PsyPARSE: Retrieval-Augmented Slow Thinking for Personalized Empathetic Counseling

Longxiang Wang[1,*], Pukun Zhao[2,*] Chen Chen[2], Jinhe Bi[3], Huacan Wang[4], Tong Zhang[5,†], Ronghao Chen[6,†]

[1]Chongqing University
[2]Guangdong University of Finance and Economics
[3]Ludwig Maximilian University of Munich
[4]University of Chinese Academy of Sciences
[5]Zhejiang University
[6]Peking University

20223610@stu.cqu.edu.cn, {zhaopukun, Allen821}@student.gdufe.edu.cn, Jinhe.bi@campus.lmu.de, wanghuacan17@mails.ucas.ac.cn, tz_zju@zju.edu.cn, chenronghao@alumni.pku.edu.cn

Abstract

The escalating global demand for mental health services highlights the potential of Large Language Models (LLMs) in psychological counseling. However, current LLM-based approaches, particularly fine-tuned models, are constrained by data distribution biases, leading to limited therapeutic diversity and personalization. Crucially, they often lack anticipatory empathetic reasoning, struggle to foresee patient emotional responses beyond immediate dialogue history, and incur substantial computational costs. To address these limitations, we propose PsyPARSE, a novel training-free framework for psychological counseling that emulates the deliberate and empathetic reasoning of human counselors. PsyPARSE integrates Multi-Therapy Retrieval-Augmented Generation (RAG) to overcome data biases and provide highly personalized therapeutic approaches tailored to individual patient attributes. Pioneering the first multi-stage slow-thinking engine in mental health LLMs, PsyPARSE employs Multi-Turn Rollouts to identify optimal therapeutic paths and through anticipating patient reactions, optimizes empathetic responses, thereby ensuring genuinely empathetic and impactful responses in complex, long-dialogue interactions. Operating as a plug-and-play solution, PsyPARSE avoids the computational burden of fine-tuning. We establish a comprehensive LLM-based patient-therapist agent simulation framework for evaluation. Extensive experiments demonstrate that PsyPARSE significantly enhances the capabilities of various LLM baselines, achieving superior personalization and deeper empathy compared to both fine-tuned and other training-free methods. This work offers an efficient, adaptable, and scalable solution to advance mental health support.

Introduction

The global rise in mental health issues, such as depression and anxiety, has significantly increased the demand for psychological services (Santomauro et al. 2021). LLMs (Liu et al. 2024; Achiam et al. 2023; Xiang et al. 2025; Xiang, Zhang, and Chen 2025; Bai et al. 2023; Zhang et al. 2025; Zhao et al. 2025), with advanced natural language capabilities, show promise in addressing this demand by simulating human dialogue and providing emotional support, thus alleviating resource scarcity and geographical limitations.

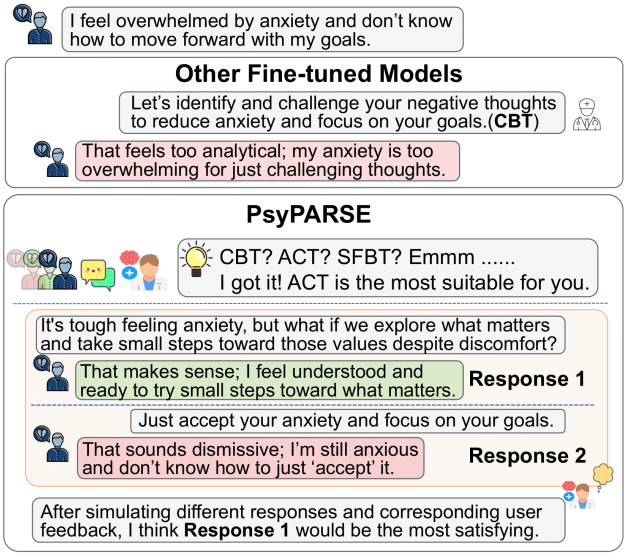


Figure 1: Comparison of PsyPARSE and other fine-tuned LLMs in psychological counseling. Typical models often exhibit therapy limitations due to data bias, whereas PsyPARSE provides personalized therapeutic approaches and more empathetic responses through anticipatory simulation.

Due to ethical and privacy constraints, real-world multi-turn mental health dialogue datasets are scarce, leading most LLM-based research (Qiu et al. 2024; Chen et al. 2023; Qiu et al. 2023; Zhang et al. 2024; Xin Yan 2023; Lee et al. 2024b; Hu et al. 2024) to rely on limited real or synthetic datasets for fine-tuning. However, these datasets often suffer from data distribution biases, causing fine-tuned models to focus on a few mainstream therapies like Cognitive

*Equal contribution
†Corresponding authors.

Behavioral Therapy (CBT) (Beck et al. 1979), while lacking diversity. Although recent efforts (Zhang et al. 2024) have used datasets with a broader range of therapeutic approaches, these models still struggle to deliver targeted, personalized interventions that account for diverse patient attributes, limiting their utility in varied counseling scenarios.

Furthermore, while some studies (Hu et al. 2024; Zhang et al. 2024; Lee et al. 2024b) have improved LLMs' semantic empathy by fine-tuning, these approaches often train models to generate responses based primarily on the immediate dialogue history. This can limit their ability to emulate the deep reasoning and reflection human counselors employ in complex, emotional, high-stakes dialogues, especially when not explicitly designed to simulate and evaluate potential conversational futures. This results in a lack of anticipatory empathetic reasoning, where models may struggle to foresee patients' potential emotional responses and provide truly empathetic support, often producing responses that prioritize surface-level semantic empathy over genuine impact. Lastly, fine-tuning approaches demand substantial computational resources, yet frequently yield suboptimal performance, highlighting the need for more efficient, training-free solutions.

To address these challenges, we propose PsyPARSE, a Psychological LLM with RAG-Enhanced Personalization and Slow-Thinking Engine. PsyPARSE is a novel training-free framework that emulates the empathetic and deliberate reasoning of human counselors through a plug-and-play design. It combines multi-therapy Retrieval-Augmented Generation (RAG) (Arslan et al. 2024) with a slow-thinking mechanism to overcome data biases, enhance personalization, and enable empathetic long-dialogue interactions(as shown in Fig 1). To construct and evaluate PsyPARSE, we establish a comprehensive patient-therapist agent dialogue simulation framework, utilizing LLMs to generate personalized patient profiles and simulate realistic multi-turn interactions with various baseline counselor models. Extensive experiments demonstrate that PsyPARSE significantly enhances the capabilities of baseline models of various scales in psychological counseling scenarios, verifying its wide adaptability and effectiveness. In summary, the key contributions of our work are as follows:

- **RAG-Enhanced Personalization for Diverse Therapeutic Needs:** PsyPARSE leverages multi-therapy Retrieval-Augmented Generation (RAG) to tailor counseling to individual patient attributes, overcoming data biases in existing models and enabling a wider range of therapeutic approaches beyond mainstream methods.
- **Pioneering a Slow-Thinking Mechanism for Empathetic and Personalized Counseling:** PsyPARSE introduces a novel slow-thinking engine—the first in LLM-based mental health research—that operates in two stages: a Multi-Turn Rollout phase to simulate dialogues based on detailed therapies retrieved by RAG and identify the most suitable approach for the patient, followed by an Empathetic Response Optimization phase to simulate therapist responses and simulate potential patient responses, enabling anticipatory thinking for empathetic, human-like counseling absent in models relying on immediate dialogue history.
- **Training-Free, Plug-and-Play Framework:** PsyPARSE operates as a training-free framework using prompt engineering and multi-agent collaboration during inference, avoiding the computational costs of fine-tuning while ensuring adaptability and effectiveness.

Related Work

LLM Applications in Mental Health Counseling. LLMs are increasingly used in mental health counseling to address growing demand. Early efforts, such as Psychat (Qiu et al. 2024), Mindchat (Xin Yan 2023), Soulchat (Chen et al. 2023), Mechat (Qiu et al. 2023), PsycoLLM (Hu et al. 2024), Cactus (Lee et al. 2024b), fine-tune LLMs on limited real or synthetic datasets to generate empathetic responses and simulate counseling dialogues. These studies demonstrate LLMs' initial promise for emotional support, particularly in resource-scarce settings. However, their reliance on datasets with limited therapeutic diversity often biases models toward mainstream approaches like CBT, limiting applicability across diverse counseling scenarios. More recent work, such as CPsyCoun (Zhang et al. 2024), improves the diversity of counseling scenarios by introducing datasets that encompass a broader range of therapeutic approaches. Despite this progress, CPsyCoun and similar efforts still struggle to deliver targeted, personalized interventions that fully account for diverse patient attributes, highlighting the need for more adaptive and empathetic models in long-dialogue contexts. In contrast, PsyPARSE employs a novel approach combining multi-therapy RAG with its Slow-thinking Engine. The engine incorporates a multi-turn rollout mechanism to dynamically select suitable therapeutic strategies based on conversational context, along with a response pruning module that filters candidate outputs for mitigating the limitations inconsistent responses in previous models.

Psychological Therapy Techniques. Psychological therapy techniques play a vital role in enhancing mental well-being and addressing emotional challenges (Grawe 2004). CBT targets the interplay of thoughts, emotions, and behaviors (Beck et al. 1979), effectively treating disorders such as anxiety and depression by modifying distorted thinking patterns. Acceptance and Commitment Therapy (ACT) promotes accepting difficult emotions while encouraging value-aligned actions (Harris 2006), which proves beneficial for conditions such as chronic pain and depression through enhanced psychological flexibility. Solution-Focused Brief Therapy (SFBT) is a goal-oriented, short-term method that emphasizes building solutions over dwelling on problems (Kim 2008), making it effective for rapid interventions such as crisis counseling. Beyond these, numerous other therapies, such as Dialectical Behavior Therapy (DBT) and Rational Emotive Behavior Therapy (REBT), offer diverse strategies to meet the varied needs of patients (Linehan 2014; Dryden 2005). However, current LLM-based counseling models often struggle to dynamically integrate these methods to deliver personalized therapeutic support, making it challenging to effectively address patients' unique psycho-

logical needs. PsyPARSE addresses this by using RAG to retrieve a multi-therapy database, ensuring responses reflect diverse, tailored psychotherapies.

Method

Overall Architecture

In this section, we present the proposed PsyPARSE Framework in detail. The framework comprises three key stages, illustrated in Fig 2: (1) Interaction and Profile Construction: We first collect patient information through simulated interaction, building a patient profile that guides subsequent therapeutic processes. (2) Multi-Therapy Retrieval-Augmented Generation (RAG): Based on the patient profile, we use the Therapist Agent to retrieve suitable therapy strategies from a structured multi-therapy database and generate an initial guidance framework. (3) Slow Thinking Strategy Refinement: To ensure contextual relevance and emotional appropriateness, the Therapist Agent performs a two-step refinement process. It first simulates multi-turn dialogues using retrieved therapeutic strategies to assess their suitability. Then it generates candidate responses, predicts patient reactions, and selects the response that best fits the therapeutic goals and emotional context. Details of the simulation and decision-making process are provided in the Appendix.

Interaction and Profile Construction

This initial stage aims to construct a foundational psychological profile of the patient through systematic data collection and structured processing. Specifically, Patient Agents are generated using LLMs to extract diverse and representative psychological scenarios from the CPsyCounR dataset (Zhang et al. 2024) (see Appendix for prompt details). These simulated patients exhibit a range of emotional states, symptom patterns, and interpersonal issues.

Then, the proposed Therapist Agent is introduced to engage in structured interactions with the Patient Agents using a set of baseline counseling strategies. These interactions are designed to elicit critical clinical information, such as emotional triggers, cognitive distortions, and therapeutic goals. The resulting patient profile forms the basis for therapy selection and dialogue planning in subsequent stages.

Multi-therapy Retrieval-Augmented Generation

For personalized therapeutic guidance, we propose a Multi-Therapy Retrieval-Augmented Generation (RAG) framework that integrates structured domain knowledge with adaptive response generation. We first construct a multi-therapy database to address distributional bias and expand therapeutic diversity. To enhance personalization, we then introduce a personalized therapy matching via RAG module that matches therapies to patient profiles and generates a tailored guidance framework for counseling simulations.

Construction of the Multi-Therapy DataBase. Existing psychological counseling datasets often exhibit a distribution bias towards mainstream therapies, such as CBT, restricting the generalization of the model to diverse patient needs and hindering the recommendation of less common but potentially more suitable therapeutic approaches. To address these limitations and provide personalized services that adequately cater to the unique traits and symptoms of individual patients, we construct a structured Multi-Therapy DataBase. We employ a LLM to extract Case Profiles (*e.g.*, core problems, emotional states, symptoms) and Therapy Tags (*e.g.*, therapy type, techniques, applicable conditions, procedural steps) from the counseling reports in CPsyCounR. This knowledge database is designed to capture a broad spectrum of therapeutic approaches, thereby overcoming data distribution biases that have limited the capacity of prior work to recommend diverse therapies. The knowledge database utilizes a hybrid indexing approach: Therapy Tags are indexed using inverted indexing to support keyword-based lookups, while BERT embeddings of Case Profiles (Issa et al. 2023) are used to capture semantic similarity for contextual retrieval.

Personalized Therapy Matching via RAG. To enhance personalization, we develop a retrieval-augmented generation (RAG) process that identifies appropriate therapies based on individual patient characteristics and generates tailored therapeutic guidance. Specifically, we first collect key patient information using a predefined question matrix. This information is then transformed into two components: a structured patient profile P (*e.g.*, symptoms and emotional states), and a set of background keywords K extracted via TF-IDF (Koloski et al. 2022). These representations jointly serve as the basis for retrieving relevant cases from the Multi-Therapy DataBase. To ensure both semantic and lexical relevance, we employ a hybrid retrieval strategy. FAISS (Douze et al. 2024) is used to compute the semantic similarity between P and each case profile C_i using BERT embeddings, while BM25 (Robertson, Zaragoza et al. 2009) measures the lexical overlap between K and the associated Therapy Tags T_i. The final similarity score for each candidate case is calculated as:

$$S(P, C_i) = \alpha \cdot \mathcal{C}(E(P), E(C_i)) + (1 - \alpha) \cdot \mathcal{B}(K, T_i), \quad (1)$$

where $E(\cdot)$ denotes BERT embeddings, $\mathcal{C}(\cdot, \cdot)$ represents the cosine similarity function, $\mathcal{B}(K, T_i)$ is the BM25 score between keywords K and Therapy Tags T_i, and α is a balancing coefficient. The top-k_1 most relevant cases are then selected for further evaluation.

Next, we introduce an Evaluation Agent that scores each associated therapy T_i with respect to the patient profile P to assess the therapeutic suitability of the retrieved cases. The overall therapy score is computed as:

$$S_{\text{th}}(P, T_i) = w_1 \cdot \mathcal{M}_s(P, T_i) + w_2 \cdot \mathcal{A}_{pp}(T_i), \quad (2)$$

where $\mathcal{M}_s(P, T_i)$ is the symptom matching score determined using Jaccard similarity between the patient's symptoms and those addressed by the therapy, $\mathcal{A}_{pp}(T_i)$ is a precomputed applicability score for therapy T_i derived from the CPsyCounR, and w_1 and w_2 are weighting factors. Therapies scoring below a predefined threshold are filtered out, and the top-k_2 highest-scoring therapies are retained.

Finally, the Therapist Agent synthesizes a personalized counseling guidance framework based on the selected therapies. This framework—comprising the recommended therapy type, core techniques, and procedural steps—serves

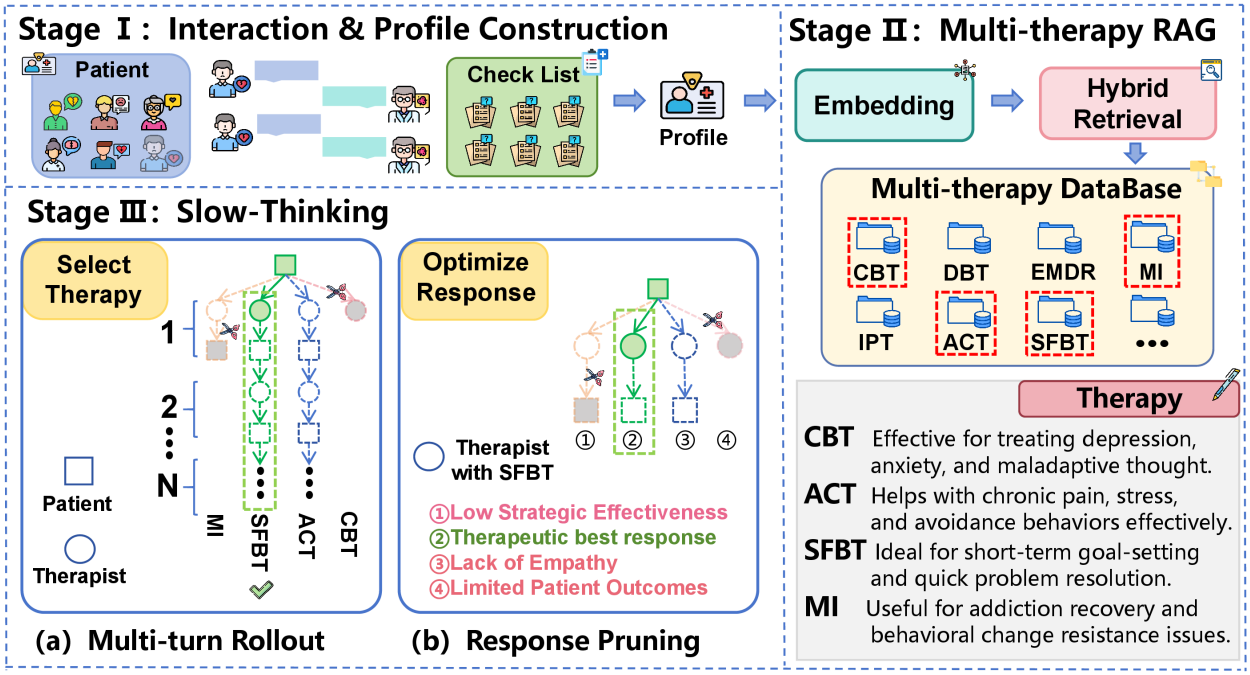


Figure 2: **Overview of the PsyPARSE workflow.** The process begins with (1) **Interaction and Profile Construction:**, where Therapist and Patient Agents gather initial patient data. Next, (2) **Multi-therapy RAG** enables the Therapist Agent to retrieve suitable therapies from the multi-therapy database. The system then enters (3) **Slow-Thinking**, a two-step deliberation process that involves **Multi-turn Rollout** to identify the optimal therapeutic approach via simulated dialogues, and subsequent **Response Pruning** to select the best therapist response by anticipating patient reactions.

as a central reference for subsequent interactions, enabling context-aware and patient-specific interventions.

Slow-Thinking Process

To improve the effectiveness of therapies retrieved by the RAG module and to support more accurate and personalized therapeutic interventions, we propose the **Slow-Thinking Process**. This framework addresses key limitations of fine-tuned psychological language models, including overreliance on recent dialogue history, limited capacity for anticipatory reasoning, and difficulty in sustaining empathy throughout extended conversations. The Slow-Thinking Process features a two-stage deliberative structure: (1) Multi-Turn Rollout for Personalized Therapy, which identifies contextually appropriate therapeutic strategies through iterative simulation; and (2) Response Pruning for Empathetic Optimization, which refines response generation to enhance empathy and maintain conversational stability.

Multi-Turn Rollout for Personalized Therapy. To determine the most contextually appropriate therapeutic strategy for an individual patient, we propose a multi-turn rollout mechanism that emulates the iterative reasoning process of a professional therapist, thereby enabling structured and progressive comparison of treatment options over multiple conversational turns. We first obtain a set of candidate therapies through a RAG process. We then evaluate these candidates based on their initial contextual relevance and select the top-k therapies. For each of the k selected therapy candidates, the Therapist Agent generates multiple response variants. These responses are subsequently evaluated by an Evaluation Agent, which assigns scores based on their empathy and alignment with predefined therapeutic goals as follows:

$$S_{\text{resp}}(R_i) = w_{e,m} \cdot \mathcal{E}(R_i) + w_{t,m} \cdot \mathcal{A}(R_i, T_j), \quad (3)$$

where $S_{\text{resp}}(R_i)$ denotes the overall score of response R_i, computed as a weighted sum of the empathy score $\mathcal{E}(R_i)$ (*e.g.*, from sentiment analysis) and the alignment score $\mathcal{A}(R_i, T_j)$ (*e.g.*, via keyword or semantic similarity). The weights $w_{e,m}$ and $w_{t,m}$ specify the contributions of empathy and alignment. Responses scoring below a predefined threshold are discarded and the remaining responses are used in multi-turn interactions with a Simulated Patient Agent. The dialogue proceeds over several turns, and the trajectory with the highest cumulative score is selected to ensure emotional consistency and alignment with the patient's therapeutic needs.

Response Pruning for Empathetic Optimization. Following the establishment of the optimal therapeutic framework through the multi-turn rollout, we propose the response optimization phase to emulate the reflective reasoning and anticipatory empathy of human counselors, ensuring each response is contextually coherent, emotionally attuned, and free from inconsistencies common in single-turn outputs. For each dialogue turn, the Therapist Agent generates n diverse candidate responses based on the chosen therapeutic approach (controlled by the temperature parameter T_{single},

a pre-defined hyperparameter that controls the randomness and diversity of the LLM's generated text). To ensure that the response is appropriate, the system simulates one step ahead by asking the Simulated Patient Agent to generate a potential reaction to each of the n therapist candidates instead of directly selecting a response. This anticipatory approach allows for an evaluation not only of the therapist's standalone utterance but, more importantly, its likely impact on the patient. Each candidate therapist response R_i, along with its corresponding simulated patient reaction P_i, is then scored based on empathy and conversational stability:

$$S_{\text{single}}(R_i, P_i) = w_{e,s} \cdot \mathcal{E}(R_i) + w_s \cdot \mathcal{S}(R_i, P_i), \quad (4)$$

where $S_{\text{single}}(R_i, P_i)$ is the score for therapist response R_i given patient response P_i, $\mathcal{E}(R_i)$ is the empathy score(e.g., via sentiment analysis), and $\mathcal{S}(R_i, P_i)$ evaluates the stability of R_i based on P_i. $w_{e,s}$ and w_s are weighting factors for empathy and stability in the single-turn rollout. The therapist candidate response with the highest score through this simulation and pruning process is selected as the actual output for that turn. This turn-by-turn optimization ensures empathetic, therapeutically focused, and controllable responses, proactively mitigating potential misunderstandings or negative patient reactions, and addressing the shortcomings of models relying on dialogue history and lacking forward-looking simulation. This process repeats for each turn until the dialogue concludes.

Experiments

Experimental Setup

Dataset. We use the CPsyCounR dataset (Zhang et al. 2024) to build the Multi-Therapy Database, independent of patient profiles. From the remaining data, 50 patient profiles are randomly sampled and used as the test set for the Patient Agent.

Baselines. We select a diverse set of baseline models to comprehensively evaluate performance of PsyPARSE as follows:

- **Closed-Source Models**: GPT-4 (Achiam et al. 2023).
- **Open-Source Models**: Qwen2.5-7B (Yang et al. 2024), Qwen2.5-32B (Yang et al. 2024), LLaMA-3.1-8B (Touvron et al. 2023), LLaMA-3.1-70B (Touvron et al. 2023), and DeepSeekV3 (Liu et al. 2024).
- **Domain-Specific Models**: MeChat (Qiu et al. 2023), PsyChat (Qiu et al. 2024), SoulChat (Chen et al. 2023), PsycoLLM (Hu et al. 2024), MindChat (Xin Yan 2023), and CPsyCounX (Zhang et al. 2024).
- **Training-Free Strategies**: Few Shot Learning (Few Shot) (Ravi and Larochelle 2017), Chain-of-Thought (CoT) (Wei et al. 2022), Chain of Empathy (CoE) (Lee et al. 2023), and Empathetic Meta-Chain (EMC) (Lee et al. 2024a).

Evaluation Metrics. We evaluate the performance of PsyPARSE using LLM-based and human evaluations, focusing on two key aspects: personalization and empathy. Definitions and rubrics for all criteria are provided in Appendix.

Given the limitations of traditional automatic metrics for multi-turn psychological counseling dialogues (Zhao et al. 2024), we apply an LLM-based evaluation (Peng et al. 2024). A prompted LLM acting as a psychological expert rates full dialogues on six criteria: Coherency (Coh.), Completeness (Cpl.), Humanoid Quality (Hum.), Technical Specificity (Tec.), Strategic Progress (Str.), and Contextual Personalization (Ctx.). Among them, Tec. and Ctx. reflect personalization, while Hum. reflects empathy.

In addition, we conduct human evaluation with certified psychological counselors, following ethical standards (Zhang et al. 2024). Evaluators rate each dialogue on a 0–3 scale using four criteria: Empathy (Emp.), Therapy Appropriateness (Thp.), Consistency (Con.), and Strategy (Str.). In this evaluation, Thp. measures personalization via therapy relevance and effectiveness, while Emp. reflects empathy by the emotional sensitivity of responses.

Implementation Details. Based on ablation studies, key parameters were set to ensure effective system performance. Therapy selection module employed 3 rounds of multi-turn rollout for iterative refinement. The response optimization module considered 4 candidate branches during empathetic response pruning and selection. In the RAG module, the top 3 therapeutic cases were retained after filtering ($k_2 = 3$). Further configuration details are provided in Appendix.

Main Results

Comparison with Training-Free Methods. Table 1 presents LLM-based evaluations of training-free methods on DeepseekV3 for domain-specific tasks. Existing training-free approaches, such as Few Shot, CoT and EMC, exhibit limited improvements in personalization (Tec. and Ctx.) and empathy (Hum.). The strongest baseline, CoE, achieves scores of 87.38 (Ctx.), 75.85 (Tec.), and 92.00 (Hum.), revealing deficiencies in capturing nuanced user interactions. In contrast, our method significantly outperforms these baselines, attaining state-of-the-art scores of 95.72 (Ctx.), 95.50 (Tec.), and 94.90 (Hum.), surpassing CoE by 8.34, 19.65, and 2.90 points, respectively. These results highlight our approach's ability to overcome the limitations of existing training-free methods, delivering superior personalization and empathy in domain-specific applications.

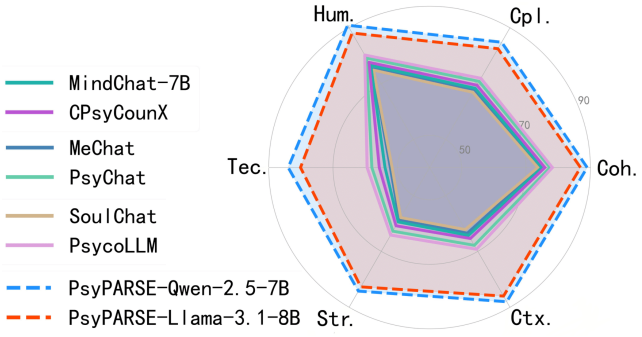


Figure 3: Model Comparison on Automatic Evaluation

Comparison with Fine-Tuned Domain-Specific Models. As illustrated in Figure 3, PsyPARSE-enhanced models significantly outperform fine-tuned domain-specific models of comparable scale. PsyPARSE elevates general-purpose

	Automatic Evaluation						Human Evaluation			
Method	**Coh.**	**Cpl.**	**Hum.**	**Tec.**	**Str.**	**Ctx.**	**Emp.**	**Thp.**	**Con.**	**Str.**
DeepseekV3	85.22	80.22	85.60	75.69	78.13	80.51	2.45	2.50	2.55	2.40
+FEWSHOT	86.55(↑1.33)	84.81(↑4.59)	89.54(↑3.94)	74.35(↓1.34)	83.73(↑5.60)	83.03(↑2.52)	2.55(↑0.10)	2.60(↑0.10)	2.65(↑0.10)	2.50(↑0.10)
+COT	86.75(↑1.53)	85.87(↑5.65)	90.99(↑5.39)	79.15(↑3.46)	85.37(↑7.24)	85.72(↑5.21)	2.60(↑0.15)	2.65(↑0.15)	2.70(↑0.15)	2.55(↑0.15)
+COE	89.70(↑4.48)	86.82(↑6.60)	92.00(↑6.40)	75.85(↑0.16)	86.45(↑8.32)	87.38(↑6.87)	2.70(↑0.25)	2.75(↑0.25)	2.80(↑0.25)	2.65(↑0.25)
+EMC	82.40(↓2.82)	85.64(↑5.42)	84.48(↓1.12)	73.66(↓2.03)	85.84(↑7.71)	78.39(↓2.12)	2.40(↓0.05)	2.45(↓0.05)	2.50(↓0.05)	2.35(↓0.05)
PsyPARSE-DeepseekV3	95.00(↑9.78)	92.64(↑12.42)	94.90(↑9.30)	95.50(↑19.81)	92.63(↑14.50)	95.72(↑15.21)	2.92(↑0.47)	2.95(↑0.45)	2.93(↑0.38)	2.90(↑0.50)

Table 1. Comparison of Training-Free and Our Method on DeepseekV3 for Domain-Specific Task (Automatic + Human Evaluation Metrics)

	Automatic Evaluation						Human Evaluation			
Model	**Coh.**	**Cpl.**	**Hum.**	**Tec.**	**Str.**	**Ctx.**	**Emp.**	**Thp.**	**Con.**	**Str.**
Closed-Source Models										
GPT4	71.00	64.73	74.07	50.83	60.83	66.93	2.20	2.15	2.25	2.10
PsyPARSE-GPT4	92.33(↑21.33)	88.00(↑23.27)	92.91(↑18.84)	87.64(↑36.81)	88.24(↑27.41)	91.27(↑24.34)	2.75(↑0.55)	2.78(↑0.63)	2.80(↑0.55)	2.72(↑0.62)
Open-Source Models										
Qwen2.5-7B	73.78	67.22	76.11	54.22	63.56	66.78	2.10	2.05	2.15	2.00
PsyPARSE-Qwen2.5-7B	89.00(↑15.22)	84.84(↑17.62)	90.93(↑14.82)	83.93(↑29.71)	84.22(↑20.66)	88.02(↑21.24)	2.60(↑0.50)	2.62(↑0.57)	2.58(↑0.43)	2.55(↑0.55)
Qwen2.5-32B	82.00	73.67	85.24	63.78	72.22	75.00	2.35	2.38	2.42	2.30
PsyPARSE-Qwen2.5-32B	91.22(↑9.22)	86.53(↑12.86)	92.33(↑7.09)	85.67(↑21.89)	86.18(↑13.96)	90.33(↑15.33)	2.70(↑0.35)	2.73(↑0.35)	2.75(↑0.33)	2.68(↑0.38)
Llama-3.1-8b	68.44	56.49	60.62	47.56	45.64	52.44	1.90	1.85	1.95	1.80
PsyPARSE-Llama-3.1-8B	86.94(↑18.50)	82.53(↑26.04)	88.13(↑27.51)	80.11(↑32.55)	82.71(↑37.07)	86.00(↑33.56)	2.55(↑0.65)	2.50(↑0.65)	2.52(↑0.57)	2.48(↑0.68)
Llama-3.3-70b	80.56	75.11	81.13	65.11	71.13	74.22	2.30	2.35	2.40	2.28
PsyPARSE-Llama-3.3-70B	90.33(↑9.77)	85.69(↑10.58)	90.62(↑9.49)	84.38(↑19.27)	85.51(↑14.38)	88.82(↑14.60)	2.65(↑0.35)	2.68(↑0.33)	2.70(↑0.30)	2.62(↑0.34)
DeepseekV3	85.22	80.22	85.60	75.69	78.13	80.51	2.45	2.50	2.55	2.40
PsyPARSE-DeepseekV3	95.00(↑9.78)	92.64(↑12.42)	94.90(↑9.30)	95.50(↑19.81)	92.63(↑14.50)	95.72(↑15.21)	2.92(↑0.47)	2.95(↑0.45)	2.93(↑0.38)	2.90(↑0.50)
Domain-Specific Models										
CPsyCounX	75.84	69.46	77.65	55.62	60.84	65.39	2.10	2.05	2.15	2.00
PsyPARSE-CPsyCounX	85.37(↑9.53)	82.91(↑13.45)	84.62(↑6.97)	82.17(↑26.55)	81.46(↑20.62)	85.82(↑20.43)	2.50(↑0.40)	2.55(↑0.50)	2.58(↑0.43)	2.48(↑0.48)
PsyChat	77.46	70.83	79.17	57.93	62.78	67.84	2.20	2.15	2.25	2.10
PsyPARSE-PsyChat	86.82(↑9.36)	83.76(↑12.93)	86.53(↑7.36)	84.73(↑26.80)	83.51(↑20.73)	87.94(↑20.10)	2.60(↑0.40)	2.65(↑0.50)	2.68(↑0.43)	2.58(↑0.48)
PsycoLLM	78.19	71.94	80.26	59.41	64.17	69.23	2.25	2.20	2.30	2.15
PsyPARSE-PsycoLLM	87.94(↑9.75)	84.82(↑12.88)	87.69(↑7.43)	86.29(↑26.88)	85.38(↑21.21)	89.47(↑20.24)	2.65(↑0.40)	2.70(↑0.50)	2.72(↑0.42)	2.62(↑0.47)
MeChat	74.29	67.95	75.28	52.19	58.63	63.51	1.95	1.90	2.00	1.85
PsyPARSE-MeChat	82.46(↑8.17)	80.28(↑12.33)	81.54(↑6.26)	78.42(↑26.23)	78.29(↑19.66)	82.46(↑18.95)	2.30(↑0.35)	2.35(↑0.45)	2.38(↑0.38)	2.28(↑0.43)
MindChat-7B	74.92	68.37	76.54	53.84	59.47	64.27	2.00	1.95	2.05	1.90
PsyPARSE-MindChat-7B	83.28(↑8.36)	81.49(↑13.12)	82.91(↑6.37)	79.58(↑25.74)	79.41(↑19.94)	83.29(↑19.02)	2.35(↑0.35)	2.40(↑0.45)	2.42(↑0.37)	2.32(↑0.42)
SoulChat	73.68	66.92	74.86	51.63	57.92	62.19	1.90	1.85	1.95	1.80
PsyPARSE-SoulChat	81.73(↑8.05)	79.82(↑12.90)	80.92(↑6.06)	77.64(↑26.01)	77.82(↑19.90)	81.73(↑19.54)	2.25(↑0.35)	2.30(↑0.45)	2.35(↑0.40)	2.22(↑0.42)

Table 2. LLM Evaluation Performance of Models (Automatic + Human Evaluation)

LLMs, enabling them to surpass specialized models in personalization and empathy. For instance, PsyPARSE-Qwen2.5-7B achieves scores of 88.02 (Ctx.), 83.93 (Tec.), and 90.93 (Hum.), outperforming representative fine-tuned models like PsycoLLM(69.23 Ctx., 59.41 Tec., 80.26 Hum.) by 18.79, 24.52, and 10.67 points, respectively. Similarly, PsyPARSE-Llama-3.1-8B attains scores of 86.00 (Ctx.), 80.11 (Tec.), and 88.13 (Hum.), consistently surpassing models like SoulChat and MindChat-7B. These results demonstrate that PsyPARSE enables general-purpose LLMs to achieve superior personalization and empathy, overcoming the limitations of fine-tuned models constrained by domain-specific training data.

Plug-and-Play Across Diverse Models. As shown in Table 2, PsyPARSE demonstrates robust plug-and-play capability, consistently enhancing performance across diverse model architectures and scales. For instance, PsyPARSE-GPT4 notably increased Ctx. by 24.34 points and Thp. by 0.63 points. Similarly, PsyPARSE-DeepseekV3 achieved a Ctx. score of 95.72, representing an increase of 15.21 points, and improved Tec. by 19.81 points to reach 95.50. PsyPARSE-Qwen2.5-32B likewise lifted its Tec. score to 85.67, a substantial gain of 21.89 points. Furthermore, domain-specific models such as PsyPARSE-CPsyCounX and PsyPARSE-SoulChat also demonstrated marked improvements in Ctx. and Emp. scores. Specifically, in empathy-related metrics, PsyPARSE-DeepseekV3 achieved a Hum. score of 94.90, an increase of 9.30 points, and raised its Emp. score by 0.47 points to 2.92. PsyPARSE-Llama-3.1-8B, moreover, showed a significant gain in Hum. of 27.51 points, reaching 88.13. These consistent improvements across models underscore PsyPARSE's substantial generalization capabilities and versatility for real-world psychological counseling applications.

Ablation Study

All ablation studies are conducted on DeepSeekV3 to evaluate the contributions of PsyPARSE's core components and parameter configurations in enhancing personalization and empathy for psychological counseling tasks.

Configuration	Coh.	Cpl.	Hum.	Tec.	Str.	Ctx.
w/o all	85.22	80.22	85.60	75.69	78.13	80.51
w/o RAG	92.75	89.46	92.19	76.82	84.29	82.95
w/o slow-thinking	86.73	83.47	84.83	89.47	85.92	87.46
all	**95.00**	**92.64**	**94.90**	**95.50**	**92.63**	**95.72**

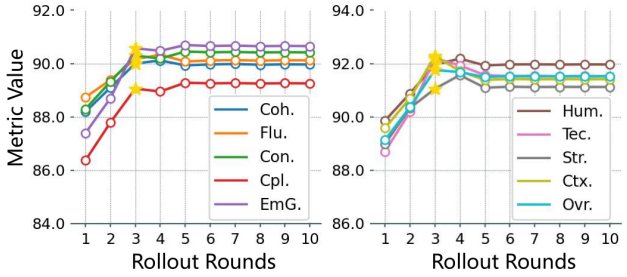
Table 3. Ablation study of key components in PsyPARSE.

Module Ablation. Table 3 details the impact of ablating key components of PsyPARSE compared to the full method. Removing the entire framework (w/o) leads to significant performance degradation, particularly in personalization where Ctx. decreases by 15.21 and Tec. by 19.81, and in empathy where Hum. decreases by 9.30, confirming the framework's integral role in psychological counseling. Excluding RAG(w/o RAG) notably impairs personalization, decreasing Tec. by 18.68 and Ctx. by 12.77, underscoring RAG's critical contribution to technical specificity and contextual understanding. Omitting slow-thinking (w/o slow-thinking) results in a 10.07 decline in Hum., highlighting their importance in fostering empathetic interactions during therapy exploration and response pruning. These results collectively affirm that Multi-therapy RAG and slow-thinking are vital components, enabling PsyPARSE to achieve superior personalization and empathy in domain-specific applications.

Parameter Ablation for Rollouts and RAG. To identify the optimal configuration of PsyPARSE's components, we conducted parameter ablation studies on multi-turn rollout rounds for therapy selection, branches for empathetic response pruning, and top-k selection for RAG. The full configuration, comprising 3 rounds for multi-turn rollout, 4 branches for response pruning, and top-3 RAG selection, achieves Ctx. 95.72, Tec. 95.50, and Hum. 94.90. These results demonstrate that the selected parameters effectively balance personalization and empathy, establishing an optimal setup for PsyPARSE's performance in psychological counseling tasks.

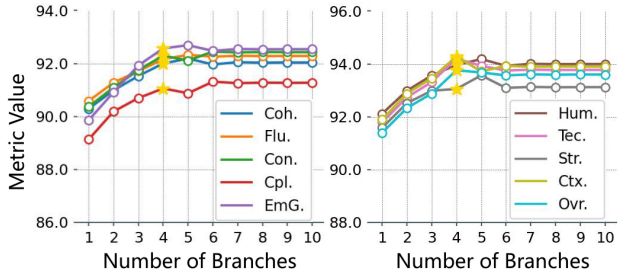
Multi-Turn Rollout Rounds for Therapy Selection. Figure 4 illustrates the impact of multi-turn rollout rounds on therapy selection in PsyPARSE, where rollouts simulate iterative therapist-patient interactions to identify optimal therapeutic strategies. The full configuration with 3 rollout rounds achieves optimal Ctx. at 95.72 and Hum. at 94.90. Reducing to Round 1 decreases Ctx. by 6.77 points to 88.95 and Hum. by 10.07 points to 84.83, indicating insufficient exploration of therapeutic options.Increasing beyond 3 rounds (*e.g.*, Ctx. of 93.53 at 10 rounds) yields stable performance but incurs higher computational costs. Thus, 3 rollout rounds optimally balance therapeutic exploration and computational efficiency for psychological counseling tasks.

Branches for Empathetic Response pruning. Figure 5 evaluates the impact of branch count in PsyPARSE's empathetic response pruning, which refines therapist responses to enhance empathy and personalization. The full configuration with 4 branches achieves Hum. 94.90 and Ctx. 95.72. Reducing to 1 branch decreases Hum. by 5.26 to 89.64 and Ctx. by 5.99 to 89.73, resulting in less empathetic and personalized responses. Increasing beyond 4 branches (*e.g.*, Hum. of 96.35 at 10 branches) yields marginal improvements at the cost of significantly higher computational overhead. Thus, 4 branches optimally balance response quality and computational efficiency for psychological counseling tasks.



Metrics: (a) Coh., Flu., Con., Cpl., EmG. (b) Hum., Tec., Str., Ctx., Ovr.

Figure 4: Ablation Study for Multi-turn Rollout Rounds.



Metrics: (a) Coh., Flu., Con., Cpl., EmG. (b) Hum., Tec., Str., Ctx., Ovr.

Figure 5: Ablation Study for Single Rollout Branches.

Conclusion

In this work, we introduced PsyPARSE, a novel training-free framework addressing key challenges in LLM-based psychological counseling: data biases, limited personalization, and crucial anticipatory empathy deficits. PsyPARSE integrates multi-therapy RAG for personalized guidance and a pioneering multi-stage slow-thinking engine. Leveraging Multi-Turn Rollouts and Empathetic Response Optimization, this engine enables deliberate, foresightful, empathetic interactions, mimicking human counselors. Evaluated in comprehensive LLM-based simulation, PsyPARSE consistently demonstrated superior personalization and deeper empathy, outperforming fine-tuned and other training-free methods. A plug-and-play solution, PsyPARSE offers an efficient, adaptable, scalable paradigm for mental health support, bypassing fine-tuning's computational burden. Future work will explore robustness in real-world settings, expanding therapeutic scope.

References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Arslan, M.; Ghanem, H.; Munawar, S.; and Cruz, C. 2024. A Survey on RAG with LLMs. *Procedia Computer Science*, 246: 3781–3790.

Bai, J.; Bai, S.; Chu, Y.; Cui, Z.; Dang, K.; Deng, X.; Fan, Y.; Ge, W.; Han, Y.; Huang, F.; et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Beck, A.; Rush, A.; Shaw, B.; and Emery, G. 1979. Cognitive therapy of depression guilford press. *New York*.

Chen, Y.; Xing, X.; Lin, J.; Zheng, H.; Wang, Z.; Liu, Q.; and Xu, X. 2023. SoulChat: Improving LLMs' empathy, listening, and comfort abilities through fine-tuning with multi-turn empathy conversations. *arXiv preprint arXiv:2311.00273*.

Douze, M.; Guzhva, A.; Deng, C.; Johnson, J.; Szilvasy, G.; Mazaré, P.-E.; Lomeli, M.; Hosseini, L.; and Jégou, H. 2024. The faiss library. *arXiv preprint arXiv:2401.08281*.

Dryden, W. 2005. Rational emotive behavior therapy. *Encyclopedia of cognitive behavior therapy*, 321–324.

Grawe, K. 2004. *Psychological therapy*. Hogrefe Publishing GmbH.

Harris, R. 2006. Embracing your demons: An overview of acceptance and commitment therapy. *Psychotherapy in Australia*, 12(4): 70–6.

Hu, J.; Dong, T.; Gang, L.; Ma, H.; Zou, P.; Sun, X.; Guo, D.; Yang, X.; and Wang, M. 2024. Psycollm: Enhancing llm for psychological understanding and evaluation. *IEEE Transactions on Computational Social Systems*.

Issa, B.; Jasser, M. B.; Chua, H. N.; and Hamzah, M. 2023. A comparative study on embedding models for keyword extraction using keybert method. In *2023 IEEE 13th International Conference on System Engineering and Technology (ICSET)*, 40–45. IEEE.

Kim, J. S. 2008. Examining the effectiveness of solution-focused brief therapy: A meta-analysis. *Research on Social Work Practice*, 18(2): 107–116.

Koloski, B.; Pollak, S.; Škrlj, B.; and Martinc, M. 2022. Extending Neural Keyword Extraction with TF-IDF tagset matching. arXiv:2102.00472.

Lee, A.; Moon, S.; Jhon, M.; Kim, J.-W.; Kim, D.-K.; Kim, J. E.; Park, K.; and Jeon, E. 2024a. Comparative Study on the Performance of LLM-based Psychological Counseling Chatbots via Prompt Engineering Techniques. In *2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 7080–7082. IEEE.

Lee, S.; Kim, S.; Kim, M.; Kang, D.; Yang, D.; Kim, H.; Kang, M.; Jung, D.; Kim, M. H.; Lee, S.; et al. 2024b. Cactus: Towards psychological counseling conversations using cognitive behavioral theory. *arXiv preprint arXiv:2407.03103*.

Lee, Y. K.; Lee, I.; Shin, M.; Bae, S.; and Hahn, S. 2023. Chain of empathy: Enhancing empathetic response of large language models based on psychotherapy models. *arXiv preprint arXiv:2311.04915*.

Linehan, M. 2014. *DBT? Skills training manual*. Guilford Publications.

Liu, A.; Feng, B.; Xue, B.; Wang, B.; Wu, B.; Lu, C.; Zhao, C.; Deng, C.; Zhang, C.; Ruan, C.; et al. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Peng, J.-L.; Cheng, S.; Diau, E.; Shih, Y.-Y.; Chen, P.-H.; Lin, Y.-T.; and Chen, Y.-N. 2024. A survey of useful llm evaluation. *arXiv preprint arXiv:2406.00936*.

Qiu, H.; He, H.; Zhang, S.; Li, A.; and Lan, Z. 2023. Smile: Single-turn to multi-turn inclusive language expansion via chatgpt for mental health support. *arXiv preprint arXiv:2305.00450*.

Qiu, H.; Li, A.; Ma, L.; and Lan, Z. 2024. Psychat: A client-centric dialogue system for mental health support. In *2024 27th International Conference on Computer Supported Cooperative Work in Design (CSCWD)*, 2979–2984. IEEE.

Ravi, S.; and Larochelle, H. 2017. Optimization as a model for few-shot learning. In *International conference on learning representations*.

Robertson, S.; Zaragoza, H.; et al. 2009. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4): 333–389.

Santomauro, D. F.; Herrera, A. M. M.; Shadid, J.; Zheng, P.; Ashbaugh, C.; Pigott, D. M.; Abbafati, C.; Adolph, C.; Amlag, J. O.; Aravkin, A. Y.; et al. 2021. Global prevalence and burden of depressive and anxiety disorders in 204 countries and territories in 2020 due to the COVID-19 pandemic. *The Lancet*, 398(10312): 1700–1712.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Xia, F.; Chi, E.; Le, Q. V.; Zhou, D.; et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35: 24824–24837.

Xiang, S.; Zhang, A.; Cao, Y.; Yang, F.; and Chen, R. 2025. Beyond surface-level patterns: An essence-driven defense framework against jailbreak attacks in llms. In *Findings of the Association for Computational Linguistics: ACL 2025*, 14727–14742.

Xiang, S.; Zhang, T.; and Chen, R. 2025. ALRPHFS: Adversarially Learned Risk Patterns with Hierarchical Fast & Slow Reasoning for Robust Agent Defense. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, 19569–19587. Association for Computational Linguistics.

Xin Yan, D. X. 2023. MindChat: Psychological Large Language Model. https://github.com/X-D-Lab/MindChat.

Yang, A.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.; Yu, B.; Li, C.; Liu, D.; Huang, F.; Wei, H.; et al. 2024. Qwen2. 5 Technical Report. *arXiv e-prints*, arXiv–2412.

Zhang, C.; Li, R.; Tan, M.; Yang, M.; Zhu, J.; Yang, D.; Zhao, J.; Ye, G.; Li, C.; and Hu, X. 2024. Cpsycoun: A report-based multi-turn dialogue reconstruction and evaluation framework for chinese psychological counseling. *arXiv preprint arXiv:2405.16433*.

Zhang, X.; Zhao, J.; Yang, Z.; Zhong, Y.; Guan, S.; Cao, L.; and Wang, Y. 2025. UORA: Uniform Orthogonal Reinitialization Adaptation in Parameter Efficient Fine-Tuning of Large Models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 11709–11728. Association for Computational Linguistics.

Zhao, H.; Li, L.; Chen, S.; Kong, S.; Wang, J.; Huang, K.; Gu, T.; Wang, Y.; Jian, W.; Liang, D.; et al. 2024. ESC-Eval: Evaluating Emotion Support Conversations in Large Language Models. *arXiv preprint arXiv:2406.14952*.

Zhao, J.; Zhang, X.; Li, J.; Niu, J.; Hu, Y.; Min, E.; and Penn, G. 2025. Tiny Budgets, Big Gains: Parameter Placement Strategy in Parameter Super-Efficient Fine-Tuning. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 6326–6344.