FineThera: Fine-grained Therapeutic Alliance Prediction with Large Language Models

Anonymous ACL submission

Abstract

Client perceptions of the therapeutic alliance are crucial predictors of counseling effectiveness, yet current methods of obtaining client feedback through questionnaires impose substantial burdens and are often impractical. This paper introduces a novel approach leveraging Large Language Models (LLMs) to automatically assess fine-grained dimensions of therapeutic relationships from counseling conversations. We collect 3241 real-world counseling sessions and develop a comprehensive framework, utilizing 551 of these sessions with client-rated alliance scores across core therapeutic dimensions (goal, approach, bond) for training. Through rationale-augmented fine-tuning, our model not only outperforms human counselors in aligning with client perceptions (0.507 vs. 0.279 correlation) but also provides interpretable explanations for its predictions. Analysis of model-generated insights reveals key patterns in counselor behaviors that influence alliance formation, offering actionable guidance for improving therapeutic relationships. Our work demonstrates the potential of LLMs to enhance counseling practice through automated, interpretable assessment while maintaining ethical considerations. The framework enables real-time understanding of client perspectives without additional burden, paving the way for more responsive and effective mental healthcare delivery.

1 Introduction

Mental health challenges affect over 25% of the global population (Organization et al., 2001), with online counseling emerging as an increasingly vital treatment modality (Mallen et al., 2005; Dowling and Rickwood, 2013). Research consistently shows that the therapeutic alliance - the collaborative relationship between counselor and client - is one of the strongest predictors of counseling outcomes (Martin et al., 2000; Lambert and Barley, 2001). Within this relationship, clients' perceptions are particularly crucial, with studies showing they robustly correlate more strongly with treatment success compared to counselor assessments (Horvath and Symonds, 1991; Piper et al., 1991). Moreover, discrepancies in their perceptions may compromise the effectiveness of therapeutic interventions (Horvath et al., 2011).

However, accurately understanding clients' perspectives remains a significant challenge in counseling practice. Current methods rely heavily on post-session questionnaires that burden clients and often yield inconsistent responses (Goldberg et al., 2020). This challenge is particularly acute in text-based online counseling, where counselors lack traditional nonverbal cues and must rely solely on written exchanges to gauge the therapeutic relationship (Kit et al., 2017; Békés et al., 2021).

Recent work has explored using NLP techniques to automatically assess therapeutic alliance from counseling transcripts (Martinez et al., 2019; Goldberg et al., 2020; Lin et al., 2023). However, these approaches have critical limitations: (1) they typically generate only overall alliance scores, missing crucial fine-grained dimensions of the relationship (Martinez et al., 2019; Goldberg et al., 2020), (2) they rely on black-box predictions without interpretable rationales (Ryu et al., 2021; Goldberg et al., 2020), and (3) they analyze individual conversation turns in isolation rather than capturing the full therapeutic context (Lin et al., 2023).

We present FineThera, a novel framework that leverages Large Language Models (LLMs) to predict fine-grained dimensions of client-perceived therapeutic relationships. By evaluating key aspects such as shared therapeutic goals, coordinated pathways to achieve goals, and attachment strength between counselors and clients, FineThera offers a more comprehensive and insightful assessment. Our approach uniquely combines dimension-specific alliance prediction with interpretable rationales extracted from conversations. Through

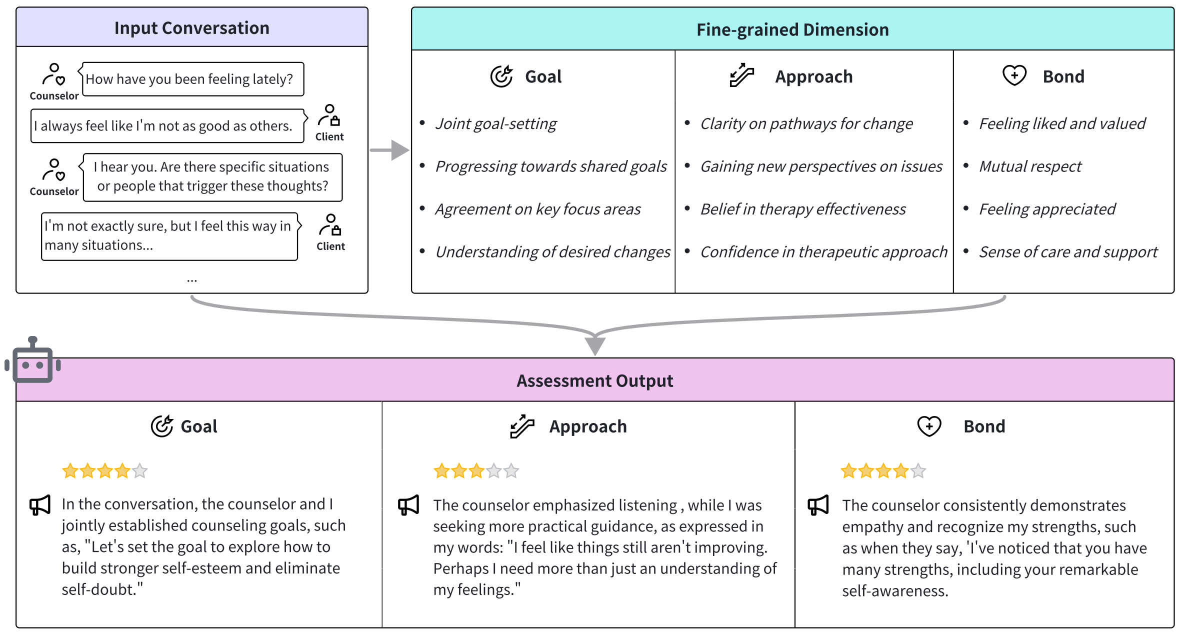

Figure 1: The LLM-based framework (FineThera) predicts fine-grained dimensions of client-perceived therapeutic relationships from textual counseling conversations. The core dimensions of the therapeutic alliance include shared therapeutic goals, coordinated problem-solving approaches, and a strong emotional bond between counselors and clients, each represented by four specific elements. LLMs predict ratings for these fine-grained dimensions, with rationale extracted directly from conversations.

rationale-augmented fine-tuning on a large-scale counseling dataset, we enable LLMs to not only predict alliance scores but also explain the specific conversational evidence supporting their assessments.

This paper makes three main contributions:

1. A novel LLM-based framework for fine-grained therapeutic alliance assessment that significantly outperforms human counselors (0.507 vs 0.279 correlation) in aligning with client perceptions across goal, approach, and bond dimensions.
2. An innovative rationale-augmented fine-tuning approach that enhances both prediction accuracy and interpretability, validated through comprehensive experiments across multiple LLM architectures.
3. An empirical demonstration of clinical utility through automated, burden-free assessment of therapeutic relationships, offering actionable insights to improve counseling practice.

The rest of this paper is organized as follows: Section 2 reviews related work in therapeutic alliance assessment and NLP applications in mental health. Section 3 introduces our framework and methodology. Section 4 presents our dataset. Section 5 details our experimental setup, results and analysis. Section 6 discusses implications. Section 7 concludes with future directions.

2 Related Work

Automatic Evaluation of Counseling using NLP. Many researchers have investigated the application of machine learning and natural language processing (NLP) techniques for the automatic evaluation of mental health counseling conversations (Calvo et al., 2017; Malgaroli et al., 2023). These studies primarily focus on analyzing individual participants' behaviors, including counselors' conversational strategies (Can et al., 2016; Gibson et al., 2016; Cao et al., 2019) and clients' reactions to interventions (Tanana et al., 2015; Li et al., 2022, 2023). There are also methods that analyze relational dynamics between counselors and clients, including their linguistic coordination in therapeutic dyads (Wadden et al., 2021; Nasir et al., 2019), their emotional convergence during counseling (Park et al., 2021; Syzdek, 2020) and therapeutic rupture (Atzil-Slonim et al., 2021) and alliance (Goldberg et al., 2020; Martinez et al., 2019). Most studies aimed at automatically predicting alliance strength rely on extracting linguistic features from conversations and applying machine learning models, limiting interpretability (Goldberg et al., 2020;

Ryu et al., 2021). These studies typically focus on predicting overall alliance scores, overlooking the fine-grained components that constitute the alliance (Martinez et al., 2019; Goldberg et al., 2020).

Our research is designed to empower LLMs to align with the fine-grained client-rated alliance based on theoretical framework, while also extracting or reasoning explanations from dialogues.

LLMs in Mental Health Conversation Analysis. With the advert of LLMs demonstrating advanced text understanding and reasoning capabilities, researchers have increasingly turned to these models for mental health-related analysis based on conversational data (Ji et al., 2023; Adhikary et al., 2024; Chiu et al., 2024). Many studies focus on leveraging LLMs to detect mental health conditions, such as anxiety, depression and suicide ideation (Lamichhane, 2023; Yang et al., 2023; Xu et al., 2024), as well as to identify the underlying causes and contributing factors of these conditions (Yang et al., 2023). Additionally, several studies explore LLMs' ability to predict the Big Five personality traits based on conversations (Yan et al., 2024; Amin et al., 2023).

Some research also explores the use of LLMs to evaluate the effectiveness of counseling conversations (Lee et al., 2024; Li et al., 2024; Wang et al., 2024). Lee et al. (2024) employ GPT models to assess the overall quality of counseling sessions as positive, neutral, or negative. While some studies, similar to ours, examine the evaluation of therapeutic relationships through dialogue, they either have yet to validate the feasibility of this method or focus on enhancing LLM capabilities through detailed prompting to align with observer-rated assessments (Li et al., 2024; Wang et al., 2024).

Different from these studies, our research aims to predict clients' self-reported therapeutic relationships, which are widely recognized as more predictive indicators of counseling outcomes. Moreover, we investigate fine-tuning techniques to develop a specialized model independent of proprietary systems, designed to be adaptable to private data.

3 Problem Definition

In this section, we introduce the definition and measurement of therapeutic alliance and outline the specific task of predicting it from dialogue text.

3.1 Dimensions of Therapeutic Alliance

The therapeutic alliance is broadly recognized as a collaborative element of the client-counselor relationship (Bordin, 1979; Ardito and Rabellino, 2011). This multifaceted concept, which integrates both cognitive and emotional interactions, is generally characterized by three crucial components: (a) mutual agreement on the goals of therapy; (b) a shared understanding that the therapeutic tasks will effectively address clients' specific concerns; and (c) the strength of the interpersonal bond between clients and counselors (Bordin, 1979).

Goal. Establishing clear counseling goals is fundamental to a successful counseling session, distinguishing it from casual conversations. Therapeutic goals involve fostering positive changes in clients' thoughts, cognition, and behaviors, facilitated by counselors' guidance and support. Both counselors and clients should collaboratively define and mutually agree on their counseling goals, ensuring their efforts are directed toward shared objectives.

Approach. Beyond setting consistent goals, reaching mutual agreement between counselors and clients on specific methods to achieve them is a critical element. Counselors typically propose tasks based on their personal styles, experience, and predispositions, but clients may find them unmanageable or unsuitable. In such instances, counselors need to provide alternative approaches to better engage their clients. Furthermore, counselors should clarify how these tasks contribute to achieving therapy goals, as this understanding is crucial for effective treatment (Horvath and Luborsky, 1993).

Bond. In addition to the cognitive aspects of the alliance that emphasize the consensus on therapy goals and tasks, the emotional attachment between counselors and clients is crucial. The bond reflects the feelings and attitudes that each party holds toward the other, fostering collaboration and trust. When clients perceive counselors' genuine care and attention, they feel secure and motivated to engage in therapy. Likewise, when both parties trust each other's abilities, a shared commitment to goals and tasks can be established.

3.2 Measurement of Therapeutic Alliance

To accurately measure clients' perceptions of the therapeutic alliance, we adopt the client version of the short revised form of the Working Alliance

Inventory (WAI), based on core alliance theory concepts (Horvath and Greenberg, 1989). The inventory includes 12 questions, with 4 questions dedicated to assessing each dimension (see Table 4 in Appendix A.2). Each question is rated on a 5-point scale: 1 = Seldom; 2 = Sometimes; 3 = Fairly Often; 4 = Very Often; 5 = Always. The reliability and validity of this inventory have been well-established across various types of psychotherapy (Hatcher and Gillaspy, 2006; Munder et al., 2010).

3.3 Task Definition

Formally, we define the task of evaluating the therapeutic relationship as follows: given the counseling conversation with each question from the measurement, predict the client's rating and extract supporting evidence from the dialogue. We use the term "client" instead of "patient" to emphasize the individual's active role and autonomy in therapy, fostering trust and collaboration.

4 Data Collection

4.1 Data Source

We gathered text-based counseling conversations between professional counselors and actual clients from an online Chinese text-based psychocounseling platform. Each session followed the standard 50-minute duration. Following each session, clients were invited to share their perspectives on the therapeutic alliance with their counselors by completing the client version of the WAI. Counselors, likewise, were required to fill out the counselor version of the same inventory, which mirrored the client version (Horvath and Greenberg, 1989). Additionally, clients were asked to complete the Outcome Rating Scale (ORS) (Miller et al., 2003) to assess the effectiveness of the counseling in terms of their physical and mental well-being, interpersonal relationships, social role functioning, and overall quality of life, with scores ranging from 0 to 100 for each aspect. Further details regarding these scales are provided in Appendix A.2.

In the end, we collected a total of 3,241 counseling sessions, with 793 including clients' self-reported working alliance scores. Of these, 569 sessions incorporated both counselors' perspectives on the alliance and clients' evaluations of the counseling outcomes. Detailed statistics are in Table 1 and Appendix A.1. We observed that only about a quarter of the sessions contained client feedback, indicating the difficulty counselors face when relying on self-reported data to evaluate clients' perspectives in real practice. This underscores the necessity for automating the prediction of clients' views on therapeutic relationships. The distribution of client ratings and average scores for each question is provided in the Appendix A, with further analysis of the findings discussed below.

	Total	Client	Counselor
# Dialogues	3,241	-	-
# Dialogues + Client WA	793	-	-
# Dialogues + All Scales	569	-	-
# Speakers	939	890	49
# Utterances	236,470	124,754	111,716
Avg. Utterances per Dialogue	72.96	38.49	34.49
Avg. Length per Utterance	29.41	26	32.46

Table 1: Statistics of the counseling conversation dataset. The balanced distribution across speakers and utterance lengths ensures comprehensive coverage of therapeutic interactions.

4.2 Data Analysis

We further validate the necessity of ensuring that counselors fully understand their clients' perspective on the working alliance.

We conducted Pearson correlation analysis between counselors and clients' ratings on each question of the working alliance and the correlation coefficient is only around 0.3 (Pleases refer to Table 5 in Appendix A.2 for detailed information) . This low correlation suggests that ***the perspectives of counselors and clients are only moderately positively correlated***, aligning with existing psychological research (Hatcher et al., 1995; Shick Tryon et al., 2007). Additionally, we examined the differences in average ratings across various dimensions and performed paired t-tests. As shown in Figure 2, our findings suggest that ***counselors may hold an overly optimistic view of their relationship with clients***, especially regarding the consensus on specific steps toward resolving the clients' mental health issues (Walfish et al., 2012; Lambert, 2013).

Furthermore, we calculated the Pearson correlations between the self-assessed working alliance scores of counselors and clients with counseling outcomes. Our analysis shows that ***clients' perceptions are a more reliable predictor of counseling success***, with a correlation nearly three times higher than that of counselors'. More details are presented in Table 6 in Appendix A.2.

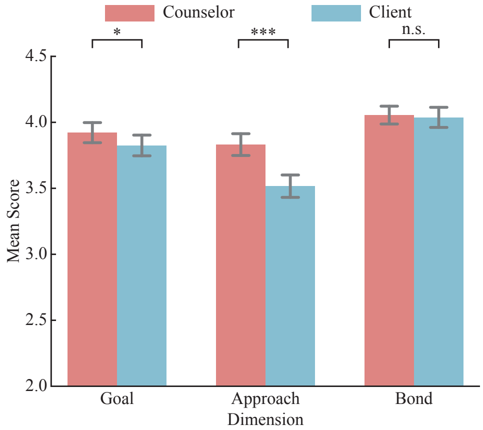

Figure 2: The histogram of the average scores and standard deviations for each dimension of the perceived alliance, for both counselors and clients respectively. *n.s.* means non-significant. $***p < 0.001$, $*p < 0.05$.

5 Automated Prediction for Client-reported Relationships

We conduct zero-shot experiments by prompting various advanced open-source and closed-source LLMs to automatically predict client-rated therapeutic relationships from text-only conversations. Moreover, to facilitate local deployment for those handling private data without relying on closed-source models, we trained specialized models using both LoRA and full-parameter fine-tuning techniques. All experiments in this work are performed using four NVIDIA A100 80G GPUs.

5.1 Data Preparation

Data Split. We randomly divided all 793 conversations with client-rated working alliance scores into a training set (70%; 551 sessions) and a validation set (30%; 242 sessions) using stratified random sampling to ensure balanced score distributions. A post hoc power analysis (Howell, 1992) was also conducted to confirm the sample size is sufficient for reliable results. Further details are provided in Appendix B.2.

Rationale-Augmented Training Data. To further enhance the interpretability of the training data, which consists solely of client-reported scores, we leverage the Qwen1.5-110B-chat model to generate underlying rationales based on the ground truth ratings provided by clients. This model has demonstrated superior performance in prompt-based approaches, as shown in the following experiments.

Training and Validation Data. For prompt-based approaches, models are applied exclusively to the validation set. For fine-tuning methods, models are trained either on the raw training set or the rationale-augmented training set, with performance evaluated on the validation set.

5.2 Prompt-based Approaches

Models We utilize several advanced LLMs to predict client-rated therapeutic relationships through prompt-based approaches. These models have been optimized to follow human instructions through instruction tuning and align with human preferences via reinforcement learning from human feedback (RLHF, (Ouyang et al., 2022)).

Closed-source LLMs. We select four top-performing, accessible open-source LLMs – Claude-3 (*Sonnet* model; Anthropic) (Anthropic, 2024), ChatGPT (*gpt-35-turbo-16k* model; OpenAI) (OpenAI, 2023a), GPT-4o-mini (*gpt-4o-mini* model; OpenAI) (OpenAI, 2024), and GPT-4 (*gpt-4-0125-preview* model; OpenAI) (OpenAI, 2023b).

Open-source LLMs. In addition, we assess 9 closed-source LLMs, including the Qwen-1.5 series (Team, 2024) with various parameter sizes from 7B to 110B (*Qwen1.5-7B-Chat*, *Qwen1.5-14B-Chat*, and *Qwen1.5-32B-Chat*, *Qwen1.5-72B-Chat* and *Qwen1.5-110B-Chat* models), the Llama-3 series models (AI@Meta, 2024) (*Meta-Llama-3-8B-Instruct* and *Meta-Llama-3-70B-Instruct* models), and GLM-4-9b (*glm-4-9b-chat* model) and Yi-1.5–34B (*Yi-1.5-34B-Chat-16K* model).

Setup We task each model to rate three times independently for every given conversation with question, and then use the average score as the final prediction. Our approach utilizes zero-shot prompting, with the temperature and nuclear sampling parameters set as 0.7 and 0.8 for all models. The template prompt is provided inAppendix B.1.

5.3 Fine-tuning Approaches

Backbone Model We select Meta-Llama-3-8B-Instruct, one of the most widely used open-source models, for fine-tuning. Its smaller scale allows for more cost-effective training, yet it has shown impressive performance in this task with prompt engineering, surpassing larger models like ChatGPT and the 14B model in the Qwen1.5 series.

Setup We utilize two types of training data: raw training data (*score only*, *SO*) and rationale-augmented training data (*score + rationale*, *SR*). We employ both full-parameter fine-tuning

Model	ICC	Kappa	Goal	Approach	Bond	Avg.
Human Counselor	-	-	0.290***	0.286***	0.261***	0.279
ChatGPT	0.458	0.297	0.220***	0.216***	0.179**	0.205
Claude-3	0.924	0.772	0.409***	0.391***	0.382***	0.394
GPT-4o-mini	0.734	0.617	0.412***	0.371***	0.426***	0.403
GPT-4	0.941	0.830	0.419***	0.481***	0.381***	0.427
Qwen1.5-7B-chat	0.500	0.392	0.148*	0.162*	0.172**	0.161
Qwen1.5-14B-chat	0.671	0.521	0.219***	0.263***	0.290***	0.257
Meta-Llama-3-8B-Instruct	0.688	0.683	0.288***	0.302***	0.334***	0.308
Qwen1.5-32B-chat	0.952	0.869	0.326***	0.307***	0.395***	0.343
Qwen1.5-72B-chat	0.711	0.550	0.247***	0.392***	0.414***	0.351
glm-4-9B-chat	0.672	0.459	0.367***	0.368***	0.327***	0.354
Yi-1.5-34B-Chat-16K	0.715	0.663	0.389***	0.341***	0.390***	0.373
Meta-Llama-3-70B-Instruct	0.945	0.854	0.424***	0.378***	0.445***	0.416
Qwen1.5-110B-chat	0.953	0.860	0.436***	0.431***	0.452***	0.440
Llama-3-8B-Instruct-LoRA-SO	0.681	0.644	0.343***	0.300***	0.380***	0.341
Llama-3-8B-Instruct-LoRA-SR	0.500	0.421	0.354***	0.321***	0.360***	0.345
Llama-3-8B-Instruct-full-SO	0.562	0.409	0.414***	0.455***	0.504***	0.458
Llama-3-8B-Instruct-full-SR	0.671	0.594	**0.505*****	**0.483*****	**0.533*****	**0.507**

Table 2: Performance comparison across models and dimensions. The Pearson correlation results between human counselors and all models with clients' self-reported assessments on the working alliance dimensions across different experimental settings, as well as intra-class correlation coefficient and Fleiss' Kappa of each model. *SO* and *SR* indicate *Score Only* and *Score + Rationale* respectively. The last column represents the average of the correlation coefficients across all three dimensions. The proposed rationale-augmented fine-tuning consistently outperforms baselines and human counselors. $^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$.

and a parameter-efficient fine-tuning method — LoRA (Hu et al., 2022) — to train our models.

In all experimental setups, we configured the training to run for 3 epochs with a learning rate of 5e-7. During inference, temperature and nucleus sampling parameters were configured to 0.7 and 0.8, respectively, for all models. Consistent with prompt-based approaches, each fine-tuned model is also tasked three times. Additional details about the experiment setup are provided in Appendix B.3.

5.4 Experimental Results

Table 2 presents models' consistency and Pearson correlation results for different alliance dimensions compared to clients' self-reported ratings across all experimental settings. Detailed results for each question can be found in Appendix B.4.

Most LLMs exhibit moderate or higher self-consistency, ensuring reliable evaluations. We use both Intra-class Correlation Coefficient (ICC) (Koo and Li, 2016) and Fleiss' Kappa (Fleiss, 1971) to calculate the consistency of the model's predictions across three trials in each setting. As demonstrated in Table 2, most models achieve moderate to high self-consistency (ICC values: $0.6 \sim 0.75$, or Fleiss' Kappa: $0.4 \sim 0.6$), meeting the necessary criteria for reliability. In this study, we focus solely on ensuring the model's internal consistency meets this threshold, and then evaluate models' performance based on correlation with the ground truth rather than inter-rater agreement.

LLMs can align better with clients' perceptions than the counselors themselves. As shown in Table 2, 14 out of the 17 models – excluding ChatGPT, Qwen1.5-7B-chat, and Qwen1.5-14B-chat – exhibit higher overall correlations with clients' perceived strength of the therapeutic alliance, compared to those with counselors' perceived ratings. All fine-tuned models surpass the alignment between human counselors and clients across all three sub-dimensions. Notably, the Llama-8B model, fine-tuned with full parameters on training data that incorporated both client ratings and rationales, significantly outperforms counselor evaluations by 81.72% in average performance. In the bond dimension, where counselors' alignment with clients is weakest, this model achieves nearly double the correlation. This highlights its superior ability to detect affective signals of liking, trust, and respect in text-based conversations. Thus, employing LLMs to predict clients' perceived therapeutic alliance can offer counselors deeperi insights into clients' perspectives.

Fine-tuning significantly enhances LLMs' ability to perceive client-rated therapeutic alliance. Continued training of the Meta-Llama-3-8B-Instruct model on task-specific data, whether through LoRA or full-parameter fine-tuning, enhances its performance across all dimensions. In the prompt-based setting, the Qwen1.5-110B-chat model shows the strongest correlation, slightly outperforming GPT-4. Nevertheless, full-parameter fine-tuning of the 8B model, even with just over 500 data points, results in a 12.06% to 17.92% improvement across all three dimensions compared to the Qwen1.5-110B-chat model.

Full-parameter fine-tuning is more effective than LoRA in enhancing LLMs' predictive ability. In this task, full-parameter fine-tuning consistently outperforms LoRA under the same training conditions. This is likely because assessing the therapeutic alliance in text-based dialogues is complex and necessitates advanced text understanding and reasoning. Full-parameter fine-tuning allows the model to fully leverage its parameter capacity, minimizing information loss and adapting effectively.

Incorporating rationales with ratings in the training data can improve the predictive performance of trained LLMs. By integrating explanations inferred from clients' self-reported scores, as generated by the Qwen1.5-110B-chat model, into the training data, the performance of the Llama-8B-Instruct model is further enhanced under both LoRA and full-parameter fine-tuning compared to training with scores alone. Notably, this improvement is more pronounced under full-parameter fine-tuning, yielding an average increase of approximately 10.70%. This suggests that including underlying rationales during training enable the model better understand the evidence in the dialogue text associated with the ratings.

5.5 Case Study

To better understand the challenges faced by the best-performing model (i.e., the Llama-3-8B-Instruct-full-SR model), we conducted a case study. We identified instances where the model's predictions significantly deviated from the client's self-reported ratings (examples are shown in Table 3) and analyzed the model's explanations to uncover the causes of these discrepancies.

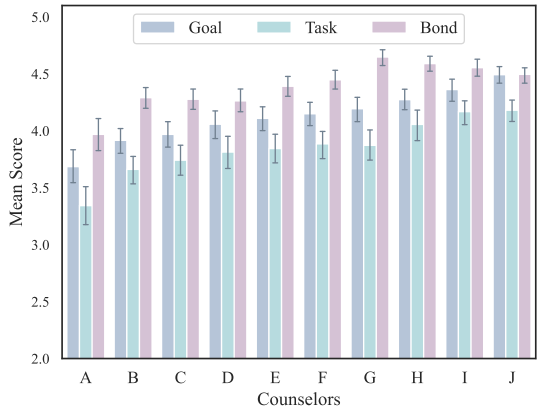

Figure 3: The average scores with standard deviations for each alliance dimension across all counseling dialogues for the top 10 counselors with the most sessions.

We found that the model may overlooks the client's actions, focusing solely on their verbal expressions. For example, in the first case, the client was late and left early, suggesting a lack of respect and weak emotional connection. However, the model interpreted the client's language as non-offensive and assumed that the client was respectful. In addition, the model may overemphasizes some positive or negative statements in counseling, while ignoring the overall interaction. In the second case, the client mentioned the positive effects of previous counseling but later gave brief responses and didn't engage in the counselor's activities. Yet, the model focused on the client's affirming statements and overlooked the interaction details.

6 Additional Insights based on LLM Predictions

We utilize the best-performing model (i.e., the *Llama-3-8B-Instruct-full-SR* model) to further explore the therapeutic alliance in text-based psychological counseling. The model is employed to predict client-perceived alliance in the remaining 2,448 sessions lacking clients' self-reports. Our analysis examines variations in counselors' abilities to establish alliances with clients, and how conversational behaviors affect the alliance.

Counselors' Varying Abilities in Establishing Relationships. We explore the differences in the strength of the therapeutic alliance across various dimensions that counselors establish with their clients. Focusing on the top 10 counselors with

Conversation	Evaluation Results
Counselor: Hello, are you online? *10 minutes later...* Client: Sorry, I fell asleep. Counselor: It's okay, let's begin now. Client: Thanks for waiting for me. *Counseling in progress...* Counselor: Do you have any plans for progress? Counselor: Are you still online? Hello?	**Dimension**: Bond **Client**: 1.25 **Our Model**: 4.25 **Explanations**: The counselor patiently waits for the client's responses, even when the client falls asleep, without showing any impatience. The client also responds in a friendly manner, such as saying "Sorry" and "Thanks". These interactions reflect mutual respect and understanding...
Counseling in progress... Client: I've been doing counseling for a while, and slowly I've started to shift my focus. I'm feeling a bit better now. *Counseling in progress...* Counselor: You don't want to have a negative impact on her? Client: Yes. Counselor: You don't want to put her in a diff-icult position? Client: I guess so. *Counseling in progress...*	**Dimension**: Task **Client**: 1.0 **Our Model**: 4.25 **Explanations**: The client directly stated that counseling has helped them clarify the areas that need change, and feel better by shifting focus.

Table 3: Example cases where the model's predictions significantly deviate from the client's self-reported scores.

the most sessions (each with at least 134 sessions), we calculate the average alliance scores across all dialogues and conduct t-tests (Kim, 2015) to assess differences in their relational effectiveness. As shown in Figure 3, nearly all counselors score highest in the bond dimension, followed by the goal and task dimensions, highlighting the fundamental of emotional connection in driving counseling progress (Horvath and Greenberg, 1989).

Even counselors with strong overall performance demonstrate specific strengths in particular dimensions. For example, Counselor G excels in fostering emotional bonds, outperforming Counselor J, while Counselor J shows greater skill in aligning goals and steps with clients. Since these three dimensions should ideally develop simultaneously (Bordin, 1979), counselors can learn from others' strengths to enhance their overall approach.

Actionable Insights for Counselors. As illustrated in Figure 3, some counselors, such as Counselor A, struggle more in building a strong therapeutic alliance compared to others, like Counselor J. To better understand the behaviors of counselors affecting relationship-building, we perform a content analysis of the explanations generated by our top-performing model. We randomly analyze around 100 explanations for counselors with poor performance (scores ≤ 2) and 100 for those with outstanding performance (scores ≥ 4) for each dimension. Examples of the generated explanations can be found in Appendix 12.

Our findings reveal that counselors who respond passively without offering concrete guidance may seem directionless. Those who use technical psychological terms rigidly, without adapting to clients' specific goals, are perceived as irrelevant. Additionally, counselors who repeatedly question clients without pausing may leave them feeling rushed and unheard. These insights highlight the importance of consistently expressing empathy, balancing guidance and support, monitoring clients' behaviors and making timely adjustments to meeting individual needs.

7 Conclusion

We developed a dataset and LLM-based approaches to assess client-rated therapeutic alliance in online text-based counseling. Our findings demonstrate that incorporating supportive evidence with ratings into the training data to fully fine-tune a smaller LLM model results in performance that exceeds that of much larger models and significantly surpasses human counselors. Additionally, we highlight the diverse abilities of counselors in forming client relationships, identify key behaviors influencing the establishment of the alliance, and offer actionable insights for improvement. Data, code, and models will be released upon paper acceptance.

8 Limitations

In this work, we focus on leveraging LLMs to assess the therapeutic alliance in text-based counseling conversations. Text-based counseling is a widely used and accessible form of therapy, Text-based counseling is a widely used and accessible form of therapy, and it is particularly well-suited for LLMs, which can precisely extract textual cues to evaluate the therapeutic alliance. Our approach can easily be adapted to face-to-face, video, or audio-based counseling formats by converting audio or voice recordings into text transcripts. In the future, a promising direction for understanding the therapeutic alliance in other counseling formats involves integrating multimodal models that combine facial expressions and vocal features to analyze video or audio counseling sessions.

Our data is sourced from a single counseling platform in China. While the demographic diversity of both clients and counselors in these conversations (see Appendix A.1) contributes to a degree of generalizability, future research can benefit from expanding the data collection to multiple counseling platforms from different regions. This would allow for a more comprehensive validation of the generalizability of our findings.

In addition, there is room for future improvement on the experimental aspects of this study. We only selected the LLaMA-8B as the backbone model for fine-tuning, based on its impressive performance in zero-shot prompt-based condition as a smaller model. While incorporating underlying rationales with ratings into the training data has notably enhanced the performance of the LLaMA-3-8B-Instruct model compared to both the base model and human counselors, there remains considerable room for improvement in terms of overall correlation. To further improve the fine-tuned model's performance on this task, we will explore additional models, parameters, and training approaches, as well as continue collecting more psychological counseling conversations with client self-ratings. Since the model's performance does not yet fully align with clients' self-ratings, analyses based on the model's predictions may still have some limitations. Nevertheless, these experiments and analyses offer valuable insights for the interdisciplinary intersection of NLP and psychology.

9 Ethics Statement

Data Privacy. This study has been approved by the Institutional Review Board. All counselors and clients provided informed consent to participate, with clear communication that conversations collected on the platform would be used for scientific research purposes and might be shared with third parties for these purposes. Following the collection of these conversations, we meticulously de-identified and anonymized the data to ensure the highest level of privacy protection for both clients and counselors. Furthermore, all collaborators involved in this research are required to formally commit to data confidentiality agreements and adhere to rigorous ethical standards.

Data Release. To encourage interdisciplinary research at the intersection of NLP and psychology, we plan to release this dataset to qualified researchers upon the acceptance of this article. Those interested in accessing the data will be evaluated based on their qualifications. We require applicants to provide a valid ID, a justification for their data request, proof of full-time employment at a non-profit academic or research institution with Institutional Review Board (IRB) approval, confirmation of their role as a full-time principal investigator, and approval from the institution's Office of Research or equivalent office. Additionally, applicants must sign a Data Non-Disclosure Agreement, committing not to share the data with any third party.

LLM-based Predictions. This study aims to introduce an automated approach for predicting client-rated therapeutic alliance with their counselors in the context of online text-based counseling. We advocate for using LLM-based predictions as an additional tool to help counselors gain a deeper understanding of their clients. Given the current limitations in LLM performance, it is crucial to exercise caution when applying these predictions in practice. Furthermore, societal acceptance must be considered to mitigate potential misuse of technology and address ethical concerns associated with LLM-generated results.

References

Prottay Kumar Adhikary, Aseem Srivastava, Shivani Kumar, Salam Michael Singh, Puneet Manuja, Jini K Gopinath, Vijay Krishnan, Swati Kedia, Koushik Sinha Deb, and Tanmoy Chakraborty. 2024.

Exploring the efficacy of large language models in summarizing mental health counseling sessions: A benchmark study. *Preprint*, arXiv:2402.19052.

AI@Meta. 2024. Llama 3 model card.

Mostafa M Amin, Erik Cambria, and Björn W Schuller. 2023. Will affective computing emerge from foundation models and general artificial intelligence? a first evaluation of chatgpt. *IEEE Intelligent Systems*, 38(2):15–23.

Anthropic. 2024. The claude 3 model family: Opus, sonnet, haiku. `https://api.semanticscholar.org/CorpusID:268232499`. [Accessed 16-04-2024].

Rita B Ardito and Daniela Rabellino. 2011. Therapeutic alliance and outcome of psychotherapy: historical excursus, measurements, and prospects for research. *Frontiers in psychology*, 2:270.

Dana Atzil-Slonim, Daniel Juravski, Eran Bar-Kalifa, Eva Gilboa-Schechtman, Rivka Tuval-Mashiach, Natalie Shapira, and Yoav Goldberg. 2021. Using topic models to identify clients' functioning levels and alliance ruptures in psychotherapy. *Psychotherapy*, 58(2):324.

Vera Békés, Katie Aafjes-van Doorn, Xiaochen Luo, Tracy A Prout, and Leon Hoffman. 2021. Psychotherapists' challenges with online therapy during covid-19: Concerns about connectedness predict therapists' negative view of online therapy and its perceived efficacy over time. *Frontiers in psychology*, 12:705699.

Edward S Bordin. 1979. The generalizability of the psychoanalytic concept of the working alliance. *Psychotherapy: Theory, research & practice*, 16(3):252.

Rafael A Calvo, David N Milne, M Sazzad Hussain, and Helen Christensen. 2017. Natural language processing in mental health applications using non-clinical texts. *Natural Language Engineering*, 23(5):649–685.

Doğan Can, Rebeca A Marín, Panayiotis G Georgiou, Zac E Imel, David C Atkins, and Shrikanth S Narayanan. 2016. "it sounds like...": A natural language processing approach to detecting counselor reflections in motivational interviewing. *Journal of counseling psychology*, 63(3):343.

Jie Cao, Michael Tanana, Zac Imel, Eric Poitras, David Atkins, and Vivek Srikumar. 2019. Observing dialogue in therapy: Categorizing and forecasting behavioral codes. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5599–5611, Florence, Italy. Association for Computational Linguistics.

Yu Ying Chiu, Ashish Sharma, Inna Wanyin Lin, and Tim Althoff. 2024. A computational framework for behavioral assessment of llm therapists. *Preprint*, arXiv:2401.00820.

Mitchell Dowling and Debra Rickwood. 2013. Online counseling and therapy for mental health problems: A systematic review of individual synchronous interventions using chat. *Journal of Technology in Human Services*, 31(1):1–21.

Franz Faul, Edgar Erdfelder, Axel Buchner, and Albert-Georg Lang. 2009. Statistical power analyses using g* power 3.1: Tests for correlation and regression analyses. *Behavior research methods*, 41(4):1149–1160.

Franz Faul, Edgar Erdfelder, Albert-Georg Lang, and Axel Buchner. 2007. G* power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior research methods*, 39(2):175–191.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

James Gibson, Dogan Can, Bo Xiao, Zac E Imel, David C Atkins, Panayiotis Georgiou, and Shrikanth Narayanan. 2016. A deep learning approach to modeling empathy in addiction counseling. *Commitment*, 111:21.

Simon B Goldberg, Nikolaos Flemotomos, Victor R Martinez, Michael J Tanana, Patty B Kuo, Brian T Pace, Jennifer L Villatte, Panayiotis G Georgiou, Jake Van Epps, Zac E Imel, et al. 2020. Machine learning and natural language processing in psychotherapy research: Alliance as example use case. *Journal of counseling psychology*, 67(4):438.

Robert L Hatcher, Alex Barends, James Hansell, and M Janice Gutfreund. 1995. Patients' and therapists' shared and unique views of the therapeutic alliance: an investigation using confirmatory factor analysis in a nested design. *Journal of Consulting and Clinical Psychology*, 63(4):636.

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.

Adam O Horvath, AC Del Re, Christoph Flückiger, and Dianne Symonds. 2011. Alliance in individual psychotherapy. *Psychotherapy*, 48(1):9.

Adam O Horvath and Leslie S Greenberg. 1989. Development and validation of the working alliance inventory. *Journal of counseling psychology*, 36(2):223.

Adam O Horvath and Lester Luborsky. 1993. The role of the therapeutic alliance in psychotherapy. *Journal of consulting and clinical psychology*, 61(4):561.

Adam O Horvath and B Dianne Symonds. 1991. Relation between working alliance and outcome in psychotherapy: A meta-analysis. *Journal of counseling psychology*, 38(2):139.

David C Howell. 1992. *Statistical methods for psychology*. PWS-Kent Publishing Co.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Shaoxiong Ji, Tianlin Zhang, Kailai Yang, Sophia Ananiadou, and Erik Cambria. 2023. Rethinking large language models in mental health applications. *arXiv preprint arXiv:2311.11267*.

Tae Kyun Kim. 2015. T test as a parametric statistic. *Korean journal of anesthesiology*, 68(6):540–546.

Phey Ling Kit, Chua Tee Teo, Meilinda Tan, and Yuhyun Park. 2017. Singaporean counsellors' online counselling experiences with children: An exploratory qualitative study. *Journal of Asia Pacific Counseling*, 7(2):141–168.

Terry K Koo and Mae Y Li. 2016. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of chiropractic medicine*, 15(2):155–163.

Michael J Lambert. 2013. Outcome in psychotherapy: the past and important advances.

Michael J Lambert and Dean E Barley. 2001. Research summary on the therapeutic relationship and psychotherapy outcome. *Psychotherapy: Theory, research, practice, training*, 38(4):357.

Bishal Lamichhane. 2023. Evaluation of chatgpt for nlp-based mental health applications. *arXiv preprint arXiv:2303.15727*.

Younghun Lee, Dan Goldwasser, and Laura Schwab Reese. 2024. Towards understanding counseling conversations: Domain knowledge and large language models. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 2032–2047, St. Julian's, Malta. Association for Computational Linguistics.

Anqi Li, Yu Lu, Nirui Song, Shuai Zhang, Lizhi Ma, and Zhenzhong Lan. 2024. Automatic evaluation for mental health counseling using llms. *arXiv preprint arXiv:2402.11958*.

Anqi Li, Jingsong Ma, Lizhi Ma, Pengfei Fang, Hongliang He, and Zhenzhong Lan. 2022. Towards automated real-time evaluation in text-based counseling. *arXiv preprint arXiv:2203.03442*.

Anqi Li, Lizhi Ma, Yaling Mei, Hongliang He, Shuai Zhang, Huachuan Qiu, and Zhenzhong Lan. 2023. Understanding client reactions in online mental health counseling. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10358–10376, Toronto, Canada. Association for Computational Linguistics.

Baihan Lin, Guillermo Cecchi, and Djallel Bouneffouf. 2023. Deep annotation of therapeutic working alliance in psychotherapy. In *International workshop on health intelligence*, pages 193–207. Springer.

Matteo Malgaroli, Thomas D Hull, James M Zech, and Tim Althoff. 2023. Natural language processing for mental health interventions: a systematic review and research framework. *Translational Psychiatry*, 13(1):309.

Michael J Mallen, David L Vogel, Aaron B Rochlen, and Susan X Day. 2005. Online counseling: Reviewing the literature from a counseling psychology framework. *The Counseling Psychologist*, 33(6):819–871.

Daniel J Martin, John P Garske, and M Katherine Davis. 2000. Relation of the therapeutic alliance with outcome and other variables: a meta-analytic review. *Journal of consulting and clinical psychology*, 68(3):438.

Victor R Martinez, Nikolaos Flemotomos, Victor Ardulov, Krishna Somandepalli, Simon B Goldberg, Zac E Imel, David C Atkins, and Shrikanth Narayanan. 2019. Identifying therapist and client personae for therapeutic alliance estimation. In *Interspeech*, volume 2019, page 1901. NIH Public Access.

Scott D Miller, BL Duncan, Jeb Brown, JA Sparks, and DA Claud. 2003. The outcome rating scale: A preliminary study of the reliability, validity, and feasibility of a brief visual analog measure. *Journal of brief Therapy*, 2(2):91–100.

Thomas Munder, Fabian Wilmers, Rainer Leonhart, Hans Wolfgang Linster, and Jürgen Barth. 2010. Working alliance inventory-short revised (wai-sr): psychometric properties in outpatients and inpatients. *Clinical Psychology & Psychotherapy: An International Journal of Theory & Practice*, 17(3):231–239.

Md Nasir, Sandeep Nallan Chakravarthula, Brian Baucom, David C Atkins, Panayiotis Georgiou, and Shrikanth Narayanan. 2019. Modeling interpersonal linguistic coordination in conversations using word mover's distance. In *Interspeech*, volume 2019, page 1423. NIH Public Access.

OpenAI. 2023a. Chatgpt.

OpenAI. 2023b. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

OpenAI. 2024. Gpt-4o mini: advancing cost-efficient intelligence.

World Health Organization et al. 2001. The world health report 2001: Mental health: new understanding, new hope.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke E. Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Francis Christiano, Jan Leike, and Ryan J. Lowe. 2022. Training language models to follow instructions with human feedback. *ArXiv*, abs/2203.02155.

Jihyun Park, Abhishek Jindal, Patty Kuo, Michael Tanana, Jennifer Elston Lafata, Ming Tai-Seale, David C Atkins, Zac E Imel, and Padhraic Smyth. 2021. Automated rating of patient and physician emotion in primary care visits. *Patient education and counseling*, 104(8):2098–2105.

William E Piper, Hassan FA Azim, Anthony S Joyce, and Mary McCallum. 1991. Transference interpretations, therapeutic alliance, and outcome in short-term individual psychotherapy. *Archives of general Psychiatry*, 48(10):946–953.

Jihan Ryu, Stephen Heisig, Caroline McLaughlin, Rebeccah Bortz, Michael Katz, and Xiaosi Gu. 2021. A natural language processing approach to modelling treatment alliance in psychotherapy transcripts. *BJPsych Open*, 7(S1):S48–S48.

Georgiana Shick Tryon, Sasha Collins Blackwell, and Elizabeth Felleman Hammel. 2007. A meta-analytic examination of client–therapist perspectives of the working alliance. *Psychotherapy research*, 17(6):629–642.

Brian M Syzdek. 2020. Client and therapist psychotherapy sentiment interaction throughout therapy. *Psychological Studies*, 65(4):520–530.

Michael Tanana, Kevin Hallgren, Zac Imel, David Atkins, Padhraic Smyth, and Vivek Srikumar. 2015. Recursive neural networks for coding therapist and patient behavior in motivational interviewing. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 71–79, Denver, Colorado. Association for Computational Linguistics.

Qwen Team. 2024. Introducing qwen1.5.

Georgiana Shick Tryon, Sasha Collins Blackwell, and Elizabeth Felleman Hammel. 2008. The magnitude of client and therapist working alliance ratings. *Psychotherapy: Theory, Research, Practice, Training*, 45(4):546.

David Wadden, Tal August, Qisheng Li, and Tim Althoff. 2021. The effect of moderation on online mental health conversations. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 15, pages 751–763.

Steven Walfish, Brian McAlister, Paul O'Donnell, and Michael J Lambert. 2012. An investigation of self-assessment bias in mental health providers. *Psychological reports*, 110(2):639–644.

Jiashuo Wang, Yang Xiao, Yanran Li, Changhe Song, Chunpu Xu, Chenhao Tan, and Wenjie Li. 2024. Towards a client-centered assessment of llm therapists by client simulation. *arXiv preprint arXiv:2406.12266*.

Xuhai Xu, Bingsheng Yao, Yuanzhe Dong, Saadia Gabriel, Hong Yu, James Hendler, Marzyeh Ghassemi, Anind K. Dey, and Dakuo Wang. 2024. Mental-llm: Leveraging large language models for mental health prediction via online text data. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 8(1).

Yang Yan, Lizhi Ma, Anqi Li, Jingsong Ma, and Zhenzhong Lan. 2024. Predicting the big five personality traits in chinese counselling dialogues using large language models. *Preprint*, arXiv:2406.17287.

Kailai Yang, Shaoxiong Ji, Tianlin Zhang, Qianqian Xie, Ziyan Kuang, and Sophia Ananiadou. 2023. Towards interpretable mental health analysis with large language models. *Preprint*, arXiv:2304.03347.

A Data Analysis

A.1 Details about Collected Counseling Conversation

The psychological counseling data we have collected comes from an online, text-based platform in China that provides free, accessible psychological support to individuals in need. All participants gave informed consent when using the platform, acknowledging that their conversations and survey responses would be anonymized and utilized for research purposes.

The clients represent a broad and diverse demographic, encompassing all genders and spanning a wide age range, from adolescents to middle-aged adults. They hail from various geographical locations, with residences extending from urban centers to rural towns and villages. Their educational backgrounds are equally varied, ranging from high school to doctoral degrees, and their relationship statuses include single, in a relationship, married, and divorced. The issues they discuss cover a wide spectrum of concerns, including academic and career challenges, interpersonal relationships and social connections, self-exploration and personal growth, family dynamics and parenting struggles, intimate relationships and marriage, existential anxiety, the search for the meaning of life, and many others.

The counselors bring a rich diversity of experience to the platform, with expertise ranging from less than three years for emerging counselors to over ten years for seasoned professionals. Their therapeutic skills cover a wide array of approaches, including Cognitive Behavioral Therapy (CBT), humanistic therapy, narrative therapy, and more, reflecting a comprehensive and flexible approach to counseling.

A.2 Different Perspectives of Counselors and Clients

The items of the client and counselor version of the revised short form of Working Alliance Inventory are shown in Table 4.

Table 5 shows the Pearson correlation coefficients between counselors and clients on each question and dimension of working alliance.

A.3 Relationship between Working Alliance Ratings and Counseling Outcomes

The Outcome Rating Scale (ORS) is designed to measure changes in clients' life functioning following psychological interventions (Miller et al., 2003). In our study, clients complete the ORS before each counseling session to assess their condition after the previous session, offering insights into the effectiveness of the prior counseling. Clients evaluate their overall quality of life over the past week across four key areas: (1) Individual Physical and Mental Well-being, (2) Interpersonal Relationships (Family or Intimate Relationships), (3) Social Life (Work, School, Friends), and (4) Overall Condition. Each aspect is rated on a scale from 0 to 100, with 0 representing the lowest point and 100 the highest, where higher scores indicate more favorable conditions.

Table 6 presents the Pearson correlation coefficients between counselors' and clients' reported working alliance scores and clients' self-reported counseling outcomes respectively.

A.4 Client Ratings and Question Scores Distribution

Figure 4 illustrates the distribution of client ratings and average scores for each question.

Overall, the score distribution exhibits a typical negative skew commonly observed in alliance assessments (Tryon et al., 2008; Goldberg et al., 2020). The average scores for the *Goal*, *Approach* and *Bond* dimensions all surpass 3.5, suggesting that a relatively robust therapeutic relationship can be established between counselors and clients in online text-based psychological counseling. The *Approach* dimension, however, received the lowest average score, primarily limited by Questions 5 and 6. This indicates that clients remain uncertain about the specific actions needed to pursue their goals with counselors. Conversely, the *Bond* dimension achieved the highest average score, exceeding 4, reflecting that clients frequently feel understood, cared for, and supported by their counselor, especially in terms of respect.

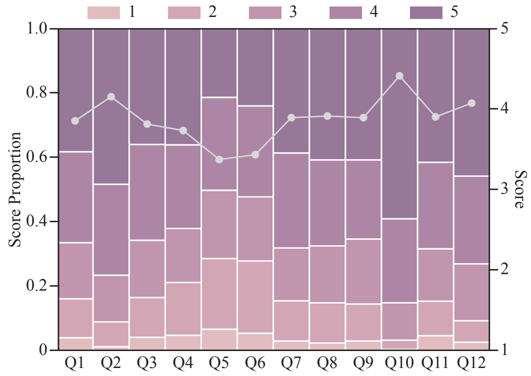

Figure 4: The stacked histogram of the distribution of client ratings for each question, as well as the line plot of the average rating across all sessions for each question. The left y-axis indicates the proportion of scores, while the right y-axis shows the scores.

Here is the psychological counseling dialogue between you as a client and your counselor.

Counselor: Hi, what would you like to talk about today?
Client: Recently, I often use sleep to escape from facing life.
Counselor: I see. What do you feel you're trying to escape from?
Client: Things I can't control, but that are still my own.
...

Before the end of this psychological counseling session, please complete the following multiple-choice question based on the conversation and your own situation. Provide the corresponding reasons with direct quotes from the dialogue between you and the counselor. (Output should be two lines, with each line indicating the choice and reason respectively):

Question: *I feel that the things I do in therapy will help me to accomplish the changes that I want.*
Options: 1. Seldom; 2. Sometimes; 3. Fairly Often; 4. Very often; 5. Always

Figure 5: The template prompt for instructing LLMs to predict clients' perceived working alliance, using an example conversation and questionnaire question (displayed in *cadetblue italic* text).

B Automatic Prediction

B.1 Template Prompt

The template prompt used to instruct LLMs to predict clients' perceived working alliance is shown in Figure 5.

B.2 Data Preparation

Table 7 presents the score distributions for the training set and the validation set.

To ensure the reliability of our results, we conducted a post hoc power analysis using

Dimension	ID	Client	Counselor
Goal	Q1	The therapist and I collaborate on setting goals for my therapy.	The client and I collaborate on setting goals for my therapy.
	Q2	The therapist and I are working towards mutually agreed upon goals.	The client and I are working towards mutually agreed upon goals.
	Q3	The therapist and I agree on what is important for me to work on.	The client and I agree on what is important for the client to work on.
	Q4	The therapist and I have established a good understanding of the kind of changes that would be good for me.	The client and I have established a good understanding of the kind of changes that would be good for the client.
Approach	Q5	As a result of this session, I am clearer as to how I might be able to change.	As a result of this session, the client is clearer as to how he/she might be able to change.
	Q6	What I am doing in therapy gives me new ways of looking at my problem.	What the client is doing in therapy gives he/she new ways of looking at his/her problem.
	Q7	I feel that the things I do in therapy will help me to accomplish the changes that I want.	I feel that the things the client do in therapy will help he/she to accomplish the changes that he/she wants.
	Q8	I believe the way we are working with my problem is correct.	I believe the way we are working with the client's problem is correct.
Bond	Q9	I believe the therapist likes me.	I believe the client likes me.
	Q10	The therapist and I respect each other.	The client and I respect each other.
	Q11	I feel that the therapist appreciates me.	I appreciates the client.
	Q12	I feel the therapist cares about me even when I do things that he/she does not approve of.	I cares about the client even when the client do things that I do not approve of.

Table 4: Core dimensions of the therapeutic alliance, with specific questions from the counselor and client versions of the Working Alliance Inventory (WAI). Each dimension captures distinct aspects of counselor-client relationship.

Question/Dimension	Correlation
Q1	0.270***
Q2	0.211***
Q3	0.235***
Q4	0.213***
Goal	***0.290******
Q5	0.244***
Q6	0.290***
Q7	0.224***
Q8	0.204***
Approach	***0.286******
Q9	0.232***
Q10	0.211***
Q11	0.204***
Q12	0.184***
Bond	***0.261******
Avg.	***0.279***

Table 5: The Pearson Correlation coefficients between counselors- and clients-reported scores on each question and dimension of the working alliance. $*p < 0.05$, $**p < 0.01$, $***p < 0.001$.

G*Power (Faul et al., 2007, 2009) after the experiment. The analysis was based on the "Correlation: Bivariate normal model" test, with an effect size of 0.5 (corresponding to the correlation between the predictions of the best-performing model and client-reported ratings) and a significance level α of 0.05. The resulting power value was nearly 1.0, confirming that the sample size of 242 in the validation set is more than sufficient to draw robust and reliable conclusions.

B.3 Experimental Settings

Table 8 shows the key hyperparameters and corresponding values used in our fine-tuning experiments.

B.4 Experimental Results

Table 9 and Table 10 show the ICC and Fleiss' Kappa values for predicting client-rated alliance across different questions and dimensions, based on various experimental settings.

Table 11 presents the Pearson correlation results between all models and clients' self-reported assessments on the alliance questions.

C Additional LLM-based Insights

C.1 Example Explanations

Table 12 presents some example explanations generated by our best-performing model.

C.2 Counselors' Varying Abilities in Establishing556 Relationships.

Figure 6 shows the heatmap results of t-tests on the working alliance scores of counselors across all counseling sessions with their clients.

	Phy. & Men.		Relationship		Social Life		Overall	
	Client	**Counselor**	**Client**	**Counselor**	**Client**	**Counselor**	**Client**	**Counselor**
Goal	0.38***	0.13**	0.39***	0.16***	0.43***	0.16***	0.40***	0.16***
Approach	0.30***	0.12**	0.35***	0.14***	0.37***	0.14***	0.34***	0.14***
Bond	0.48***	0.11**	0.46***	0.14**	0.51***	0.15***	0.48***	0.13**

Table 6: The Pearson Correlation coefficients between clients' and counselors' reported scores on the working alliance and clients' self-reported counseling outcomes.

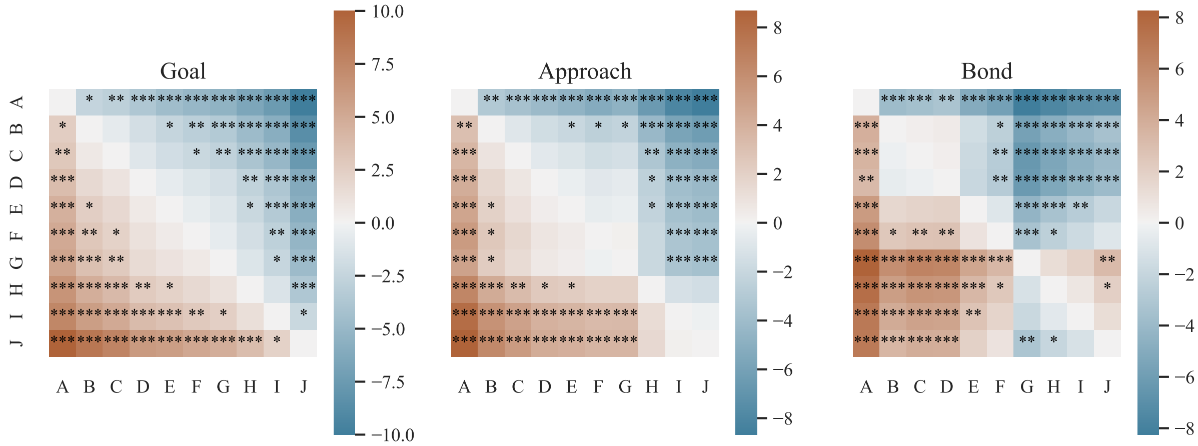

Figure 6: The heatmap results of t-tests on the working alliance scores of counselors across all counseling sessions with their clients, where each element in the heatmap represents the t value with significance. $^*p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

Question	**Data Set**	**Score** 1	2	3	4	5	**Sum**
Q1	Train	22	65	98	162	204	551
	Test	9	31	40	62	100	242
Q2	Train	7	39	79	163	263	551
	Test	2	22	36	61	121	242
Q3	Train	20	61	105	172	193	551
	Test	12	37	36	64	93	242
Q4	Train	22	82	99	152	196	551
	Test	15	48	34	54	91	242
Q5	Train	37	112	127	171	104	551
	Test	15	62	41	58	66	242
Q6	Train	29	123	110	166	123	551
	Test	13	55	48	58	68	242
Q7	Train	15	64	91	173	208	551
	Test	8	35	39	61	99	242
Q8	Train	10	61	108	155	217	551
	Test	8	38	32	57	107	242
Q9	Train	16	60	120	133	222	551
	Test	7	31	40	62	102	242
Q10	Train	1	15	64	149	322	551
	Test	1	8	28	58	147	242
Q11	Train	24	53	94	157	223	551
	Test	12	32	35	56	107	242
Q12	Train	10	34	105	151	251	551
	Test	10	19	35	65	113	242

Table 7: Score distributions for each question in the training and validation sets.

Hyperparameters	**Value**
Per-device Train Batchsize	1
Gradient Accumulation Steps	2
Warmup Ratio	0.1
LR Scheduler Type	cosine
Learning Rate	5e-7
Data Type	bfloat16
Optimizer	adamw
Epoch	3
LoRA Rank	8
LoRA α	16

Table 8: The hyperparameters with values used in our fine-tuning experiments.

Model	Q1	Q2	Q3	Q4	Goal	Q5	Q6	Q7	Q8	Task	Q9	Q10	Q11	Q12	Bond	Overall
ChatGPT	0.479	0.627	0.563	0.390	0.515	0.427	0.427	0.289	0.492	0.409	0.557	0.471	0.532	0.237	0.449	0.458
GPT-4o-mini	0.912	0.940	0.928	0.940	0.930	0.938	0.903	0.974	0.935	0.938	0.961	0.818	0.959	0.883	0.905	0.924
Claude-3	0.638	0.776	0.667	0.583	0.666	0.792	0.763	0.830	0.736	0.780	0.706	0.845	0.805	0.671	0.757	0.734
GPT-4	0.946	0.917	0.946	0.943	0.938	0.945	0.942	0.938	0.932	0.939	0.909	0.960	0.943	0.976	0.947	0.941
Qwen1.5-7B-chat	0.330	0.503	0.529	0.552	0.479	0.604	0.552	0.424	0.643	0.556	0.608	0.433	0.518	0.308	0.467	0.500
Qwen1.5-14B-chat	0.672	0.711	0.646	0.717	0.686	0.769	0.772	0.664	0.747	0.738	0.631	0.392	0.749	0.580	0.588	0.671
Meta-Llama-3-8B-Instruct	0.622	0.741	0.549	0.825	0.684	0.764	0.690	0.689	0.738	0.720	0.862	0.635	0.785	0.358	0.660	0.688
Qwen1.5-32B-chat	0.952	0.966	0.962	0.962	0.960	0.988	0.990	0.966	0.970	0.978	0.966	0.901	0.966	0.834	0.917	0.952
Yi-1.5-34B-Chat-16K	0.724	0.760	0.691	0.721	0.724	0.706	0.683	0.791	0.750	0.733	0.516	0.758	0.848	0.585	0.677	0.711
glm-4-9b-chat	0.452	0.773	0.655	0.567	0.612	0.653	0.736	0.785	0.702	0.719	0.871	0.736	0.656	0.477	0.685	0.672
Qwen1.5-72B-chat	0.707	0.519	0.793	0.673	0.673	0.750	0.833	0.907	0.724	0.804	0.810	0.342	0.910	0.608	0.668	0.715
Meta-Llama-3-70B-Instruct	0.894	0.948	0.942	0.963	0.937	0.956	0.968	0.971	0.971	0.967	0.948	0.921	0.922	0.937	0.932	0.945
Qwen1.5-110B-chat	0.955	0.965	0.936	0.946	0.950	0.955	0.974	0.959	0.942	0.958	0.972	0.960	0.958	0.910	0.950	0.953
Llama-3-8B-Instruct-LoRA-SO	0.694	0.642	0.585	0.721	0.660	0.747	0.735	0.730	0.788	0.750	0.786	0.501	0.762	0.486	0.634	0.681
Llama-3-8B-Instruct-LoRA-SR	0.325	0.429	0.474	0.565	0.448	0.607	0.659	0.702	0.568	0.634	0.629	0.487	0.702	0.237	0.417	0.500
Llama-3-8B-Instruct-full-SO	0.417	0.485	0.523	0.547	0.493	0.702	0.729	0.692	0.695	0.705	0.495	0.321	0.696	0.439	0.488	0.562
Llama-3-8B-Instruct-full-SR	0.533	0.677	0.663	0.710	0.646	0.777	0.799	0.752	0.764	0.773	0.601	0.517	0.724	0.540	0.595	0.671

Table 9: The intra-class correlation (ICC) of models in evaluating each question and dimension across different experimental settings.

Model	Q1	Q2	Q3	Q4	Goal	Q5	Q6	Q7	Q8	Task	Q9	Q10	Q11	Q12	Bond	Overall
ChatGPT	0.218	0.181	0.131	0.065	0.285	0.166	0.185	0.106	0.150	0.211	0.272	0.208	0.196	0.082	0.216	0.297
GPT-4o-mini	0.599	0.696	0.694	0.707	0.694	0.714	0.627	0.840	0.708	0.722	0.687	0.509	0.731	0.585	0.782	0.772
Claude-3	0.245	0.371	0.329	0.317	0.366	0.430	0.428	0.465	0.335	0.520	0.534	0.336	0.416	0.212	0.346	0.617
GPT-4	0.582	0.535	0.599	0.578	0.639	0.641	0.666	0.695	0.547	0.691	0.621	0.666	0.591	0.856	0.652	0.830
Qwen1.5-7B-chat	0.105	0.256	0.273	0.289	0.279	0.328	0.296	0.202	0.308	0.347	0.201	0.188	0.245	0.155	0.281	0.392
Qwen1.5-14B-chat	0.337	0.310	0.254	0.310	0.386	0.373	0.304	0.318	0.298	0.451	0.265	0.174	0.371	0.225	0.316	0.521
Meta-Llama-3-8B-Instruct	0.353	0.487	0.279	0.610	0.385	0.489	0.314	0.416	0.452	0.596	0.624	0.365	0.530	0.142	0.551	0.683
Qwen1.5-32B-chat	0.792	0.910	0.882	0.895	0.821	0.964	0.968	0.908	0.930	0.922	0.893	0.731	0.894	0.563	0.684	0.869
Yi-1.5-34B-Chat-16K	0.352	0.445	0.425	0.423	0.508	0.434	0.401	0.492	0.379	0.547	0.370	0.297	0.592	0.286	0.366	0.550
glm-4-9b-chat	0.189	0.468	0.261	0.299	0.370	0.354	0.459	0.456	0.392	0.510	0.610	0.287	0.311	0.110	0.398	0.459
Qwen1.5-72B-chat	0.373	0.259	0.536	0.386	0.346	0.549	0.622	0.682	0.404	0.679	0.445	0.598	0.711	0.327	0.309	0.663
Meta-Llama-3-70B-Instruct	0.580	0.707	0.672	0.740	0.695	0.762	0.789	0.827	0.770	0.794	0.741	0.711	0.653	0.694	0.689	0.854
Qwen1.5-110B-chat	0.824	0.883	0.749	0.768	0.805	0.836	0.888	0.860	0.864	0.870	0.852	0.785	0.849	0.660	0.898	0.860
Llama-3-8B-Instruct-LoRA-SO	0.396	0.298	0.286	0.454	0.462	0.346	0.308	0.399	0.348	0.497	0.474	0.197	0.475	0.118	0.388	0.644
Llama-3-8B-Instruct-LoRA-SR	0.130	0.160	0.120	0.185	0.203	0.196	0.196	0.208	0.180	0.356	0.197	0.018	0.232	0.074	0.189	0.421
Llama-3-8B-Instruct-full-SO	0.077	0.079	0.097	0.070	0.173	0.136	0.158	0.166	0.140	0.375	0.094	0.056	0.125	0.054	0.267	0.409
Llama-3-8B-Instruct-full-SR	0.043	0.142	0.134	0.171	0.260	0.158	0.211	0.139	0.132	0.381	0.108	0.075	0.151	0.077	0.333	0.594

Table 10: The Fleiss' Kappa value of models in evaluating each question and dimension across different experimental settings.

Model	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12
ChatGPT	0.184**	0.166**	0.102	0.11	0.111	0.097	0.134*	0.130*	0.182**	0.064	0.205**	0.07
GPT-4o-mini	**0.300*** **	0.248***	0.203**	0.361***	0.272***	0.360***	0.271***	0.227***	0.396***	0.099	0.415***	0.187**
Claude-3	0.245***	0.297***	0.146*	0.249***	0.324***	0.331***	0.328***	0.202**	0.126	0.182**	0.437***	0.281***
GPT-4	0.147*	0.376***	0.299***	**0.416*** **	0.456***	**0.459*** **	0.340***	0.294***	0.271***	0.259***	0.414***	0.074
Qwen1.5-7B-chat	0.084	0.134*	0.109	0.088	0.113	0.049	0.141*	0.079	0.232***	-0.054	0.226***	0.07
Qwen1.5-14B-chat	0.162*	0.112	0.182**	0.191**	0.225***	0.222***	0.175**	0.111	0.239***	-0.052	0.334***	0.148*
Meta-Llama-3-8B-Instruct	0.125	0.199**	0.206**	0.171**	0.306***	0.168**	0.329***	0.148*	0.407***	0.080	0.244***	0.171**
Qwen1.5-32B-chat	0.076	0.181**	0.306***	0.199**	0.205**	0.197**	0.272***	0.175**	0.263***	0.137*	0.322***	**0.352*** **
Yi-1.5-34B-Chat-16K	0.185**	0.261***	0.314***	0.358***	0.337***	0.307***	0.246***	0.087	0.217***	0.316***	0.269***	0.310***
glm-4-9b-chat	0.163*	0.196**	0.125	0.267***	0.286***	0.295***	0.281***	0.156*	0.316***	0.023	0.340***	0.097
Qwen1.5-72B-chat	0.032	0.216***	0.161*	0.186**	0.273***	0.286***	0.309***	0.246***	0.337***	0.288***	0.387***	0.133*
Meta-Llama-3-70B-Instruct	0.201**	0.216***	0.363***	0.387***	0.327***	0.386***	**0.360*** **	0.182**	0.385***	0.277***	0.398***	0.302***
Qwen1.5-110B-chat	0.229***	0.309***	0.353***	0.372***	0.362***	0.351***	0.294***	0.271***	0.330***	0.229***	0.453***	0.221***
Llama-3-8B-Instruct-LoRA-SO	0.142*	0.213***	0.266***	0.258***	0.289***	0.202**	0.288***	0.145*	0.379***	0.100	0.284***	0.206**
Llama-3-8B-Instruct-LoRA-SR	0.228***	0.083	0.286***	0.272***	0.322***	0.146*	0.306***	0.172**	0.351***	0.180**	0.320***	0.064
Llama-3-8B-Instruct-full-SO	0.271***	0.288***	0.269***	0.253***	0.414***	0.391***	0.323***	0.301***	0.439***	0.275***	0.445***	0.229***
Llama-3-8B-Instruct-full-SR	0.285***	**0.333*** **	**0.363*** **	**0.416*** **	**0.461*** **	0.432***	0.304***	**0.387*** **	**0.455*** **	**0.373*** **	**0.487*** **	0.302***

Table 11: The Pearson correlation results between all models and clients' self-reported assessments on the working alliance questions across different experimental settings. *SO* and *SR* indicate *Score Only* and *Score + Rationale* respectively. The best results are in bold. $^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$.

Dimension	Evaluation	Example
Goal	bad	In our conversations, although the counselor tried to guide me in exploring issues and feelings, we did not clearly agree on specific areas for change. For example, when I mentioned how to deal with a difficult roommate, the counselor focused more on understanding my feelings and situation rather than collaboratively determining specific change strategies. My responses, such as "No thanks" and later mentioning "I might not be suited for this," indicate that my understanding of the counseling process may differ from the counselor's expectations. As a result, we rarely reached a consensus on the areas where I needed to make efforts for change.
	good	In the conversation, the counselor and I jointly established counseling goals, such as, "Let's set the goal to increase daily sleep time to 6 hours." This indicates that we were indeed actively involved in the goal-setting process, which was a continuous part of the counseling sessions, thus aligning with the description of "always" collaboratively setting goals.
Approach	bad	In the conversation, I expressed my feelings about the counselor's questions, saying, "To be honest, I speak directly and bluntly, and this is a free service. I shouldn't have any expectations. I really shouldn't have too many demands and should be content." This indicates that I did not feel the counseling provided new perspectives or methods for addressing my relationship issues. The counselor focused more on listening and understanding, whereas I was expecting concrete guidance and solutions, such as, "Shouldn't you be better and have more ideas?" This shows that I didn't frequently find new ways to view my problems during the counseling sessions.
	good	In the conversation, I mentioned trying to make changes but not finding the right approach, which indicates that I recognize the value of the counseling process. The counselor helped me reflect on my concerns about effort and boundaries, for example, "It seems like you feel that the evaluation of your goodness comes from what you do rather than your traits or personal charisma in relationships." This helped me start to understand my behavioral patterns more deeply. Although I am still exploring, I feel that counseling has indeed guided me toward the changes I want.
Bond	bad	In the conversation, I expressed doubts about the counselor's professionalism and mentioned, "You gave me the impression of being a bit rushed, not allowing me time to process my changes," as well as "I have a lot of anger towards you." This indicates that I did not feel that the counselor was attuned to my personal preferences but rather sensed a mismatch in professional approach.
	good	During the counseling process, the counselor demonstrated patience and understanding, such as by asking, "Can you describe the sense of presence you have experienced from childhood to now?" This indicates that the counselor actively listens and is concerned about my feelings. Although personal preferences were not directly addressed, the counselor's positive feedback and in-depth exploration, such as, "The sense of presence you described—when you feel that doing something genuinely brings joy to others—seems to be part of your value system. It looks like you have a direction, but the path is blocked," show the counselor's deep understanding and care for me, making me feel valued and supported.

Table 12: Example explanations generated by our best-performing model.