Multi-Session Client-Centered Treatment Outcome Evaluation in Psychotherapy

Hongbin Na[1], **Tao Shen**[1], **Shumao Yu**[2], **Ling Chen**[1]
[1]Australian AI Institute, University of Technology Sydney, Australia
[2]KU Leuven, Belgium
hongbin.na@student.uts.edu.au
🤗 https://huggingface.co/datasets/UTSNLPGroup/TheraPhase

Abstract

In psychotherapy, therapeutic outcome assessment, or treatment outcome evaluation, is essential to mental health care by systematically evaluating therapeutic processes and outcomes. Existing large language model approaches often focus on therapist-centered, single-session evaluations, neglecting the client's subjective experience and longitudinal progress across multiple sessions. To address these limitations, we propose IPAEval, a client-Informed Psychological Assessment-based Evaluation framework, which automates treatment outcome evaluations from the client's perspective using clinical interviews. It integrates cross-session client-contextual assessment and session-focused client-dynamics assessment for a comprehensive understanding of therapeutic progress. Specifically, IPAEval employs a two-stage prompt scheme that maps client information onto psychometric test items, enabling interpretable and structured psychological assessments. Experiments on our new TheraPhase dataset, comprising 400 paired initial and completion stage client records, demonstrate that IPAEval effectively tracks symptom severity and treatment outcomes over multiple sessions, outperforming baseline approaches across both closed-source and open-source models, and validating the benefits of items-aware reasoning mechanisms.

Keywords: Large Language Model, Mental Health Support, Evaluation

1. Introduction

In psychotherapy, therapeutic outcome assessment, a.k.a treatment outcome (see Figure 1) evaluation under clinical settings, refers to the systematic evaluation of therapeutic processes and outcomes (Groth-Marnat, 2009), focusing on factors such as therapist effectiveness (Johns et al., 2019) and treatment efficacy (Jensen-Doss et al., 2018) to improve mental health care delivery. It plays a significant role in enhancing the quality and effectiveness of mental health care by providing actionable insights that guide therapists in refining their treatment approaches (Wampold and Imel, 2015), ultimately leading to better client outcomes and improved therapeutic relationships in real-world clinical practice (Maruish and Leahy, 2000).

Over the last years, the emergence of large language models (LLMs) has demonstrated their effectiveness in automatic evaluations, showing a high degree of alignment with human judgment when provided with proper instruction and contextual guidance (Liu et al., 2023; Li et al., 2024b; Kim et al., 2024). This aligns with 'LLMs-as-a-judge' paradigm, where LLMs are employed to simulate human evaluators by providing assessments upon natural language input (Zheng et al., 2023; Wang et al., 2024b). This was extended to therapeutic outcome assessment by harnessing LLMs' ability to model complex therapeutic procedures and interactions, offering a novel pathway for automating the assessment of therapeutic efficacy (Chiu et al., 2024; Lee et al., 2024; Li et al., 2024a).

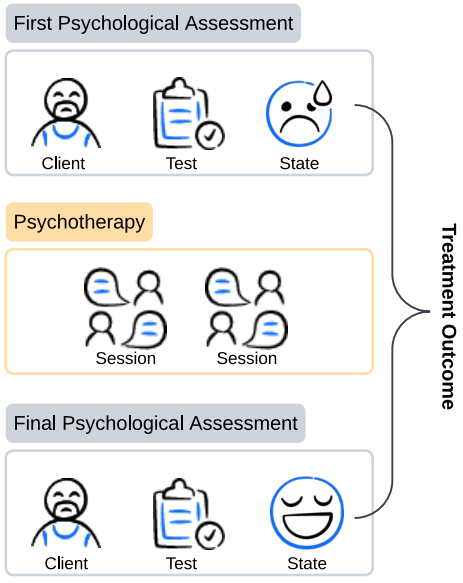

Figure 1: What is Treatment Outcome?

In the assessment, compared to psychometric tests (Furr, 2020) that are often constrained by the limitations of self-reported data, susceptibility to social desirability biases (Braun et al., 2001; Paulhus, 2017), clinical interviews not only provide richer, more nuanced insights into the client's emotional and behavioral states but also offer data that is more readily obtainable through natural, conversational interactions. Therefore, many recent works leverage clinical interviews, potentially enriched

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 2952–2968
11–16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

Method	Perspective	Theory Adherence	Reasoning	Evaluation Target
CPsyCoun (Zhang et al., 2024)	Therapist	✗	✗	Single Session
Cactus (Lee et al., 2024)	Therapist	✓	✗	Single Session
ClientCAST (Wang et al., 2024a)	Client	✓	✗	Single Session
IPAEval (Ours)	Client	✓	✓	Multiple Sessions

Table 1: Comparisons of IPAEval with other counterparts. **Perspective** indicates whether the evaluation is conducted from the therapist's or the client's point of view. **Theory Adherence** signifies whether the method is grounded in established psychological theories. **Reasoning** denotes whether the method involves generating intermediate reasoning steps before arriving at the final evaluation results. **Evaluation Target** refers to whether the method evaluates a single session or multiple sessions.

by the client's profile (Lee et al., 2024), to evaluate therapists from multiple perspectives, including behavioral labels (Chiu et al., 2024), skills adherence (Lee et al., 2024), and therapeutic rapport (Li et al., 2024a; Yosef et al., 2024), offering a holistic view of their effectiveness in psychotherapy.

While the above therapist-centered assessments focus on evaluating the therapist's techniques and adherence to therapeutic models, they often overlook the subjective experience and evolving needs of the client, limiting the depth of the evaluation (Wang et al., 2024a; Yosef et al., 2024). In contrast, client-centered assessments, such as *treatment outcome evaluation* in common practice, prioritize the client's perspective, offering a more comprehensive understanding of therapy's impact by capturing changes in the client's emotional, cognitive, and behavioral states across sessions (Hatfield and Ogles, 2004; Rogers, 2012). Although a concurrent work, ClientCAST (Wang et al., 2024a), presents an LLM-based client simulator for treatment outcome evaluations, which focuses on reducing harmful outputs and improving answering consistency, we stand fundamentally apart and never fabricate client responses that could distort the evaluation of treatment outcomes. What's worse, almost all previous approaches focus on evaluating individual therapy sessions in isolation, without considering the broader context of the client's journey across multiple sessions. This narrow scope limits the ability to assess longitudinal progress or capture the dynamic shifts in a client's mental state and therapeutic needs over time, which are crucial for a comprehensive treatment outcome evaluation (Hayes and Andrews, 2020).

Motivated by the above therapist-centered and single-session limitations (please see Table 1 for comparisons), we design a new evaluation framework, dubbed client-**I**nformed **P**sychological **A**ssessment-based **Eval**uation (IPAEval), for treatment outcomes in the format of clinical interviews.

Specifically, to achieve treatment outcome evaluation, we formulate an information extraction task that leverages clinical interviews to automatically populate psychometric tests for psychological assessments, bridging the gap between subjective client dialogues and standardized metrics. As such, treatment outcomes are evaluated through these assessments of clients conducted both before and after therapy, allowing for a more comprehensive understanding of therapeutic progress. Upon this new framework, we first propose a cross-session client-contextual assessment module that integrates client history and contextual information across multiple sessions to enhance the accuracy of psychological assessments. Then, we present a session-focused client-dynamics assessment module that evaluates the effectiveness of individual therapy sessions by tracking real-time client responses and treatment outcomes within each session. In the meantime, to boost reasoning capability in the extraction, we also present an items-aware reasoning prompt technique for psychometric test-oriented rationale generation.

To evaluate the proposed framework, we first develop a new dataset, called TheraPhase, based on CPsyCoun (Zhang et al., 2024), which includes transcripts from initial and final therapy sessions. This dataset offers valuable insights into therapy progress and serves as a key resource for evaluating psychological assessments and treatment outcomes across multiple sessions. Then, we tested nine LLMs, including both open- and closed-source models. These models were evaluated for their performance in psychological assessments and treatment outcome prediction, particularly in multi-session evaluations. IPAEval consistently tracked symptom severity and treatment outcomes across multiple sessions, a capability lacking in previous single-session models. Our ablation study confirmed that the items-aware reasoning mechanism significantly boosts model performance in both symptom detection and outcome prediction.

2. Methodology

Starting with a task definition (§2.1), we elaborate on our evaluation framework, called client-Informed Psychological Assessment-based Evaluation (IPAEval), which is mainly composed of 1) a *cross-session client-contextual assessment* module (§2.2) for client-tracking psychological assessment and 2) a *session-focused client-dynamics assessment* module (§2.3) to derive session-informed

treatment outcome evaluation. Please see Figure 2 for an overall illustration of our framework. Last but not least, as there is no precursor in clinical interviews-based treatment outcome evaluation, we curate a new dataset, called TheraPhase, as a testbed for IPAEval (§2.4).

2.1. Task Definition

Consider a client with profile p undergoing a series of therapy sessions $s_1, s_2, \ldots$. Our goal is to evaluate the treatment outcome of a given session s_k (or a combination of consecutive sessions). We denote the client information available after session s_k as $c_k = (p, s_k)$, which may be further enriched with prior session history when available (§2.2). We decompose the evaluation into two sequential subtasks, both performed by an LLM $\mathcal{M}$ instantiated differently for each stage.

Psychological Assessment. After session s_k, we conduct a psychological assessment based on the client information c_k and a set of psychometric tests $\mathcal{T}$:

$$\mathbf{a}_k = \mathcal{M}^{(\mathsf{a})}(c_k,\ \mathcal{T}), \tag{1}$$

where $\mathbf{a}_k$ denotes the resulting assessment scores.

Treatment Outcome Evaluation. By comparing assessments from different stages, we quantify the treatment outcome:

$$\mathbf{e}_k = \mathcal{M}^{(\mathsf{e})}(\mathbf{a}_k,\ \mathbf{a}_{<k}), \tag{2}$$

where $\mathbf{a}_{<k}$ refers to one or more prior assessments serving as the baseline. In the simplest case, $\mathbf{e}_k$ measures the change between an initial assessment and a post-treatment assessment. We detail the concrete instantiation in §2.3.

2.2. Cross-session Client-contextual Assessment

Existing research using client information with LLMs for psychological assessment, particularly for depression and PTSD, shows promising results (Galatzer-Levy et al., 2023; Arcan et al., 2024; Rosenman et al., 2024). However, these studies typically focus on specific symptoms and lack broad coverage of psychological conditions and transparency in interpreting scale results, which may erode trust among clinicians and clients, limiting clinical applications (Martin and Rouas, 2024).

To address these gaps, we introduce a two-stage prompt scheme that populates information from clinical interviews to fill psychometric tests by making the best of LLMs' capability in natural language understanding (Zhao et al., 2023a; Hua et al., 2024). It is applicable to various psychometric tests and specifically designed to provide interpretable psychological assessments. Without sacrificing generality, in this work we utilize a widely used and comprehensive psychometric test for screening psychological symptoms.

Stage 1: Items-Aware Reasoning. This stage extracts structured symptom evidence from client information. Inspired by Schulhoff et al. (2024), we design a prompt $p^{(\mathsf{ir})}$ (**i**tems-aware **r**easoning) that instructs the LLM to act as a psychologist, map client information onto individual items of the psychometric test $\mathcal{T}$, and produce an explanation for each item. Concretely, the LLM generates items-aware reasoning results:

$$\hat{\mathcal{X}} = \underset{\mathcal{X}}{\operatorname{argmax}}\ P_{\mathsf{LLM}}(\mathcal{X} \mid c_k,\ p^{(\mathsf{ir})}), \tag{3}$$

where c_k is the client information defined in §2.1 and $\hat{\mathcal{X}}$ is a set of structured records, each consisting of the extracted evidence, symptom category, specific symptom, presence judgment, and a detailed explanation. This approach helps clinicians quickly trace the source of evidence and understand the relevance of various symptoms. The detailed prompt and an example are provided in Appendix D and Appendix F, respectively.

Stage 2: Psychological Assessment. Building on the reasoning results, the LLM then scores the client across N symptom dimensions. We design a second prompt $p^{(\mathsf{sa})}$ (**s**ymptom **a**ssessment) that provides the psychometric test $\mathcal{T}$ together with simplified scoring criteria (at the dimension level rather than individual items, to account for the practical constraint that not all items are addressed in a session). The assessment scores are obtained as:

$$\hat{\mathbf{a}} = \underset{\mathbf{a}}{\operatorname{argmax}}\ P_{\mathsf{LLM}}(\mathbf{a} \mid c_k,\ \hat{\mathcal{X}},\ p^{(\mathsf{sa})}), \tag{4}$$

where $\hat{\mathbf{a}} \in \mathbb{R}^N$ contains the estimated score for each of the N symptom dimensions. The detailed prompt is provided in Appendix E.

Remark: Avoiding Excessive Speculation. In contrast to ClientCAST (Wang et al., 2024a), which simulates the client's own estimation of psychometric test scores, our approach adjusts score ranges to account for items not yet addressed by the client. This more accurately reflects the gradual disclosure of information over sessions and avoids biased assessments caused by unmentioned items.

2.3. Session-focused Client-dynamics Assessment

Given the assessment scores $\hat{\mathbf{a}} \in \mathbb{R}^N$ over N symptom dimensions, we compute the *Positive Symp-*

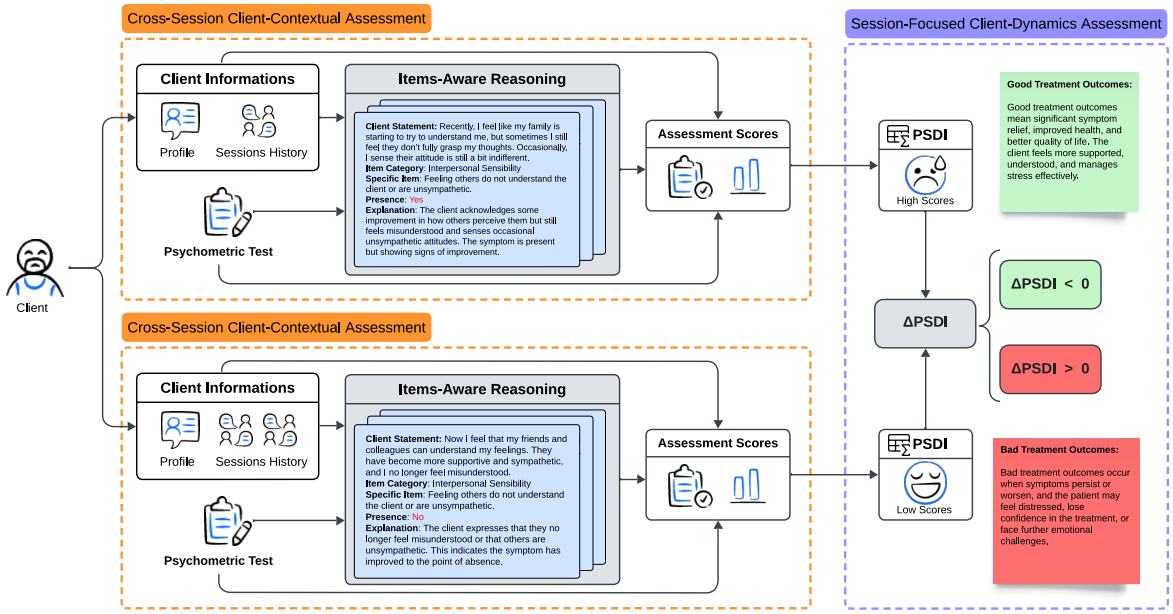

Figure 2: An illustration of client-informed psychological assessment-based evaluation (IPAEval).

tom Distress Index (PSDI) (Derogatis and Unger, 2010), which summarizes the average distress level across the dimensions where positive symptoms are detected. Let $\mathcal{P} \subseteq \{1, \ldots, N\}$ denote the subset of dimensions with positive scores (i.e., $\hat{\mathbf{a}}_i > 0$). The PSDI is defined as:

$$\text{PSDI} = \frac{1}{|\mathcal{P}|} \sum_{i \in \mathcal{P}} \hat{\mathbf{a}}_i, \tag{5}$$

where $|\mathcal{P}|$ denotes the number of positive dimensions.

To evaluate treatment outcomes, we apply the two-stage assessment (Eq. 3–4) to the client information at the initial stage c_i and the final stage c_f, yielding PSDI_i and PSDI_f respectively. The treatment outcome is then defined as:

$$\mathbf{e} := \Delta\text{PSDI} = \text{PSDI}_f - \text{PSDI}_i. \tag{6}$$

A negative ΔPSDI indicates symptom improvement after treatment.

Remark: Advantages and Versatility of PSDI. Although PSDI originates from the Symptom Checklist-90 (SCL-90) (Derogatis et al., 1973), the idea of averaging scores over positive items generalizes naturally to other psychometric tests, providing a flexible tool for tracking treatment progress.

2.4. Brief on TheraPhase Dataset Design

Popular datasets such as High-Low Quality (Pérez-Rosas et al., 2019) and AnnoMI (Wu et al., 2023) contain only single-session client information sourced from public videos, with no data for subsequent stages. To assess cross-stage changes, we construct the TheraPhase Dataset based on CPsyCoun (Zhang et al., 2024), which exhibits significant within-session changes. Our dataset includes 400 pairs of client information from the initial and completion stages of treatment.

To construct the dataset, we use 5-shot prompting with GPT-4 to extract the initial-stage information from each client's comprehensive record, forming paired data (initial vs. full) that enables analytical comparison between pre- and post-treatment conditions. Please see §3.2 and §A for details.

3. Experiments

Starting with the settings of IPAEval framework and involved models, we elaborate on datasets constructions and their auto eval metrics (plus human alignment results), followed by empirical results, ablations, and error analysis.

3.1. Experimental Settings

IPAEval Setup. The IPAEval framework is capable of handling various forms of client information, such as user profiles and interaction histories. However, due to data acquisition limitations, we primarily utilized consultation dialogue data as the main source of client information. Furthermore, IPAEval supports a variety of symptom-based psychometric tests, such as the General Health Questionnaire (GHQ) series (Montazeri et al., 2003), the Symptom Checklist (SCL) series, and the Brief Symptom Inventory (BSI) (Derogatis and Melisaratos, 1983). In this experiment, we utilized the SCL-90 (Derogatis et al., 1973), a widely recognized and

comprehensive tool for assessing a broad range of psychological symptoms. The scoring criteria for assessing symptoms are outlined in Table 2. Additionally, to ensure structured output, our code utilizes LangChain[1] and Pydantic[2] for better LLMs integration and data validation.

Score	Description
-1	Symptom not addressed.
0	Symptom addressed, but no symptoms found; no signs of distress or dysfunction.
1	Minimal symptoms, minor indications of distress but no significant dysfunction.
2	Clear symptoms, clear indications of distress, and significant dysfunction.

Table 2: Scoring Criteria for Symptom Assessment

Involved Models. We conducted an investigation into the performance of several closed and open-source LLMs. The closed-source models we tested include GPT-4 (Team, 2024b), GPT4o, GPT-4-turbo, and GPT-4o-mini, which represent the latest advancements in proprietary LLMs developed by OpenAI[3]. Additionally, we tested a variety of open-source models, such as Llama3.1-405B (Team, 2024a), Llama3.1-70B (Team, 2024a), Qwen2-72B (Yang et al., 2024), Mistral-8X22B (Jiang et al., 2024), and Mistral-8X7B (Jiang et al., 2024). These models vary significantly in terms of architecture, parameter size, and training data, providing a comprehensive overview of both commercial and community-driven LLM development. All of these models were invoked through API platforms[4].

3.2. Datasets Construction

For **psychological assessment**, we selected 2 datasets, High-Low Quality Counseling (Pérez-Rosas et al., 2019) and AnnoMI (Wu et al., 2023), consisting of counseling therapy transcripts extracted from publicly available videos on online platforms such as YouTube and Vimeo. But, there are issues of data duplication between these two datasets. Given the higher quality of data in AnnoMI, we have chosen to retain the AnnoMI data from the same sources. Furthermore, considering the context window limitation of one of our test models, GPT-4, the maximum number of dialogue turns is set to 102. To increase the challenge and ensure the dialogues are sufficiently complex for evaluating the model's capability in handling extended therapeutic conversations, the minimum number of turns is set at 25. Based on these, we have selected 110 client dialogue entries as our test data.

For **treatment outcomes**, we selected the TheraPhase Dataset. This dataset comprises treatment session transcripts that encompass two distinct phases of client interactions. Its advantage lies in the clear changes observable in clients across these phases, which aids in observing the treatment outcomes. The statistics of the resulting datasets are listed in Appendix A.

3.3. Auto Evaluation and Human Alignment

For **psychological assessment** (e.g., symptom detection), we assessed the model's ability to identify symptoms from client data using classification metrics such as Accuracy, Precision, Recall, and F1 scores (Binary, Macro, and Weighted). The scoring system assigned a value of -1 for a negative class and 0, 1, or 2 for positive classes. For symptom severity, we computed a PSDI score for each client and reported error metrics – Mean Squared Error (MSE) and Mean Absolute Error (MAE) – across three runs to ensure consistency.

For **treatment outcomes**, we measured changes in positive symptom severity using ΔPSDI, which reflects the difference in mean positive symptom scores between two assessments. A ΔPSDI greater than 0 indicates worsening or new symptoms, while a value of 0 or less suggests improvement or stability. We also evaluated the accuracy of predicting these changes using the same classification metrics.

References Generation. For psychological assessment, we manually annotated 30 client sessions with a Cohen's kappa of 0.73. Among the models evaluated, GPT-4o achieved the best performance, with an accuracy of 78.33% and a binary F1 score of 72.95%. Accordingly, GPT-4o was selected as the gold model for generating reference scores in this task. Similarly, for treatment outcomes evaluation, we annotated 60 sessions, reaching a Cohen's kappa of 0.81. GPT-4 delivered the top results with an accuracy of 74.44% and a binary F1 score of 83.44%, leading us to choose GPT-4 as the gold model for this evaluation. For further experimental details regarding References Generation, please refer to the Appendix C.

[1] https://www.langchain.com/

[2] https://docs.pydantic.dev/

[3] Specific versions of the OpenAI models used in the tests were `gpt-4-0613`, `gpt-4o-2024-05-13`, `gpt-4-turbo-2024-04-09`, `gpt-4o-mini-2024-07-18`.

[4] For the OpenAI models, we invoked them via https://platform.openai.com, Mistral models through https://console.mistral.ai/, Llama3.1 models via https://fireworks.ai/, and Qwen2 through https://www.together.ai/.

3.4. Evaluation Results across LLMs

Psychological Assessments. As shown in Table 3, GPT-4 achieved the best performance in symptom detection, excelling in both accuracy and binary F1 score, highlighting its strong ability to accurately identify symptoms. GPT-4-turbo demonstrated a more conservative approach with higher precision but lower recall, indicating it was more cautious in detecting symptoms but missed more cases. GPT-4o-mini excelled in recall but had reduced overall reliability due to a higher rate of false positives. Among open-source models, Qwen2-72B and Llama3.1-70B showed the closest performance to GPT-4, though they still fell short. Notably, Mistral-8X7B's extremely low recall was caused by a significant number of output formatting errors, leading to evaluation failures. We will further discuss these formatting issues in Appendix G.

In symptom severity assessment, GPT-4 once again stood out with the lowest MSE and MAE, making it the most accurate model. Although GPT-4o-mini and GPT-4-turbo showed more balanced results, they were less precise compared to GPT-4. Among open-source models, Llama3.1-70B performed the best, though the gap between open-source and closed-source models remained substantial. Furthermore, GPT-4 exhibited the greatest consistency and reliability, with minimal variance across runs, indicating robust performance. In contrast, GPT-4o-mini showed more variability in MAE and MSE, and open-source models generally exhibited less stability compared to their closed-source counterparts.

Treatment Outcomes. Table 4 compares the performance of closed-source and open-source models on treatment outcome evaluation tasks. Among the closed-source models, GPT-4-turbo achieved the highest scores across multiple metrics, making it the most effective model in treatment outcome prediction. GPT-4o and GPT-4o-mini displayed competitive performance but lagged slightly behind GPT-4-turbo. For the open-source models, Llama3.1-405B led the group with the highest accuracy and macro F1, demonstrating superior performance in treatment outcome tasks. Qwen2-72B and Llama3.1-70B also performed well, while Mistral-8X7B had the highest recall but struggled with lower F1 scores, indicating higher sensitivity but less consistent overall performance. Overall, both closed-source and open-source models showed strong capabilities, with GPT-4-turbo and Llama3.1-405B emerging as the top performers in their respective categories.

3.5. Further Analysis

Impact of Parameters on Performance. As shown in Figure 3, model parameter size has a clear impact on performance across tasks such as symptom detection, symptom severity evaluation, and treatment outcome prediction. Larger models consistently outperform smaller models, exhibiting higher F1 (Weighted) scores and lower MAE. This trend indicates that increasing model size enhances the model's ability to handle complex tasks (Wen et al., 2024), especially in identifying subtle patterns related to psychological symptoms and predicting treatment outcomes.

Impact of Items-aware Reasoning. The ablation study (Figure 4) demonstrates that items-aware reasoning is crucial for both psychological assessment and treatment outcomes evaluation. Removing this feature significantly decreased performance across all models. For psychological assessment, models like GPT-4o and GPT-4 less accurately detected symptoms and assessed severity, which lowered F1 scores and increased error metrics. Similarly, performance in treatment outcomes evaluation also dropped, though the impact was less pronounced. These results underscore that items-aware reasoning improves the models' precision for these tasks.

4. Related Work

Therapist Assessment using LLMs. LLMs' role-playing capabilities have led to increased interest in developing Role-Play Therapists (Chen et al., 2023; Chiu et al., 2024; Lee et al., 2024), but the lack of automated metrics for evaluating t herapist is a significant challenge. CPsyCoun (Zhang et al., 2024) employs an LLM-based evaluation method from the therapist's perspective to assess single session, specifically evaluating the therapist's comprehensiveness, professionalism, authenticity, and safety. Lee et al. (2024), Li et al. (2024a), and Yosef et al. (2024) similarly adopt a therapist's perspective with LLM-based evaluation, but they address CPsyCoun's lack of support from psychological theories by employing the Cognitive Therapy Rating Scale (Goldberg et al., 2020) for CBT skills assessment and the Working Alliance Inventory (Hatcher and Gillaspy, 2006) for evaluating the therapeutic relationship. Notably, BOLT (Chiu et al., 2024) applied LLMs to identify therapist behaviors, evaluating the quality of dialogue sessions based on the frequency and sequence of LLM therapist behaviors. Clinical evidence (Goodson et al., 2017; Mason et al., 2016) shows that better therapists are linked to improved outcomes, but evaluating therapists alone may miss how much the client is benefiting (Robinson, 2009). The treatment outcome

Models	Accuracy ↑	Precision ↑	Recall ↑	$F1_{Binary}$ ↑	$F1_{Macro}$ ↑	$F1_{Weighted}$ ↑	MSE ↓	MAE ↓
Closed-Source Models								
GPT-4	0.7973±0.01	0.7852±0.01	0.7121±0.01	0.7469±0.01	0.7889±0.01	0.7956±0.01	0.2100±0.02	0.3292±0.03
GPT-4-turbo	0.7561±0.00	0.8726±0.02	0.4913±0.01	0.6285±0.01	0.7234±0.00	0.7386±0.00	0.4055±0.05	0.4490±0.03
GPT-4o-mini	0.4915±0.00	0.4467±0.02	0.8824±0.01	0.5931±0.00	0.4576±0.01	0.4359±0.01	0.2245±0.01	0.3329±0.02
Open-Source Models								
Llama3.1-405B	0.7291±0.00	0.6960±0.01	0.6306±0.00	0.6616±0.00	0.7179±0.00	0.7269±0.00	0.3922±0.03	0.4476±0.01
Qwen2-72B	0.7385±0.00	0.7405±0.01	0.5815±0.01	0.6513±0.01	0.7210±0.00	0.7322±0.00	0.3962±0.01	0.4559±0.00
Llama3.1-70B	0.7333±0.01	0.7201±0.01	0.5974±0.01	0.6529±0.01	0.7182±0.01	0.7286±0.01	0.3379±0.01	0.4041±0.00
Mistral-8X22B	0.6215±0.00	0.5405±0.00	0.6616±0.02	0.5948±0.00	0.6198±0.00	0.6238±0.00	0.5205±0.03	0.5452±0.02
Mistral-8X7B	0.6070±0.00	0.6158±0.01	0.1710±0.02	0.2672±0.02	0.4993±0.01	0.5364±0.01	1.5711±0.02	1.0927±0.01

Table 3: Performance comparison between closed-source and open-source models across various evaluation metrics in **psychological assessment**. Metrics with an upward arrow ↑ indicate higher values are better, while metrics with a downward arrow ↓ indicate lower values are better. The results show mean values along with standard deviations for each metric. Cells highlighted in blue represent the best-performing results.

Models	Accuracy ↑	Precision ↑	Recall ↑	$F1_{Binary}$ ↑	$F1_{Macro}$ ↑	$F1_{Weighted}$ ↑
Closed-Source Models						
GPT-4o	0.6375±0.02	0.7706±0.01	0.7356±0.02	0.7526±0.01	0.5370±0.02	0.6448±0.02
GPT-4-turbo	0.6800±0.01	0.7824±0.01	0.7944±0.01	0.7883±0.00	0.5660±0.01	0.6772±0.01
GPT-4o-mini	0.6317±0.01	0.7727±0.01	0.7211±0.01	0.7459±0.01	0.5380±0.01	0.6420±0.01
Open-Source Models						
Llama3.1-405B	0.6958±0.01	0.7965±0.01	0.7989±0.01	0.7976±0.00	0.5925±0.02	0.6951±0.01
Qwen2-72B	0.6725±0.01	0.7747±0.01	0.7944±0.01	0.7844±0.01	0.5515±0.01	0.6679±0.01
Llama3.1-70B	0.6708±0.01	0.7796±0.01	0.7822±0.01	0.7809±0.01	0.5597±0.02	0.6703±0.01
Mistral-8X22B	0.6383±0.01	0.7544±0.00	0.7678±0.01	0.7610±0.01	0.5089±0.01	0.6350±0.01
Mistral-8X7B	0.6825±0.01	0.7469±0.00	0.8722±0.02	0.8046±0.01	0.4779±0.00	0.6413±0.00

Table 4: Comparison between closed and open-source models across various evaluation metrics in **treatment outcomes**.

evaluation based on client-centered psychological assessment focuses more on results, specifically determining whether the therapy has brought about meaningful changes in the client's life, which is the ultimate goal of the treatment (Groth-Marnat, 2009).

Client-centered Psychological Assessment. Client-centered psychological assessment combines psychometric tests with clinical interviews to provide a comprehensive understanding of individuals (Spoto et al., 2013). While tests offer standardized data on psychological traits, interviews yield deeper insights into personal experiences, addressing nuances that tests might miss (Groth-Marnat, 2009). Using multiple methods ensures a complete client assessment in clinical practice (Meyer et al., 2001; Groth-Marnat, 2009). Moreover, leveraging LLMs' advanced language processing capabilities (Luo et al., 2023; Zhao et al., 2023b) enables complex and diverse assessments, contrasting with earlier approaches that only detected individual symptoms (Ji et al., 2022; Zhai et al., 2024). A substantial body of research supports this advancement (Galatzer-Levy et al., 2023; Arcan et al., 2024; Rosenman et al., 2024). For example, several studies have employed LLMs to analyze interviews (Gratch et al., 2014), assessing depression and PTSD scores via widely used tests like (Kroenke et al., 2009) and PCL-C (Weathers et al., 1994). However, precise psychological assessments enable therapists to grasp the client's psychological state, but a psychological assessment alone cannot determine whether the treatment has brought about positive changes for the client.

Treatment outcomes evaluation complements psychological assessment by measuring the effectiveness of interventions over time (Maruish and Leahy, 2000). While psychological assessments provide a snapshot of the client's mental state, as shown in the Figure 1, treatment outcomes evaluation focuses on tracking changes in symptoms and overall well-being throughout the therapeutic process. This dynamic evaluation allows therapists to determine whether the treatment has been successful and as needed for improvement.

5. Conclusion

We presented IPAEval to overcome limitations of current therapeutic outcome evaluations by shifting the focus from therapist-centered, single-session assessments to a comprehensive, client-informed framework. By using clinical interviews and integrating cross-session client-contextual and session-focused client-dynamics assessments, it offers a

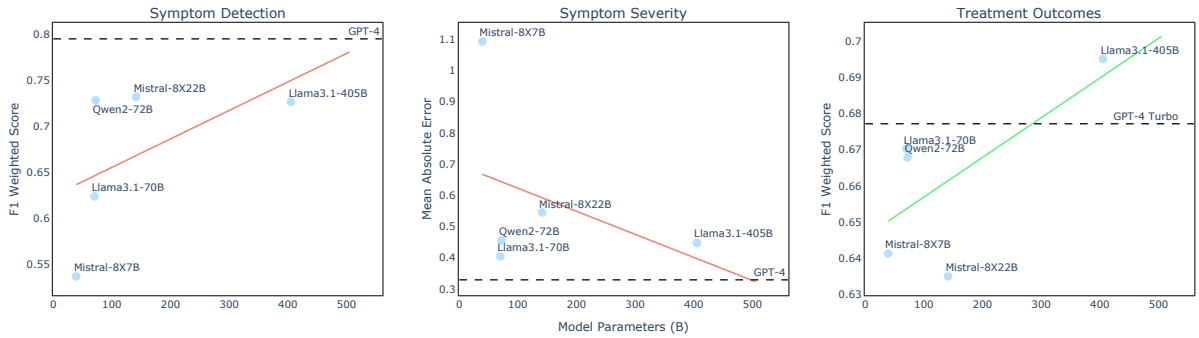

Figure 3: The impact of model parameters on symptom detection, symptom severity evaluation, and treatment outcome prediction. Dashed lines represent the best-performing closed-source models.

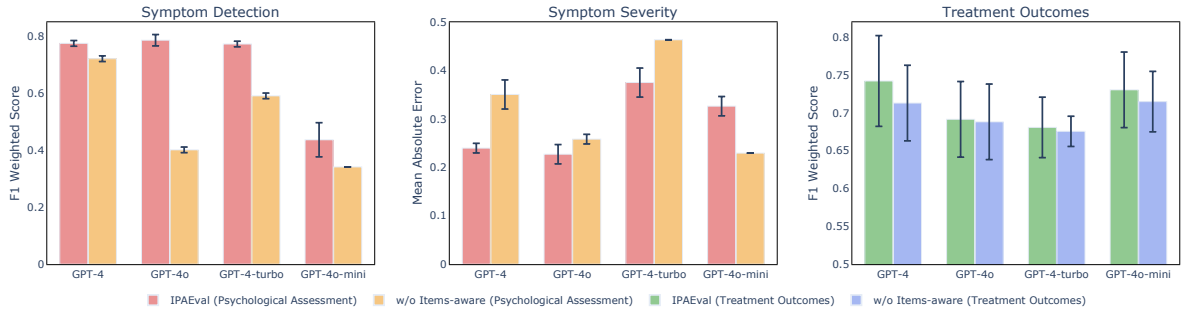

Figure 4: The impact of items-aware reasoning on psychological assessment and treatment outcomes evaluation using human-annotated data across four OpenAI models.

holistic evaluation of treatment outcomes. Experiments on our TheraPhase dataset validate its effectiveness in tracking symptom severity and progress across sessions. Overall, our results underscore the value of a client-centered, multi-session evaluation strategy for personalized and effective mental health interventions.

Limitations

The limitations of this paper are as follows: (1) Due to the shortage of professional psychological annotators, only two individuals were involved in a limited amount of data labeling. This resulted in fewer human-aligned experimental data. Future research should focus on developing more multi-session datasets that include psychological assessment scores. (2) As the amount of client information increases, smaller models with fewer parameters struggle to follow instructions effectively. This limits the scalability and performance of these models in more complex scenarios. Future research should explore strategies to enhance model adaptability in handling larger client information inputs. (3) Our multi-session dataset was derived by splitting single multi-turn conversations, which can represent client changes but cannot fully capture the characteristics of true multi-session data. In the future, we should develop authentic multi-session datasets.

Ethics Statement

All datasets used in this study are either fully synthetic—thus free of personally identifiable information—or have been rigorously anonymized and are employed strictly under their original usage agreements, eliminating privacy concerns. Nonetheless, large language models can generate inaccurate or biased outputs; relying on these results without professional oversight could mislead or harm clients. Consequently, IPAEval is intended solely as a *research-level* evaluation tool, never a substitute for clinical diagnosis or treatment, and every automated conclusion must be reviewed by a licensed mental-health professional. Because psychological symptoms manifest differently across cultures, genders, and age groups—and our current training data are primarily English- and Chinese-centric—we recognize limitations in cross-cultural generalization.

Bibliographical References

Mihael Arcan, David-Paul Niland, and Fionn Delahunty. 2024. An assessment on comprehending mental health through large language models.

Henry I Braun, Douglas N Jackson, and David E

Wiley. 2001. Socially desirable responding: The evolution of a construct. In *The role of constructs in psychological and educational measurement*, pages 61–84. Routledge.

Siyuan Chen, Mengyue Wu, Kenny Q. Zhu, Kunyao Lan, Zhiling Zhang, and Lyuchun Cui. 2023. Llm-empowered chatbots for psychiatrist and patient simulation: Application and evaluation.

Yu Ying Chiu, Ashish Sharma, Inna Wanyin Lin, and Tim Althoff. 2024. A computational framework for behavioral assessment of llm therapists.

Leonard R Derogatis, Ronald S Lipman, and Lino Covi. 1973. Scl-90: an outpatient psychiatric rating scale–preliminary report. *Psychopharmacol bull*, 9(1):13–28.

Leonard R Derogatis and Nick Melisaratos. 1983. The brief symptom inventory: an introductory report. *Psychological medicine*, 13(3):595–605.

Leonard R Derogatis and Rachael Unger. 2010. Symptom checklist-90-revised. *The Corsini encyclopedia of psychology*, pages 1–2.

R. Michael Furr. 2020. *Psychometrics in Clinical Psychological Research*, Cambridge Handbooks in Psychology, page 54–65. Cambridge University Press.

Isaac R. Galatzer-Levy, Daniel McDuff, Vivek Natarajan, Alan Karthikesalingam, and Matteo Malgaroli. 2023. The capability of large language models to measure psychiatric functioning.

Simon B Goldberg, Scott A Baldwin, Kritzia Merced, Derek D Caperton, Zac E Imel, David C Atkins, and Torrey Creed. 2020. The structure of competence: Evaluating the factor structure of the cognitive therapy rating scale. *Behavior Therapy*, 51(1):113–122.

Jason T Goodson, Amy W Helstrom, Emily J Marino, and Rachel V Smith. 2017. The impact of service-connected disability and therapist experience on outcomes from prolonged exposure therapy with veterans. *Psychological Trauma: Theory, Research, Practice, and Policy*, 9(6):647.

Jonathan Gratch, Ron Artstein, Gale Lucas, Giota Stratou, Stefan Scherer, Angela Nazarian, Rachel Wood, Jill Boberg, David DeVault, Stacy Marsella, David Traum, Skip Rizzo, and Louis-Philippe Morency. 2014. The distress analysis interview corpus of human and computer interviews. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3123–3128, Reykjavik, Iceland. European Language Resources Association (ELRA).

Gary Groth-Marnat. 2009. *Handbook of psychological assessment*. John Wiley & Sons.

Robert L Hatcher and J Arthur Gillaspy. 2006. Development and validation of a revised short version of the working alliance inventory. *Psychotherapy research*, 16(1):12–25.

Derek R Hatfield and Benjamin M Ogles. 2004. The use of outcome measures by psychologists in clinical practice. *Professional Psychology: Research and Practice*, 35(5):485.

Adele M Hayes and Leigh A Andrews. 2020. A complex systems approach to the study of change in psychotherapy. *BMC medicine*, 18:1–13.

Yining Hua, Hongbin Na, Zehan Li, Fenglin Liu, Xiao Fang, David Clifton, and John Torous. 2024. Applying and evaluating large language models in mental health care: A scoping review of human-assessed generative tasks.

Amanda Jensen-Doss, Emily M Becker Haimes, Ashley M Smith, Aaron R Lyon, Cara C Lewis, Cameo F Stanick, and Kristin M Hawley. 2018. Monitoring treatment progress and providing feedback is viewed favorably but rarely used in practice. *Administration and Policy in Mental Health and Mental Health Services Research*, 45:48–61.

Shaoxiong Ji, Tianlin Zhang, Luna Ansari, Jie Fu, Prayag Tiwari, and Erik Cambria. 2022. MentalBERT: Publicly available pretrained language models for mental healthcare. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 7184–7190, Marseille, France. European Language Resources Association.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2024. Mixtral of experts.

Robert G Johns, Michael Barkham, Stephen Kellett, and David Saxon. 2019. A systematic review of therapist effects: A critical narrative update and refinement to review. *Clinical Psychology Review*, 67:78–93.

Seungone Kim, Juyoung Suk, Shayne Longpre, Bill Yuchen Lin, Jamin Shin, Sean Welleck, Graham Neubig, Moontae Lee, Kyungjae Lee, and

Minjoon Seo. 2024. Prometheus 2: An open source language model specialized in evaluating other language models.

Kurt Kroenke, Tara W Strine, Robert L Spitzer, Janet BW Williams, Joyce T Berry, and Ali H Mokdad. 2009. The phq-8 as a measure of current depression in the general population. *Journal of affective disorders*, 114(1-3):163–173.

Suyeon Lee, Sunghwan Kim, Minju Kim, Dongjin Kang, Dongil Yang, Harim Kim, Minseok Kang, Dayi Jung, Min Hee Kim, Seungbeen Lee, Kyoung-Mee Chung, Youngjae Yu, Dongha Lee, and Jinyoung Yeo. 2024. Cactus: Towards psychological counseling conversations using cognitive behavioral theory.

Anqi Li, Yu Lu, Nirui Song, Shuai Zhang, Lizhi Ma, and Zhenzhong Lan. 2024a. Automatic evaluation for mental health counseling using llms.

Zhen Li, Xiaohan Xu, Tao Shen, Can Xu, Jia-Chen Gu, Yuxuan Lai, Chongyang Tao, and Shuai Ma. 2024b. Leveraging large language models for nlg evaluation: Advances and challenges.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: NLG evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.

Zheheng Luo, Qianqian Xie, and Sophia Ananiadou. 2023. Chatgpt as a factual inconsistency evaluator for text summarization.

Vincent P. Martin and Jean-Luc Rouas. 2024. Why voice biomarkers of psychiatric disorders are not used in clinical practice? deconstructing the myth of the need for objective diagnosis. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 17603–17613, Torino, Italia. ELRA and ICCL.

Mark E. Maruish and Robert L. Leahy. 2000. The use of psychological testing for treatment planning and outcome assessment. *Journal of Cognitive Psychotherapy*, 14:205 – 206.

Liam Mason, Nick Grey, and David Veale. 2016. My therapist is a student? the impact of therapist experience and client severity on cognitive behavioural therapy outcomes for people with anxiety disorders. *Behavioural and Cognitive Psychotherapy*, 44(2):193–202.

Gregory J Meyer, Stephen E Finn, Lorraine D Eyde, Gary G Kay, Kevin L Moreland, Robert R Dies, Elena J Eisman, Tom W Kubiszyn, and Geoffrey M Reed. 2001. Psychological testing and psychological assessment: A review of evidence and issues. *American psychologist*, 56(2):128.

Ali Montazeri, Amir Mahmood Harirchi, Mohammad Shariati, Gholamreza Garmaroudi, Mehdi Ebadi, and Abolfazl Fateh. 2003. The 12-item general health questionnaire (ghq-12): translation and validation study of the iranian version. *Health and quality of life outcomes*, 1:1–4.

Delroy L Paulhus. 2017. Socially desirable responding on self-reports. *Encyclopedia of personality and individual differences*, 1(5).

Verónica Pérez-Rosas, Xinyi Wu, Kenneth Resnicow, and Rada Mihalcea. 2019. What makes a good counselor? learning to distinguish between high-quality and low-quality counseling conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 926–935, Florence, Italy. Association for Computational Linguistics.

Bill Robinson. 2009. When therapist variables and the client's theory of change meet. *Psychotherapy in Australia*, 15(4):60–65.

Carl Rogers. 2012. *Client centered therapy (new ed)*. Hachette UK.

Gony Rosenman, Lior Wolf, and Talma Hendler. 2024. Llm questionnaire completion for automatic psychiatric assessment.

Sander Schulhoff, Michael Ilie, Nishant Balepur, Konstantine Kahadze, Amanda Liu, Chenglei Si, Yinheng Li, Aayush Gupta, HyoJung Han, Sevien Schulhoff, Pranav Sandeep Dulepet, Saurav Vidyadhara, Dayeon Ki, Sweta Agrawal, Chau Pham, Gerson Kroiz, Feileen Li, Hudson Tao, Ashay Srivastava, Hevander Da Costa, Saloni Gupta, Megan L. Rogers, Inna Goncearenco, Giuseppe Sarli, Igor Galynker, Denis Peskoff, Marine Carpuat, Jules White, Shyamal Anadkat, Alexander Hoyle, and Philip Resnik. 2024. The prompt report: A systematic survey of prompting techniques.

Andrea Spoto, Gioia Bottesi, Ezio Sanavio, and Giulio Vidotto. 2013. Theoretical foundations and clinical implications of formal psychological assessment. *Psychotherapy and psychosomatics*, 82(3):197–199.

Llama Team. 2024a. The llama 3 herd of models.

OpenAI Team. 2024b. Gpt-4 technical report.

Bruce E Wampold and Zac E Imel. 2015. *The great psychotherapy debate: The evidence for what makes psychotherapy work*. Routledge.

Jiashuo Wang, Yang Xiao, Yanran Li, Changhe Song, Chunpu Xu, Chenhao Tan, and Wenjie Li. 2024a. Towards a client-centered assessment of llm therapists by client simulation.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, and Zhifang Sui. 2024b. Large language models are not fair evaluators. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9440–9450, Bangkok, Thailand. Association for Computational Linguistics.

Frank W Weathers, B Litz, D Herman, J Juska, and T Keane. 1994. Ptsd checklist—civilian version. *Journal of Occupational Health Psychology*.

Bosi Wen, Pei Ke, Xiaotao Gu, Lindong Wu, Hao Huang, Jinfeng Zhou, Wenchuang Li, Binxin Hu, Wendy Gao, Jiaxin Xu, Yiming Liu, Jie Tang, Hongning Wang, and Minlie Huang. 2024. Benchmarking complex instruction-following with multiple constraints composition.

Zixiu Wu, Simone Balloccu, Vivek Kumar, Rim Helaoui, Diego Reforgiato Recupero, and Daniele Riboni. 2023. Creation, analysis and evaluation of annomi, a dataset of expert-annotated counselling dialogues. *Future Internet*, 15(3).

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jianxin Yang, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Xuejing Liu, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, Zhifang Guo, and Zhihao Fan. 2024. Qwen2 technical report.

Stav Yosef, Moreah Zisquit, Ben Cohen, Anat Klomek Brunstein, Kfir Bar, and Doron Friedman. 2024. Assessing motivational interviewing sessions with AI-generated patient simulations. In *Proceedings of the 9th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2024)*, pages 1–11, St. Julians, Malta. Association for Computational Linguistics.

Wei Zhai, Hongzhi Qi, Qing Zhao, Jianqiang Li, Ziqi Wang, Han Wang, Bing Xiang Yang, and Guanghui Fu. 2024. Chinese mentalbert: Domain-adaptive pre-training on social media for chinese mental health text analysis.

Chenhao Zhang, Renhao Li, Minghuan Tan, Min Yang, Jingwei Zhu, Di Yang, Jiahao Zhao, Guancheng Ye, Chengming Li, and Xiping Hu. 2024. Cpsycoun: A report-based multi-turn dialogue reconstruction and evaluation framework for chinese psychological counseling.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. 2023a. A survey of large language models.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. 2023b. A survey of large language models.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Language Resource References

Verónica Pérez-Rosas, Xinyi Wu, Kenneth Resnicow, and Rada Mihalcea. 2019. High- vs. low-quality counseling conversations corpus. Dataset derived from public video sources. Transcripts from YouTube/Vimeo demonstration videos; availability varies; see paper for details.

Zixiu Wu, Simone Balloccu, Vivek Kumar, Rim Helaoui, Diego Reforgiato Recupero, and Daniele Riboni. 2023. Annomi: Expert-annotated counselling dialogues. Dataset. Professionally transcribed MI dialogues with expert annotations; openly available.

A. Dataset Statistics

As shown in Table 5, the datasets used in our study consist of both English and Chinese dialogue data. The High-Low Quality Counseling + AnnoMI dataset provides a challenging benchmark due to its long conversations, averaging around 80 utterances per session. In contrast, the TheraPhase dataset, which is constructed to simulate multi-session counseling, consists of shorter but more focused interactions, with an average of 11.5 utterances per session and a significantly higher word count per utterance. These structural differences highlight the diverse nature of our datasets, ensuring that models are tested under varying conversational conditions.

B. Scoring Criteria for Symptom Assessment

Table 2 provides an overview of the scoring system for symptom assessment. It is worth noting that, unlike traditional questionnaires where clients can provide imaginative responses to questions that are not directly addressed during a consultation, our evaluation process may leave certain symptoms unaddressed. Consequently, we include a score of -1 to indicate that the symptom was not addressed.

C. Detailed Experimental Setup and Results for References Generation

In this appendix, we provide detailed experimental settings and results for the References Generation task, covering both the psychological assessment and treatment outcomes evaluation.

C.1. Psychological Assessment

For the psychological assessment task, we manually annotated 30 client sessions, achieving a Cohen's kappa of 0.73. Table 6 summarizes the performance of the evaluated models on various metrics including Accuracy, Precision, Recall, Binary F1, Macro F1, Weighted F1, Mean Squared Error (MSE), and Mean Absolute Error (MAE). The best performing results for each metric are highlighted in blue. The results show that GPT-4o achieved the highest Accuracy (78.33%) and Binary F1 score (72.95%), and was therefore selected as the gold standard model for generating reference scores in this task.

C.2. Treatment Outcomes Evaluation

For the treatment outcomes evaluation, we annotated 60 sessions, obtaining a Cohen's kappa of 0.81. Table 7 presents a detailed comparison of the models on Accuracy, Precision, Recall, Binary F1, Macro F1, and Weighted F1. The experimental results indicate that GPT-4 delivered the best performance with an Accuracy of 74.44%, along with superior Recall and Binary F1 scores. Consequently, GPT-4 was chosen as the gold standard model for this evaluation.

Overall, these experiments provide a comprehensive comparison of the four models (GPT-4, GPT-4o, GPT-4-turbo, and GPT-4o-mini) across different tasks, thereby justifying the selection of the gold standard models for References Generation.

Datasets	Language	# of Clients	# of Sessions	Avg. # of Utterances	Words per Utterance
High-Low Quality Counseling AnnoMI	English	110	110	79.8 (std = 26.1)	22.2 (std = 27.1)
TheraPhase	Chinese	400	800	11.5 (std = 6.3)	41.7 (std = 20.9)

Table 5: Summary of key characteristics of the selected datasets, including language, number of clients, sessions, average number of utterances per session, and the average word count per utterance.

Models	Accuracy ↑	Precision ↑	Recall ↑	$F1_{Binary}$ ↑	$F1_{Macro}$ ↑	$F1_{Weighted}$ ↑	MSE ↓	MAE ↓
GPT-4	0.7744±0.01	0.6792±0.01	0.7187±0.01	0.6984±0.01	0.7591±0.01	0.7757±0.01	0.1369±0.02	0.2398±0.01
GPT-4o	0.7833±0.02	0.6674±0.02	0.8043±0.03	0.7295±0.02	0.7744±0.02	0.7867±0.02	0.1207±0.01	0.2272±0.02
GPT-4-turbo	0.7800±0.01	0.7503±0.03	0.5933±0.01	0.6623±0.01	0.7495±0.01	0.7734±0.01	0.2379±0.03	0.3754±0.03
GPT-4o-mini	0.4844±0.04	0.4079±0.02	0.9144±0.04	0.5634±0.01	0.4641±0.05	0.4370±0.06	0.1962±0.03	0.3265±0.02

Table 6: Comparison of different models on various performance metrics using human-annotated data in psychological assessment. Metrics with an upward arrow ↑ indicate higher values are better, while metrics with a downward arrow ↓ indicate lower values are better. The results show mean values along with standard deviations for each metric. Cells highlighted in blue represent the best-performing results.

Models	Accuracy ↑	Precision ↑	Recall ↑	$F1_{Binary}$ ↑	$F1_{Macro}$ ↑	$F1_{Weighted}$ ↑
GPT-4	0.7444±0.06	0.8285±0.04	0.8406±0.04	0.8344±0.04	0.6370±0.08	0.7423±0.06
GPT-4o	0.6778±0.06	0.8219±0.02	0.7391±0.07	0.7770±0.05	0.5939±0.05	0.6916±0.05
GPT-4-turbo	0.6778±0.06	0.8046±0.02	0.7681±0.11	0.7815±0.05	0.5660±0.04	0.6809±0.04
GPT-4o-mini	0.7222±0.04	0.8625±0.06	0.7681±0.05	0.8090±0.03	0.6410±0.08	0.7306±0.05

Table 7: Comparison of different models on various performance metrics using human-annotated data in treatment outcomes.

D. Items-Aware Reasoning Prompts

Prompt: Items-Aware Reasoning

Role:
Imagine you are a skilled psychologist adept at identifying potential symptoms from interview. You can explain how these symptoms relate to the SCL-90 symptom checklist and specific items within it.

Directives:
Your task is to determine the presence or absence of symptoms from the Client's statements and provide detailed reasons for your assessment. Extract specific parts related to SCL-90 symptoms from the Client's statements. For each extracted part, indicate whether the symptom is present or not, and explain why this text is related to the SCL-90 symptom and specific item. If a symptom is mentioned but not present, extract that part and explain why the symptom is not present.SCL-90 is a psychological symptom assessment tool with 90 items, each evaluating different aspects of psychological distress.

Additional Information:
Symptom Checklist-90:
<Psychometric Test>
Presence of Symptoms: Extract the relevant part of the Client's statement. Indicate that the symptom is present. Explain why this text indicates the presence of the SCL-90 symptom and specify the item.
Absence of Symptoms: Extract the part where the symptom is mentioned but not present. Indicate that the symptom is not present. Explain why this text does not indicate the presence of the SCL-90 symptom despite the mention.

Output Formatting:
<Format Instructions>

Client Information:
<Interview>

Please extract specific parts related to SCL-90 symptoms from the Client's statements. For each extracted part, indicate whether the symptom is present or not, and explain why this text is related to the SCL-90 symptom and specific item. If a symptom is mentioned but not present, extract that part and explain why the symptom is not present.

E. Psychological Assessment Prompts

Prompt: Psychological Assessment

Role:
As a psychologist specializing in this evaluation task, based on the following interview and the extracted Symptom Checklist-90 (SCL-90) symptom-related content and explanation, provide a qualitative score (-1-2) for each symptom category.

Score Criteria:
Scoring criteria: -1 (Symptom not addressed in the interview), 0 (Symptom addressed in the interview, but no symptoms found, no signs of distress or dysfunction), 1 (Minimal symptoms, minor indications of distress but no significant dysfunction), 2 (Clear symptoms, clear indications of distress and significant dysfunction).

Directives:
Please note that this qualitative assessment is based on the state at the end of the interview. There may be noticeable symptoms during the interview, but these symptoms may become clarified or alleviated as the discussion progresses.

Additional Information:
<Psychometric Test>

Output Formatting:
<Format Instructions>

Client Information:
<Interview>
<Item-aware Reasoning Result>

Please extract specific parts related to SCL-90 symptoms from the Client's statements. For each extracted part, indicate whether the symptom is present or not, and explain why this text is related to the SCL-90 symptom and specific item. If a symptom is mentioned but not present, extract that part and explain why the symptom is not present.

F. Items-Aware Reasoning Output

Table 8 shows an example of the Items-Aware Reasoning output derived from a therapy session transcript. In this example, the system extracts a key client statement—specifically, the client's admission of consistently prioritizing others over themselves—and categorizes it under the symptom category "Interpersonal Sensibility." The output further identifies a specific symptom ("Feeling others do not understand the client or are unsympathetic") and marks its presence as "Yes." An explanation is provided to clarify that the client's behavior might indicate feelings of being misunderstood or a lack of empathy from others. Additionally, an assessment score for Interpersonal Sensitivity is generated, demonstrating how the system quantifies this symptom. This example illustrates the system's capability to reason about session content and map client statements to relevant psychological constructs.

G. Output Formatting Errors

In our two experiments, OpenAI series models produced no errors in output formatting, whereas open-source models encountered numerous issues. Specifically, the Figure 5 below shows the error statistics for open-source models during the Assessment task, with the main issue being incorrect output that did not follow the Pydantic-defined JSON format.

SESSION:
Therapist: So, thank you for coming in today.
Client: Yes.
Therapist: How are you feeling today?
Client: I feel great actually.
Therapist: Yeah? Good.
Client: Yeah.
Therapist: Good.
Client: I feel good.
Therapist: And so you did your clarifications, value clarifications-
Client: Yeah.
Therapist: -and what are your top five?
Client: Yes. It was a good, uh, experience for me. It was different. It was different than usual. There were several things that were different, and, uh, the number one value that I put was self-respect. And I-I don't even know if self-respect has ever been in my top five let alone my number one.
Therapist: Really?
Client: Yeah. And, um—
Therapist: Do you have any idea why that is?
Client: I do have an idea, I think, why that is. Um, I think that there's been a few things that have happened recently and something that really came to my awareness, when I visited with my family, is that **I have consistently through my whole life, probably, put other people first. And I have consistently, uh, almost not even considered myself in the equation.** It was, uh, kind of sad in a way, at the time that I realized it. Uh, I didn't realize how severe it actually was, but I was kind of glad that I realized it because I feel like it's never too late to change-
Therapist: True.
Client: -and I feel like I can- I can, uh, respect and value myself just as much as I have other people. I know that's important. And I feel like when I do that, I'm a better person for other people as well.
Therapist: Mm-hmm. By not putting yourself on the back burner so much?
.....

ITEMS-AWARE REASONING RESULT:
Client Statement: **I have consistently through my whole life, probably, put other people first. And I have consistently, uh, almost not even considered myself in the equation.**
Symptom Category: Interpersonal Sensibility
Specific Symptom: Feeling others do not understand the client or are unsympathetic.
Presence: Yes
Explanation: The client's statement indicates that they have been prioritizing others over themselves, which could be a sign of feeling misunderstood or not receiving empathy from others.
.....

ASSESSMENT SCORE:
.....; Interpersonal Sensitivity: 1;.....

Table 8: Items-Aware Reasoning Output Example

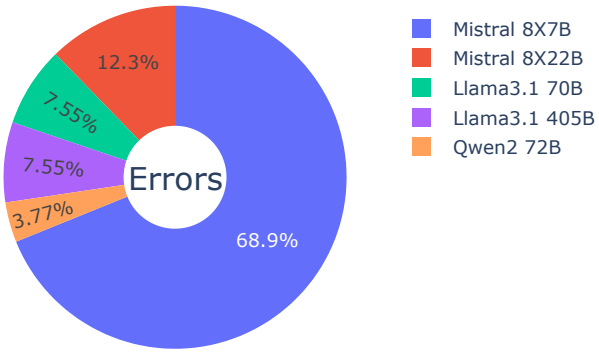

Figure 5: Error distribution across different models.