Sarti, S., Lingiardi, V., & Bond, M. (2004). Italian version of the Defense Style Questionnaire. *Comprehensive Psychiatry*, *45*(6), 483–494. https://doi.org/10.1016/j.comppsych.2004.07.012

Schafer, R. (1954). *Psychoanalytic interpretation in Rorschach testing: Theory and application.* Grune & Stratton.

Scholich, T., Barr, M., Wiltsey Stirman, S., & Raj, S. (2025). A comparison of responses from human therapists and large language model–based chatbots to assess therapeutic communication: Mixed methods study. *JMIR Mental Health, 12*, Article e69709. https://doi.org/10.2196/69709

Stern, B. L., Caligor, E., Clarkin, J. F., Critchfield, K. L., Horz, S., MacCornack, V., Lenzenweger, M. F., & Kernberg, O. F. (2010). Structured Interview of Personality Organization (STIPO): Preliminary psychometrics in a clinical sample. *Journal of Personality Assessment, 92*(1), 35–44. https://doi.org/10.1080/00223890903379308

Tasca, A. N., Carlucci, S., Wiley, J. C., Holden, M., El-Roby, A., & Tasca, G. A. (2023). Detecting defense mechanisms from Adult Attachment Interview (AAI) transcripts using machine learning. *Psychotherapy Research, 33*(6), 757–767. https://doi.org/10.1080/10503307.2022.2156306

Tay, Y., Dehghani, M., Bahri, D., & Metzler, D. (2023). Efficient transformers: A survey. *ACM Computing Surveys, 55*(6), 1–28. https://doi.org/10.1145/3530811

Vaillant, G. E. (1998). *Adaptation to life*. Harvard University Press.

Wang, L., Bhanushali, T., Huang, Z., Yang, J., Badami, S., & Hightow-Weidman, L. (2025). Evaluating generative AI in mental health: Systematic review of capabilities and limitations. *JMIR Mental*.