MCTSr-Zero: Self-Reflective Psychological Counseling Dialogues Generation via Principles and Adaptive Exploration

Hao Lu[1], **Yanchi Gu**[1], **Haoyuan Huang**[1], **Yulin Zhou**[1], **Ningxin Zhu**[1*], **Chen Li**[1*]

[1]JianChengXingYun Technology Co., Ltd., Shenzhen, China
{luhao, zhuningxin, lichen}@jianchengxingyun.com

Abstract

The integration of Monte Carlo Tree Search (MCTS) with Large Language Models (LLMs) has demonstrated significant success in structured, problem-oriented tasks. However, applying these methods to open-ended dialogues, such as those in psychological counseling, presents unique challenges. Unlike tasks with objective correctness, success in therapeutic conversations depends on subjective factors like empathetic engagement, ethical adherence, and alignment with human preferences, for which strict "correctness" criteria are ill-defined. Existing result-oriented MCTS approaches can therefore produce misaligned responses. To address this, we introduce MCTSr-Zero, an MCTS framework designed for open-ended, human-centric dialogues. Its core innovation is "domain alignment", which shifts the MCTS search objective from predefined end-states towards conversational trajectories that conform to target domain principles (e.g., empathy in counseling). Furthermore, MCTSr-Zero incorporates "Regeneration" and "Meta-Prompt Adaptation" mechanisms to substantially broaden exploration by allowing the MCTS to consider fundamentally different initial dialogue strategies. We evaluate MCTSr-Zero in psychological counseling by generating multi-turn dialogue data, which is used to fine-tune an LLM, PsyLLM. We also introduce PsyEval, a benchmark for assessing multi-turn psychological counseling dialogues. Experiments demonstrate that PsyLLM achieves state-of-the-art performance on PsyEval and other relevant metrics, validating MCTSr-Zero's effectiveness in generating high-quality, principle-aligned conversational data for human-centric domains and addressing the LLM challenge of consistently adhering to complex psychological standards.

Code — https://github.com/JianChengXingYun/Mctsr-Zero

1 Introduction

The integration of Monte Carlo Tree Search (MCTS) with Large Language Models (LLMs) has recently achieved significant breakthroughs in structured, problem-oriented tasks such as mathematics (Zhang et al. 2024b,c; Guan et al. 2025; Wang et al. 2025; Chen et al. 2024; Wu et al. 2024). These methods leverage the planning capabilities of MCTS to guide the generative power of LLMs towards optimal solutions.

*Corresponding author

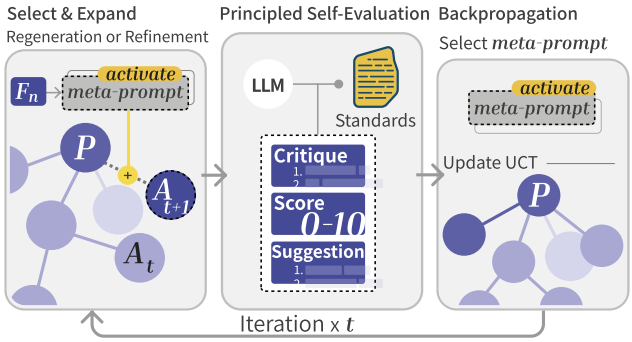

Figure 1: Iterative workflow of MCTSr-Zero: (1) **Select & Expand** responses using meta-prompts; (2) **Principled Self-Evaluation** against standards; (3) **Backpropagation** updating UCT & meta-prompts. Iterated t times.

On another front, recent LLM advancements (e.g., GPT-4 (OpenAI 2025)) have spurred their application in mental health, leading to specialized models like PsyChat (Qiu et al. 2024), CPsyCounX (Zhang et al. 2024a), Interactive Agents (Qiu and Lan 2024), and PsyDT (Xie et al. 2024). These models often rely on synthesized multi-turn dialogue datasets due to real-data scarcity. A challenge in this domain, however, is that LLMs often struggle to deeply understand and consistently adhere to the complex, abstract, and open-ended psychological standards or principles essential for effective counseling. Given the reliance on synthetic data, enhancing its quality by ensuring better alignment with these principles becomes a key research focus. This points towards the potential of employing search mechanisms—like those offered by MCTS—to actively discover replies more concordant with human preferences and established therapeutic guidelines.

While this potential to leverage MCTS for improved alignment in therapeutic dialogue is promising, applying these methods to open-ended dialogue, especially in domains like psychological counseling, is challenging. Unlike tasks with objective, verifiable answers, success in therapeutic conversations hinges on factors such as empathetic engagement, adherence to ethical guidelines, and alignment with human preferences, for which strict "correctness" cri-

teria are ill-defined. Consequently, existing result-oriented MCTS-based methods may produce responses misaligned with human expectations or domain-specific conversational goals.

To address the challenge of applying MCTS-enhanced LLMs to open-ended, human-centric dialogues, we introduce MCTSr-Zero. Its core innovation is domain alignment, shifting the search objective from predefined end-states towards conversational trajectories that conform to target domain principles (e.g., empathy in counseling). Furthermore, to broaden MCTS exploration, we propose mechanisms for "Regeneration" and "Meta-Prompt Adaptation." These allow the MCTS to explore fundamentally different initial dialogue strategies by modifying the guiding meta-prompt, thereby substantially expanding the search space beyond variations of a single initial approach.

We evaluate MCTSr-Zero in psychological counseling. Specifically, we collected topics from online mental health resources, organized them into case scenarios of psychological counseling, and used MCTSr-Zero to convert these case scenarios into multi-turn dialogue data. This data was used to fine-tune an open-source LLM, named PsyLLM. To address the need for standardized evaluation in this domain, we also developed PsyEval, a benchmark for assessing multi-turn psychological counseling dialogues. Experiments show that PsyLLM, developed using our approach, achieves state-of-the-art (SOTA) performance on PsyEval and other relevant metrics, demonstrating the effectiveness of MCTSr-Zero.

Our contributions are summarized as follows:

- We propose MCTSr-Zero, an MCTS framework incorporating domain alignment, Regeneration, and Meta-Prompt Adaptation techniques for improved search in open-ended dialogue generation.
- We construct PsyEval, a benchmark for the automated evaluation of multi-turn dialogues in psychological counseling, addressing a need for standardized assessment.
- We develop PsyLLM, an LLM for psychological counseling fine-tuned with MCTSr-Zero-generated data, which achieves SOTA performance, demonstrating our approach's effectiveness.

2 Related Work

2.1 MCTS-Enhanced LLMs

Monte Carlo Tree Search (MCTS) has proven highly effective across diverse complex problem domains, from multi-agent pathfinding (Pitanov et al. 2023) and train timetabling (Yang 2023) to SAT solving (Li et al. 2023) and robotics (Vagadia et al. 2024). Recently, MCTS has been integrated with Large Language Models (LLMs) to enhance their reasoning, particularly in structured tasks like mathematics (Chen et al. 2024; Zhang et al. 2024b,c; Guan et al. 2025; Wang et al. 2025), leveraging its planning capabilities to guide LLM generation. However, applying these MCTS-LLM methods to open-ended dialogue, especially in domains like psychological counseling, is challenging because success hinges on subjective factors like empathetic engagement and ethical adherence rather than objectively verifiable correctness, for which strict "correctness" criteria are ill-defined, potentially leading to responses misaligned with human expectations or domain-specific conversational goals. To address this, our work introduces MCTSr-Zero, which innovatively employs principle-guided domain alignment, shifting the MCTS search objective from predefined end-states towards conversational trajectories that conform to target domain principles.

2.2 LLMs in Mental Health Support

Recent LLM advancements (e.g., GPT-4 (OpenAI 2025)) have spurred their application in mental health Support, leading to specialized models like PsyChat (Qiu et al. 2024), CPsyCounX (Zhang et al. 2024a) and Interactive Agents (Qiu and Lan 2024), which often rely on synthesized multi-turn dialogue datasets due to real-data scarcity. The PsyDT framework (Xie et al. 2024) further advanced this by personalizing counseling styles. However, a critical challenge is that LLMs often struggle to deeply understand and consistently adhere to the complex, abstract, and open-ended psychological standards or principles essential for effective counseling. Our proposed method employs a search mechanism explicitly guided by these principles to actively discover replies that are more concordant with human preferences and established therapeutic guidelines. This focus on a guided search aims to enhance the LLM's alignment with nuanced therapeutic criteria and thereby improve the intrinsic quality of support provided.

3 Preliminary

This section establishes the necessary background on the foundational Monte Carlo Tree Search (MCTS) algorithm, its Upper Confidence Bound (UCB) selection strategy, and the general Monte Carlo Tree Self-Refine (MCTSr) adaptation, which are preliminary to understanding the proposed MCTSr-Zero algorithm detailed in Section 4.

3.1 Monte Carlo Tree Search (MCTS)

Monte Carlo Tree Search (MCTS) (Browne et al. 2012) is a heuristic search algorithm widely used in AI for decision-making in complex domains. It builds a search tree iteratively through four phases: **Selection** traverses the tree using strategies like UCB, **Expansion** adds new child nodes, **Simulation** evaluates outcomes, and **Backpropagation** updates node statistics.

The Upper Confidence Bound (UCB) formula, particularly UCB1, is commonly used in the Selection phase to balance exploration and exploitation. For a node a and its child j, the UCB value is calculated as:

$$UCB(a, j) = Q(j) + C\sqrt{\frac{2\ln N(a)}{N(j)}} \tag{1}$$

where $Q(j)$ is the estimated value of child node j, $N(j)$ and $N(a)$ are visit counts, and C controls the exploration-exploitation trade-off.

3.2 Dynamic Monte Carlo Tree Self-Refine (MCTSr)

Building upon MCTS, the Dynamic Monte Carlo Tree Self-Refine (MCTSr) algorithm adapts this framework for iterative text refinement using LLMs. In MCTSr, nodes represent versions of textual output, and edges represent refinement actions. MCTSr reinterprets the four MCTS phases for text refinement:

- **Selection:** An existing output node is selected for refinement, balancing exploration of different refinement paths (using a strategy like UCB) with exploiting paths that previously yielded high-quality outputs.
- **Expansion:** The selected node is expanded by prompting the LLM with refinement instructions, generating new child nodes with improved text.
- **Evaluation:** Instead of a rollout simulation, MCTSr directly evaluates the quality of a newly generated child node using an LLM-based self-reward function.
- **Backpropagation:** The evaluation outcome (reward/quality) from the evaluated node is propagated up the tree, updating statistics (visit counts and estimated values) for itself and its ancestors.

The "Dynamic" aspect highlights MCTSr's strategies for handling the potentially vast and stochastic nature of LLM-based generation and refinement. Specific notations and the detailed adaptation for psychological consultation dialogue in MCTSr-Zero are presented in Section 4.

4 MCTSr-Zero: A Framework for Principled, Large-Scale Dialogue Exploration

This section details MCTSr-Zero, an advanced Monte Carlo Tree Search algorithm tailored for generating and refining high-quality, iterative open-ended dialogue. Building upon MCTSr (Zhang et al. 2024b), MCTSr-Zero introduces two primary innovations significantly enhancing its capabilities:

- Constitutional AI Principles for Self-Alignment and Reflective Iteration: Inspired by Constitutional AI (CAI) (Bai et al. 2022), MCTSr-Zero empowers the language model with principled self-evaluation and self-refinement. Leveraging predefined psychological standards as a guiding "constitution," the AI critically assesses its dialogue, identifies areas for improvement, and uses this feedback for iterative refinement. This fosters reflective, self-aligned thinking within the LLM, enhancing dialogue quality and safety without direct human intervention in the core refinement loop.
- Vastly Expanded Search Space via Adaptive Exploration: Beyond standard MCTS branch deepening, MCTSr-Zero incorporates mechanisms for exploring a significantly larger, higher-dimensional search space. A UCB-driven selection criterion can choose to refine an existing response or regenerate a new initial response. Coupled with dynamic Meta-Prompt Adaptation, informed by self-evaluation feedback from recently generated initial responses, the algorithm strategically explores fundamentally different initial dialogue strategies. This dual mode—deepening paths via refinement and broadening via new starting points—results in an exponentially larger potential search space than methods limited to variations from a fixed initial strategy.

Integrating these innovations allows MCTSr-Zero to not only iteratively improve dialogue quality but also learn and adapt its fundamental generation strategy, guided by explicit principles and empirical self-evaluation. The conceptual structure, illustrating the exploration tree and information flow, is depicted in Figure 2.

The main workflow operates in iterative cycles, driving this principled, large-scale exploration through the following stages:

- Initialization: Setup the root node representing the query and an initial meta-prompt.
- Selection: Choose a node (either the root or an existing answer node) based on its Upper Confidence Bound for Trees (UCT) value.
- Action:
  - If the root is selected: A candidate meta-prompt is generated and used to produce a new initial response.
  - If an answer node is selected: That answer undergoes reflective self-refinement.
- Evaluation: The newly generated or refined response is assessed using Principled Self-Evaluation, yielding a score and feedback.
- Update and Adaptation: The score is backpropagated to update node Q-values. If a new response from a candidate meta-prompt improves the root's value, the active meta-prompt is updated to the candidate. UCT values are then recomputed for the next selection.

The algorithm iterates until a termination condition is met.

4.1 Initialization

MCTSr-Zero starts with the node P as the root, representing user query P. An initial system prompt named meta prompt m_0 is set. The language model generates $\mathcal{M}$ initial children $A_0 \dots A_k \sim \mathcal{M}(P \parallel m_0)$, forming $\mathcal{A}_{initial}$. The first generated node, A_0, undergoes an immediate evaluation using the Principled Self-Evaluation (Section 4.5). P's initial Q value is then set equal to the reward value obtained from the Principled Self-Evaluation of A_0.

4.2 Selection and Adaptive Exploration Strategy

In each cycle, UCT (Section 4.7) selects node $a \in \mathcal{C}$ (answer nodes $\mathcal{A}$ and P), embodying the expanded exploration strategy:

This dynamic UCT choice, $a = \arg\max_{a \in \mathcal{C}} UCT(\mathcal{C})$, balances deepening and broadening the search space.

- If $a \in \mathcal{A}$: Self-Refine action on a, deepening a path.
- If $a = P$: Regeneration action, broadening search via new initial strategy (Meta-Prompt Adaptation).

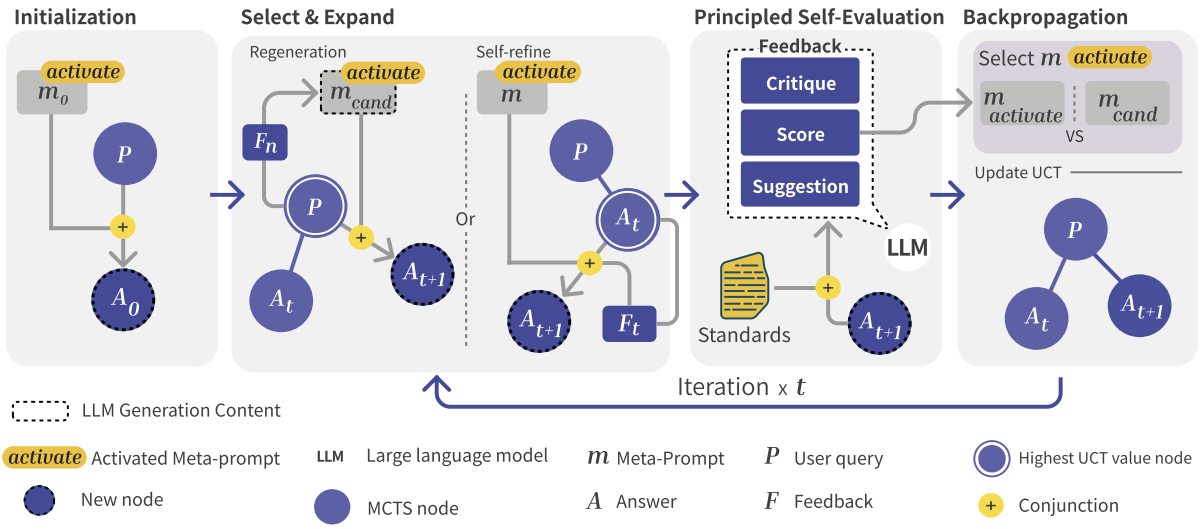

Figure 2: The operational workflow of MCTSr-Zero. (1) **Initialization**: The meta-prompt m_0 is activated to generate initial responses A_0 for the user query P. (2) **Select & Expand**: The system uses the UCT value to either trigger *Regeneration*—using a candidate meta-prompt m_{cand} as the basis for generating a new answer node A_{t+1}—or trigger *Refinement* to further improve an existing answer node. (3) **Principled Self-Evaluation**: The new answer is evaluated against predefined standards, producing a critique, score, and actionable suggestions. (4) **Backpropagation**: Evaluation results are propagated to update UCT scores and guide meta-prompt selection.

4.3 Meta-Prompt Adaptation and Search Space Expansion

A pivotal mechanism for exploring the higher-order search space is Meta-Prompt Adaptation, triggered when P is selected. It allows learning and adjusting the fundamental approach to generating initial responses.

When P is selected at iteration $t+1$, signaling an intent to generate a new initial response, a candidate meta-prompt m_{cand} is synthesized. This synthesis uses the current active meta-prompt $m_{activate}$ and the self-evaluation feedback $\mathcal{F}_n$ (critique and suggestions) from a relevant, recently evaluated child node $A_n \in \mathcal{A}_{initial}$ of P:

$$m_{cand} \leftarrow \mathcal{M}(m_{activate} \parallel \mathcal{F}_n) \tag{2}$$

This feedback provides targeted insights from recent attempts to generate good initial responses under P. Using m_{cand}, a new initial response $A_{t+1} = \mathcal{M}(P \parallel m_{cand})$ is generated. This new response A_{t+1} is added to $\mathcal{A}_{initial}$

After A_{t+1} is generated and undergoes Principled Self-Evaluation to obtain its quality score $Q(A_{t+1})$, the active meta-prompt $m_{activate}$ is conditionally updated. If the quality of the new response $Q(A_{t+1})$ is greater than the current quality of the parent node P, $Q(P)$ (which represents the average quality of P's children before A_{t+1}'s inclusion and subsequent $Q(P)$ update), then $m_{activate}$ is updated to m_{cand}:

$$m_{activate} = \begin{cases} m_{cand} & if Q(A_{t+1}) >= Q(P) \\ m_{activate} & otherwise \end{cases} \tag{3}$$

This mechanism allows the system to autonomously learn and switch to better initial generation strategies if they prove more effective than the current average performance of initial responses.

Unlike standard MCTS from a fixed m_0, MCTSr-Zero explores from a growing $\mathcal{A}_{initial}$, elevating the search to different distributions $\mathcal{P}(\mathcal{A}^{(t)}_{initial} | P, m^{(t)}_{activate})$. This constitutes exploring a higher-order search space.

4.4 Reflective Self-Refine

When a response node a (not P) is selected, Reflective Self-Refine transforms a to a'. Inspired by Self-Refine (Madaan et al. 2023), this uses iterative interaction with the LLM, guided by feedback from a's Principled Self-Evaluation. The model uses concrete, standards-based critique and suggestions $\mathcal{F}$ and meta prompt $m_{activate}$ as explicit guidance to produce $A'_{t+1} = \mathcal{M}(A_t \parallel \mathcal{F}_t \parallel m_{activate})$. This enables targeted improvement based on AI's principled assessment and Self-Refine techniques.

4.5 Principled Self-Evaluation: Applying the Constitution

Principled Self-Evaluation is the cornerstone, providing crucial feedback for refinement and adaptation. It rigorously assesses new/refined response a. Inspired by Constitutional AI (Bai et al. 2022), which uses principles for self-improvement, we use 16 psychological standards as our AI's "constitution."

For response a, the LLM performs structured evaluation guided by standards:

- Critique based on Constitution: Analyzes a against 16 standards, identifying strengths/weaknesses relative to

these principles. This critique guides subsequent steps towards standard fulfillment.

- Scoring (0-10): Assigns a reward score based on critique and adherence to standards.
- Actionable Suggestions: Provides specific suggestions for improving a to better meet standards. These are used as guidance for Refinement and the feedback for Meta-Prompt Adaptation as described in Eq. 2.

This embodies Constitutional AI: AI uses principles to evaluate output and generate feedback for self-improvement, fostering self-alignment. Evaluation adheres strictly to standards and is sampled multiple times for robustness. $Q(a)$ from sampled scores R_a uses:

$$Q(a) = \frac{1}{2}\left(\min R_a + \frac{1}{|R_a|}\Sigma_{i=1}^{|R_a|} R_a^i\right) \tag{4}$$

This balances average quality and low-score robustness.

4.6 Backpropagation

Following Principled Self-Evaluation, Q value and insights propagate upwards.

- For evaluated response a, $Q(a)$ is set using Eq. 4.
- For parent answer node p of a, $Q'(p)$ updates recursively:

$$Q'(p) = \frac{1}{2}\left(Q(p) + \max_{c \in Children} Q(c)\right) \tag{5}$$

- For user query node P, $Q(P)$ is the average of its children's Q values:

$$Q(P) = \frac{1}{|\mathcal{A}_{initial}|} \sum_{a \in \mathcal{A}_{initial}} Q(a) \tag{6}$$

  This $Q(P)$ value is used in the condition for updating the active meta-prompt (Eq. 3).

Backpropagation ensures performance signals inform future decisions.

4.7 Update UCT and Guiding Large-Scale Exploration

After Backpropagation, UCT values recompute for selectable nodes $\mathcal{C}$ answer nodes and P). UCT is computed using UCB:

$$UCT_s = Q(s) + c\sqrt{\frac{\ln N(Parent(s)) + 1}{N(s) + \epsilon}} \tag{7}$$

where $Q(s)$ is quality (Eq. 4 or Eq. 6), $N(s)$ is visit count, $N(Parent(s))$ is parent visit count (adjusted for P), c balances exploration/exploitation, ϵ prevents division by zero. $a \in \mathcal{C}$ with highest UCT is selected. If a is answer node, high UCT signals refining this path. If a is P, high UCT signals exploring a new initial strategy via Regeneration/Meta-Prompt Adaptation. This UCT-driven selection intelligently navigates the massively expanded search space, balancing deepening paths with exploring new starting points, guided by principled self-evaluation feedback.

4.8 Termination Function

Iteration continues until condition T is met (e.g., budget, depth, quality threshold, stagnation). Final output is the dialogue response node with the highest Q value.

5 PsyEval Benchmark: Evaluating AI Empathy in Psychological Support

The growing use of AI for psychological support demands evaluation beyond standard metrics. Conventional benchmarks often fail to assess crucial competencies like accurate empathy, re-framing, and meaningful questioning. To bridge this gap, we introduce PsyEval—a specialized, multi-dimensional benchmark designed to gauge AI competence in simulated counseling dialogues grounded in realistic scenarios.

5.1 Systematic Scenario and Data Generation

PsyEval creates diverse psychological scenarios by synthesizing detailed case reports, rather than relying on existing human data. Based on these reports, a Language Model (LLM) simulates a multi-turn counseling dialogue. An independent AI Judge then evaluates this dialogue against 16 predefined criteria. This systematic process enables rigorous testing of AI capabilities across various psychological contexts. Details on scenarios and criteria are in Appendix A and B, which are provided in the supplementary material.

5.2 Comprehensive Multi-Dimensional Evaluation

The core of PsyEval is its novel 16-dimension evaluation framework, which assesses an AI therapist's ability to provide empathic support in multi-turn interactions. Adopting a third-party observational perspective similar to the ESHCC benchmark (Concannon and Tomalin 2024), it focuses on an AI's observable expressions of counseling skill. The rubric uses inferential language (e.g., "the system seems to…") to capture a neutral observer's perception of the interaction quality. The 16 dimensions synthesize insights from established frameworks, including the Therapist Empathy Scale (TES) (Decker et al. 2014), ESHCC, Motivational Interviewing (MI) (Bolton et al. 2021), and Person-centered therapy (Rogers 2007). We optimized existing dimensions and integrated six new ones crucial for psychological counseling: Dialogical Logical Consistency, Conversational Continuity, Resistance Handling, Ethics/Prosocial Guidance, Summarizing, and Dialogue Pacing/Process Attunement. We also redefine the "Fallacy Avoidance" dimension to evaluate hallucination control—the AI's ability to remain coherent and factually grounded while maintaining a consistent persona. Unlike clinical-only scales like TES, our benchmark has broader applicability, evaluating AI in both therapeutic and everyday emotional support contexts. This "dual role" enhances its ecological validity, assessing AI as both a structured therapist and an empathetic companion.

5.3 AI-based Judging Mechanism

PsyEval employs an independent AI Judge for evaluation, configured with our 16-dimension rubric. This approach en-

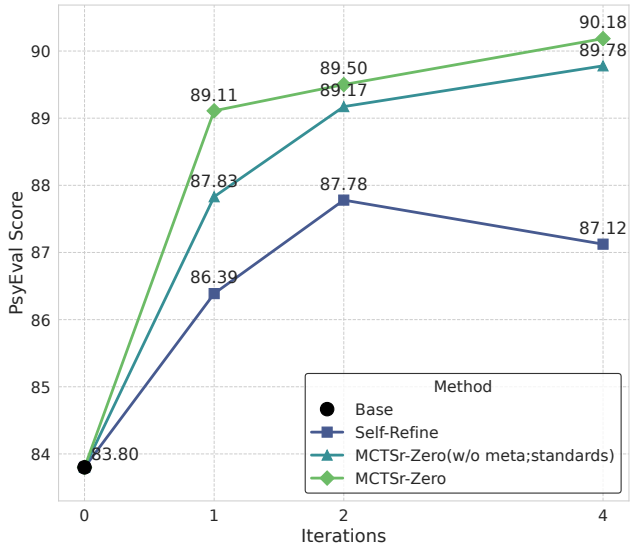

Figure 3: PsyEval scores for gpt-4.1-mini with four different refinement methods across iterations. MCTSr-Zero demonstrates the highest performance.

sures scalable, efficient, and consistent scoring across numerous dialogues, minimizing the human rater variability common in subjective assessments. Details of the AI judge setup are in Appendix C. These standards, developed by human experts and tailored for LLM comprehension, are designed to ensure high alignment between the AI's evaluation and human judgment, thereby minimizing potential evaluation bias.

6 Experiment

This section presents the experimental evaluation of various language models on the PsyEval Benchmark. We report results from two key experiments: (1) a comprehensive benchmark comparison of our PsyLLM variants against leading baseline models, including other psychological domain-specific models, evaluated using PsyEval, and (2) an ablation study demonstrating the impact of iterative refinement methods aligned with principles of frameworks like MCTSr-Zero.

6.1 MCTSr-Zero-Psy

To validate the effectiveness of our proposed dialogue reconstruction method, we generated dialogues using MCTSr-Zero. We denote this collection as MCTSr-Zero-Psy, which comprises 4,000 multi-turn counseling dialogues. These dialogues cover 16 distinct categories of psychological distress (as detailed in Section 5.2), with each category containing approximately 250 entries, and an average of 20 turns per dialogue.

6.2 PsyLLM

Our PsyLLM model, developed in Large (based on GLM-4-32B-0414) and Mini (based on GLM-4-9B-0414) variants, underwent a two-stage training process on 4 NVIDIA A800 GPUs. We used MCTSr-Zero-Psy as our training data. First, for supervised fine-tuning (SFT), each respective pre-trained GLM-4 base model was trained for 2 epochs with a learning rate of 1×10^{-4}, a 0.1 linear warmup, a per-device batch size of 1, and the AdamW (Loshchilov and Hutter 2019) optimizer. Subsequently, the SFT-trained models underwent SimPO (Meng, Xia, and Chen 2024) alignment for 3 epochs, employing a reduced learning rate of 5×10^{-7}, maintaining a 0.1 linear warmup and a per-device batch size of 1 with the AdamW optimizer, ultimately yielding the final PsyLLM models.

6.3 Experimental Setup

Our evaluation utilizes the PsyEval Benchmark, which simulates psychological counseling dialogues for 64 unique case scenarios. An AI Judge evaluates responses against **16** specific psychological consultation criteria to calculate a total score. Further details on the AI Judge are in Appendix C.

We conduct two main experiments:

- **Main Benchmark Comparison on PsyEval:** We comprehensively evaluate a wide array of models, including leading commercial, general-purpose open-source, and other psychological domain-specific models. This includes our PsyLLM-Large-250519 (based on GLM-4-32B-0414(Team et al. 2024)) and PsyLLM-Mini-250519 (based on GLM-4-9B-0414) models, which were developed using techniques aligned with MCTSr-Zero's principles of iterative refinement. Results are presented in Table 1.
- **Ablation Study on Iterative Refinement Methods:** We investigate the effectiveness of iterative refinement on a `gpt-4.1-mini` base model. We apply `Self-Refine`, `MCTSr-Zero(w/o meta;standards)` (a variant using `Self-Refine` for node refinement instead of standard-based evaluation), and the full `MCTSr-Zero` framework. These methods are tested across 1, 2, and 4 iterations and compared to a non-refined baseline (0 iterations), with results shown in Figure 3.

6.4 Analysis of Benchmark Results

Table 1 details the performance of all models on the PsyEval Benchmark, as assessed by the AI Judge across all 16 criteria. The data indicates that our proposed models, PsyLLM-Large and PsyLLM-Mini, achieved the highest overall scores of 90.93 and 90.72, respectively. This demonstrates a discernible advantage over other evaluated models, including the next-highest score of 88.89.

A deeper analysis reveals that this strong performance is not limited to a single capability but reflects a balanced and holistic profile. For instance, the PsyLLM variants achieved high scores in empathetic and human-centered communication (consolidated as 'ESHCC REVISED' in the table for brevity) while also excelling in maintaining logical consistency, conversational continuity, and effectively handling user resistance.

Type	Model Name	Total Score	ESHCC REVISED	DLC	CC	RH	Sum.	EPG	DPPA
Ours	**PsyLLM-Large-250519**	**90.93**	**54.53**	**9.16**	4.57	4.56	**4.47**	4.53	**4.55**
	PsyLLM-Mini-250519	90.72	54.46	9.15	**4.58**	**4.57**	4.43	4.47	4.51
Other	claude-3-7-sonnet-20250219	88.89	53.13	9.03	4.51	4.44	4.28	**4.56**	4.49
	gemini-2.5-pro-exp-03-25	88.62	53.01	9.06	4.53	4.48	4.33	4.34	4.36
	gemini-2.5-flash-preview-04-17	88.07	52.59	8.94	4.50	4.47	4.27	4.39	4.39
	gpt-4.1	85.65	50.87	8.77	4.44	4.44	4.04	4.32	4.38
	gpt-4.1-mini	83.80	49.82	8.69	4.38	4.19	3.94	4.18	4.21
	gpt-4o-2024-11-20	82.31	48.71	8.52	4.28	4.18	3.87	4.25	4.24
	Qwen3-32B-w/o Reasoning	81.40	48.66	8.45	4.28	4.16	3.78	4.06	4.10
	GLM-4-32B-0414	80.92	48.00	8.39	4.25	3.97	3.82	4.05	4.13
	gpt-4.1-nano	80.72	47.58	8.52	4.25	3.96	3.83	4.14	4.10
	qwen-max-2025-01-25	79.59	47.11	8.32	4.22	3.99	3.65	4.14	3.93
	gpt-4o-mini-2024-07-18	78.76	46.48	8.21	4.25	3.82	3.59	4.06	4.10
	doubao-1-5-pro-32k-250115	78.71	46.80	8.19	4.17	3.89	3.66	3.97	3.90
	Qwen2.5-72B-Instruct	76.41	45.53	7.97	3.98	3.61	3.48	3.92	3.82
	GLM-4-9B-0414	75.74	44.99	7.94	3.96	3.76	3.41	3.91	3.79
	Qwen2.5-32B-Instruct	75.70	44.54	8.08	4.02	3.67	3.47	4.03	3.91
	doubao-1-5-lite-32k-250115	74.99	44.72	7.88	4.00	3.55	3.32	3.91	3.59
Domain	simpsybot_D	77.92	45.87	8.36	4.15	3.74	3.60	4.13	3.92
	SoulChat2_0-Qwen2-7B	77.38	45.55	8.08	4.20	3.84	3.56	4.10	3.98
	Xinyuan-LLM-14B-0428	76.41	45.26	8.02	4.02	3.72	3.50	4.07	3.78
	CPsyCounX	66.00	39.99	6.76	3.37	3.24	3.01	3.82	3.31

Table 1: Results on the PsyEval Benchmark. Scores for each criterion reflect model performance in simulated psychological counseling dialogues. The criteria are abbreviated as follows: ESHCC-R (Revised Empathic Support, Human Connection, and Care), DLC (Dialogical Logical Consistency), CC (Conversational Continuity), RH (Resistance Handling), Sum. (Summarizing), EPG (Ethics and Prosocial Guidance), and DPPA (Dialogue Pacing, Process, and Attunement). The best result in each column is highlighted in **bold**.

We attribute this robust and well-rounded performance to the direct alignment between our development methodology and the evaluation benchmark. The performance disparity seen in the results stems from a core design principle: the PsyEval criteria are derived from the same 16 psychological standards that guide the iterative refinement within our proposed framework, MCTSr-Zero. The high scores of the PsyLLM variants underscore the value of a synergistic approach that combines domain-specific evaluation (PsyEval) with a tailored development framework (MCTSr-Zero). This deep alignment is fundamental to building effective and responsible AI for psychological support.

6.5 Ablation Study

Figure 3 shows our ablation study results, demonstrating that iterative refinement methods substantially improve the performance of a base model (gpt-4.1-mini) on PsyEval.

The baseline model's score of 83.60 improves significantly with just one iteration of Self-Refine (to 86.39) or an MCTSr-Zero variant (over 87). Performance generally increases with more iterations. The full `MCTSr-Zero` framework consistently outperforms simpler methods and its ablated variant, reaching a peak score of 90.18 after 4 iterations. This study validates that iterative processes and strategic search—core mechanisms of frameworks like MCTSr-Zero—effectively enhance an AI's capabilities for psychological support by leveraging self-reflection and evaluation against explicit standards.

7 Conclusion

We introduced MCTSr-Zero, an advanced Monte Carlo Tree Search algorithm that uses dynamic meta-prompts and principles-based self-evaluation to generate high-quality dialogues for training PsyLLM, our specialized language model for psychological counseling.

To assess such models, we also developed the PsyEval Benchmark, which uses an AI Judge to evaluate dialogues against 16 psychological criteria. Our results show that PsyLLM models, developed using MCTSr-Zero principles, significantly outperform leading general and domain-specific models. This underscores the value of our specialized framework and benchmark for the nuanced demands of psychological consultation.

Key limitations include MCTSr-Zero's computational cost and potential biases in PsyEval's AI Judge. Future work will focus on improving MCTSr-Zero's search efficiency and refinement techniques, while simultaneously advancing PsyEval by mitigating bias, expanding its scenarios, and integrating human evaluation.

References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2024. GPT-4 Technical Report. arXiv:2303.08774.

Anthropic. 2025. Claude 3.7 Sonnet and Claude Code. https://www.anthropic.com/claude/sonnet. Accessed: 2025-02-24.

Bai, Y.; Kadavath, S.; Kundu, S.; Askell, A.; Kernion, J.; Jones, A.; Chen, A.; Goldie, A.; Mirhoseini, A.; McKinnon, C.; Chen, C.; Olsson, C.; et al. 2022. Constitutional AI: Harmlessness from AI Feedback. arXiv:2212.08073.

Bolton, K. W.; Hall, J. C.; Lehmann, P.; et al. 2021. *Theoretical perspectives for direct social work practice: A generalist-eclectic approach*. Springer Publishing Company.

Browne, C.; Powley, E. J.; Whitehouse, D.; Lucas, S. M. M.; Cowling, P. I.; Rohlfshagen, P.; Tavener, S.; Liebana, D. P.; Samothrakis, S.; and Colton, S. 2012. A Survey of Monte Carlo Tree Search Methods. *IEEE Transactions on Computational Intelligence and AI in Games*, 4: 1–43.

bytedance. 2025. Doubao-1.5-pro. https://seed.bytedance.com/zh/special/doubao_1_5_pro. Accessed: 2025-01-22.

Chen, G.; Liao, M.; Li, C.; and Fan, K. 2024. AlphaMath Almost Zero: Process Supervision without Process. arXiv:2405.03553.

Concannon, S.; and Tomalin, M. 2024. Measuring perceived empathy in dialogue systems. *Ai & Society*, 39(5): 2233–2247.

Cylingo. 2025. Xinyuan-LLM-14B-0428. https://huggingface.co/Cylingo/Xinyuan-LLM-14B-0428. Accessed: 2025-04-28.

Decker, S. E.; Nich, C.; Carroll, K. M.; and Martino, S. 2014. Development of the therapist empathy scale. *Behavioural and cognitive psychotherapy*, 42(3): 339–354.

Google. 2025. Gemini 2.5 Pro Preview Model Card. https://storage.googleapis.com/model-cards/documents/gemini-2.5-pro-preview.pdf. Accessed: 2025-05-09.

Guan, X.; Zhang, L. L.; Liu, Y.; Shang, N.; Sun, Y.; Zhu, Y.; Yang, F.; and Yang, M. 2025. rStar-Math: Small LLMs Can Master Math Reasoning with Self-Evolved Deep Thinking. arXiv:2501.04519.

Li, A.; Han, C.; Guo, T.; Li, H.; and Li, B. 2023. General Method for Solving Four Types of SAT Problems. arXiv:2312.16423.

Loshchilov, I.; and Hutter, F. 2019. Decoupled Weight Decay Regularization. arXiv:1711.05101.

Madaan, A.; Tandon, N.; Gupta, P.; Hallinan, S.; Gao, L.; Wiegreffe, S.; Alon, U.; Dziri, N.; Prabhumoye, S.; Yang, Y.; Gupta, S.; Majumder, B. P.; Hermann, K.; Welleck, S.; Yazdanbakhsh, A.; and Clark, P. 2023. Self-Refine: Iterative Refinement with Self-Feedback. arXiv:2303.17651.

Meng, Y.; Xia, M.; and Chen, D. 2024. SimPO: Simple Preference Optimization with a Reference-Free Reward. In Globerson, A.; Mackey, L.; Belgrave, D.; Fan, A.; Paquet, U.; Tomczak, J.; and Zhang, C., eds., *Advances in Neural Information Processing Systems*, volume 37, 124198–124235. Curran Associates, Inc.

OpenAI. 2025. Introducing GPT-4.1 in the API. https://openai.com/index/gpt-4-1/. Accessed: 2025-04-14.

Pitanov, Y.; Skrynnik, A.; Andreychuk, A.; Yakovlev, K.; and Panov, A. 2023. Monte-carlo tree search for multi-agent pathfinding: Preliminary results. In *International Conference on Hybrid Artificial Intelligence Systems*, 649–660. Springer.

Qiu, H.; and Lan, Z. 2024. Interactive Agents: Simulating Counselor-Client Psychological Counseling via Role-Playing LLM-to-LLM Interactions. arXiv:2408.15787.

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.

Qwen; :; Yang, A.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.; Yu, B.; Li, C.; Liu, D.; Huang, F.; Wei, H.; Lin, H.; Yang, J.; et al. 2025. Qwen2.5 Technical Report. arXiv:2412.15115.

Rogers, C. 2007. *Counseling and Psychotherapy*. Read Books. ISBN 9781406760873.

Team, G.; Zeng, A.; Xu, B.; Wang, B.; Zhang, C.; Yin, D.; Rojas, D.; Feng, G.; Zhao, H.; Lai, H.; et al. 2024. ChatGLM: A Family of Large Language Models from GLM-130B to GLM-4 All Tools. arXiv:2406.12793.

Vagadia, H.; Chopra, M.; Barnawal, A.; Banerjee, T.; Tuli, S.; Chakraborty, S.; and Paul, R. 2024. PhyPlan: Compositional and Adaptive Physical Task Reasoning with Physics-Informed Skill Networks for Robot Manipulators. arXiv:2402.15767.

Wang, Y.; Ji, P.; Yang, C.; Li, K.; Hu, M.; Li, J.; and Sartoretti, G. 2025. MCTS-Judge: Test-Time Scaling in LLM-as-a-Judge for Code Correctness Evaluation. arXiv:2502.12468.

Wu, J.; Feng, M.; Zhang, S.; Che, F.; Wen, Z.; and Tao, J. 2024. Beyond Examples: High-level Automated Reasoning Paradigm in In-Context Learning via MCTS. arXiv:2411.18478.

Xie, H.; Chen, Y.; Xing, X.; Lin, J.; and Xu, X. 2024. PsyDT: Using LLMs to Construct the Digital Twin of Psychological Counselor with Personalized Counseling Style for Psychological Counseling. arXiv:2412.13660.

Xu, H. 2023. No Train Still Gain. Unleash Mathematical Reasoning of Large Language Models with Monte Carlo Tree Search Guided by Energy Function. arXiv:2309.03224.

Yang, A.; Li, A.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.; Yu, B.; Gao, C.; Huang, C.; Lv, C.; Zheng, C.; Liu, D.; et al. 2025. Qwen3 Technical Report. arXiv:2505.09388.

Yang, F. 2023. An Integrated Framework Integrating Monte Carlo Tree Search and Supervised Learning for Train Timetabling Problem. arXiv:2311.00971.

Zhang, C.; Li, R.; Tan, M.; Yang, M.; Zhu, J.; Yang, D.; Zhao, J.; Ye, G.; Li, C.; and Hu, X. 2024a. CPsyCoun: A Report-based Multi-turn Dialogue Reconstruction and Evaluation Framework for Chinese Psychological Counseling. arXiv:2405.16433.

Zhang, D.; Huang, X.; Zhou, D.; Li, Y.; and Ouyang, W. 2024b. Accessing GPT-4 level Mathematical Olympiad Solutions via Monte Carlo Tree Self-refine with LLaMa-3 8B. arXiv:2406.07394.

Zhang, D.; Wu, J.; Lei, J.; Che, T.; Li, J.; Xie, T.; Huang, X.; Zhang, S.; Pavone, M.; Li, Y.; Ouyang, W.; and Zhou, D. 2024c. LLaMA-Berry: Pairwise Optimization for O1-like Olympiad-Level Mathematical Reasoning. arXiv:2410.02884.