PRMB: A Benchmark for Evaluating Reward Models in Long-Horizon Psychological Counseling Dialogues with Negative Emotional Triggers

Anonymous ACL submission

Abstract

Reward models (RMs) are central to large language model (LLM) alignment, yet their reliability in emotionally sensitive long-horizon interactions remains poorly understood. Existing RM evaluations mainly focus on short-context, general-domain preferences while overlooking subtle emotional harms that emerge over multi-session conversations. In this work, we present a comprehensive evaluation of RMs in long-horizon psychological counseling dialogues. We show that state-of-the-art RMs, despite strong performance on standard benchmarks, struggle to distinguish beneficial counseling responses from emotionally harmful yet preference-aligned alternatives. To support rigorous evaluation, we introduce PRMB, a benchmark built from validated CBT counseling conversations with 13,893 samples spanning six-session interactions. PRMB evaluates whether RMs can detect subtle emotional harms under extended conversational contexts through pairwise ranking and Best-of-N evaluation. Experiments across diverse RMs show that strong performance on existing RM benchmarks does not reliably transfer to counseling settings. Furthermore, PRMB performance correlates with downstream generation quality, indicating strong consistency between benchmark and generation performance. These findings establish PRMB as a challenging benchmark for developing safer and more emotionally aware RMs.

1 Introduction

Large language models (LLMs) are increasingly adopted in emotionally sensitive applications, including emotional support and psychotherapy counseling dialogues (Qiu et al., 2024; Chen et al., 2023b; He et al., 2025; Sharma et al., 2024). Among these, Cognitive Behavioral Therapy (CBT) provides a structured and goal-oriented counseling framework, making it a promising setting for LLM-based counseling systems (Lee et al., 2024).

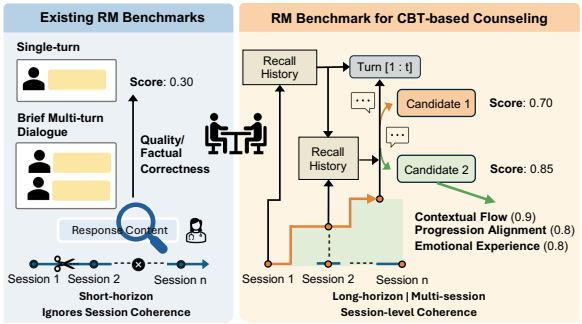

Figure 1: Comparison of reward model evaluation paradigms. CBT-based counseling requires maintaining session-level coherence and emotionally appropriate behavior across extended interactions.

Aligning LLMs with appropriate counseling behaviors in such scenarios commonly relies on reward models (RMs), which guide models toward human-preferred responses through reinforcement learning (Zhang et al., 2025a; Wang et al., 2025).

However, existing RM evaluations predominantly focus on short-context, general-domain preference judgments (Lambert et al., 2025; Zhou et al., 2024; Malik et al., 2025), implicitly assuming that strong performance on these benchmarks reflects alignment quality in real-world applications. This assumption becomes questionable in process-oriented scenarios such as psychological counseling, where responses must remain coherent, context-aware, and emotionally appropriate over extended interactions. In CBT-based counseling, models are expected to maintain consistency across sessions while adapting to evolving client states (Persons, 2012; Hayes and Hofmann, 2021). As illustrated in Figure 1, alignment in such settings depends not only on local response quality, but also on maintaining emotionally safe and context-sensitive behavior throughout the interaction trajectory. As a result, responses that appear appropriate in isolated evaluations may still accumulate subtle emotional harms over extended interactions (Lee et al., 2026; Tang et al., 2026).

Moreover, current RM benchmarks largely overlook subtle emotional harms that may emerge during extended interactions. Even when responses are appropriate, minor deviations in tone or phrasing can elicit negative emotional reactions and undermine the counseling process (Vybíral et al., 2024). In CBT settings, avoiding such subtle harms is essential for maintaining emotional regulation and a stable therapeutic alliance (Beck, 2020). This highlights a key limitation of existing preference-based RM evaluations: they do not adequately assess whether RMs can recognize subtle emotional harms in emotionally sensitive conversations.

To investigate this limitation, we introduce **PRMB**, a benchmark for evaluating RMs in multi-session counseling dialogues with a focus on subtle emotional harms. PRMB is constructed from publicly available CBT cases, and incorporates 21 categories of emotionally salient triggers to create challenging pairwise ranking and Best-of-N (BoN) evaluation instances. To better reflect realistic long-horizon interactions, we integrate summarized historical context across sessions with a progressive summarization framework. The resulting benchmark contains 13,893 samples generated by ten representative LLMs. Using PRMB, we conduct a comprehensive evaluation of state-of-the-art RMs. We find that, while these RMs perform well on existing RM benchmarks, their effectiveness drops in counseling settings, particularly when distinguishing between subtly different responses involving emotional harms. Furthermore, PRMB rankings correlate with downstream generation quality, indicating strong consistency between our benchmark and generation performance. Our findings suggest that current preference-based alignment pipelines remain insufficient for ensuring emotional safety in long-horizon interactions. The main contributions of this paper are as follows:

- We introduce **PRMB**, the first benchmark for evaluating reward models in long-horizon psychological counseling dialogues, with a focus on subtle emotional harms.
- We conduct comprehensive analysis on both discriminative and generative reward models across diverse evaluation settings and negative emotional experience categories.
- We show that PRMB rankings correlate with downstream response quality, offering insights into the reliability of reward models in mental health applications.

2 Related Work

Simulated Psychological Dialogue Large language models (LLMs) have advanced therapy-oriented dialogue systems by enabling more coherent and emotionally supportive interactions (Liu et al., 2023; Chen et al., 2023c; Jo et al., 2023; Wei et al., 2024). Subsequent studies have incorporated domain-specific counseling knowledge and structured therapeutic strategies to improve psychological support quality (Hsu et al., 2023; Chen et al., 2023a; Na, 2024). More recently, reinforcement learning (RL) has been introduced to align conversational agents with therapeutic objectives (Zhang et al., 2025b; Wang et al., 2025). Affect-aware reward signals have been introduced to optimize empathetic behavior and perceived support quality, shifting the optimization target from surface-level text similarity to higher-level psychological outcomes. Consequently, reward models (RMs) have become key components for behavioral alignment in counseling dialogue systems.

Reward Model Reward models (RMs) are central to LLM alignment, providing preference signals that guide models toward human-preferred responses (Ouyang et al., 2022; Askell et al., 2021; Bai et al., 2022). They can be broadly categorized into discriminative and generative RMs (Zhang et al., 2024; Zhou et al., 2024). Discriminative RMs are typically trained on human preference datasets to output scalar scores that reflect response quality and are widely used in reinforcement learning from human feedback (RLHF) pipelines (Ouyang et al., 2022; Lambert et al., 2025; Christiano et al., 2017). Generative RMs, often referred to as LLM-as-a-judge, leverage powerful pretrained LLMs to evaluate responses directly via incorporating explicit evaluation criteria, which can produce preference judgments or detailed rationales without additional fine-tuning (Zhu et al., 2023; Zheng et al., 2023; Kim et al., 2024a; Yuan et al., 2024; Sun et al., 2023). Nevertheless, how these RMs in domain-specific and long-horizon evaluation settings, particularly within mental health dialogue, remains less understood.

Benchmarks of Reward Models Several benchmarks have been proposed to systematically evaluate RMs. Early efforts, RewardBench (Lambert et al., 2025), provided a unified infrastructure for as-

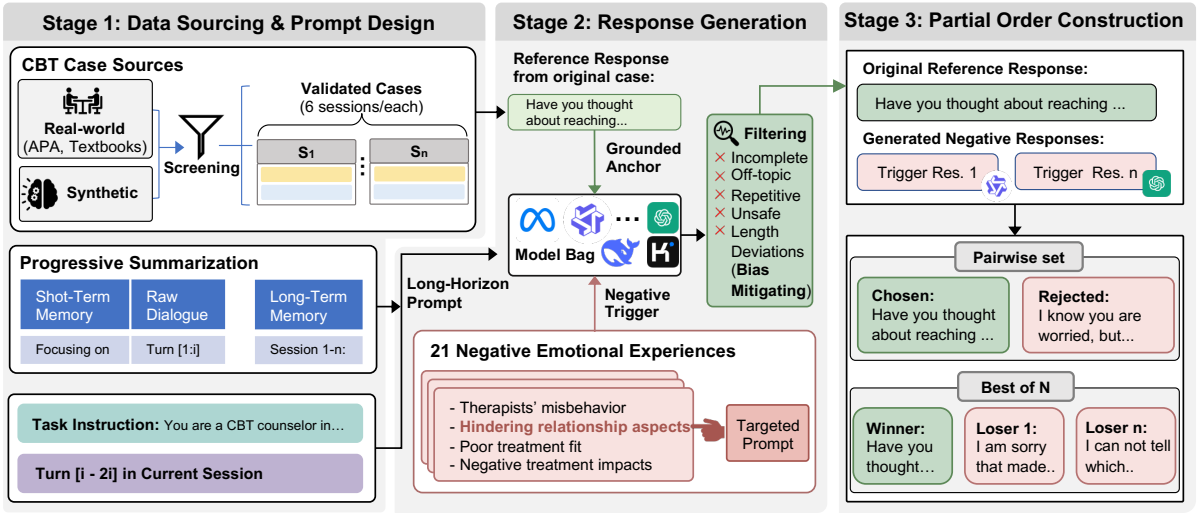

Figure 2: An overview of the PRMB data construction pipeline. The process consists of three stages: (1) Data sourcing and prompt design with progressive summarization for long-horizon contexts. (2) Response generation including clinically grounded positive responses and 21 types of subtle negative responses that elicit client negative emotions. (3) Partial order construction for pairwise preference pairs and BoN sets.

sessing RMs across diverse domains ranging from open-ended chat to reasoning task, thereby helping to formalize RM benchmarking as a distinct research area. Subsequent work, including RMB (Zhou et al., 2024), propose the BoN as a new more challenging evaluation setting. Later efforts have extended evaluation to specialized settings, such as multilinguality (Gureja et al., 2025), mathematical reasoning (Kim et al., 2024b), and agentic behavior (Lù et al., 2025). RoleRMBench (Ding et al., 2025) further pioneer RM benchmarks for role-playing dialogues, where human preferences are nuanced and highly context-dependent. Despite their comprehensiveness, existing benchmarks predominantly focus on short-context interactions, leaving long-horizon preference modeling, especially in process-oriented affective domains such as CBT counseling, largely underexplored.

3 Dataset Construction

To evaluate whether reward models (RMs) can capture subtle negative emotional triggers in process-oriented psychological counseling, we construct **PRMB** through a three-stage pipeline. Unlike standard preference benchmarks, PRMB is designed to reflect the complexity of CBT-based counseling, where responses must remain context-aware and emotionally appropriate across extended interactions. Figure 2 presents an overview of the pipeline, which consists of: (1) data sourcing and progressive prompt design for trajectory-level contexts, (2) response generation including clinically grounded positive responses and carefully constructed subtle negative responses, and (3) partial order construction with pairwise preferences, BoN sets, and human validation.

3.1 Data Sourcing and Prompt Design

CBT Case Data Sources We collect CBT counseling cases from publicly available resources, including real-world examples from the American Psychological Association (APA) website[1], CBT textbooks, and simulated multi-session cases from DiaCBT (Zhou et al., 2025), a large-scale synthetic dataset created under CBT-guided constraints. All cases are used exclusively for research and evaluation. Each case is screened to ensure coverage to core CBT stages, including problem formulation, identification of automatic thoughts, cognitive restructuring, and behavioral interventions. After filtering and verification, we retain 118 validated cases, each comprising six sessions.

Long-Horizon Prompt Design To capture the trajectory-dependent nature of CBT interactions, we design a hierarchical prompt structure that preserves therapeutic continuity without exceeding computational context limits. Each prompt is synthesized into a unified format comprising four components: (1) a task instruction defining the counselor role and goals, (2) a long-term summary capturing cross-session client trajectories, (3) a short-term summary capturing the immediate clinical focus of the current session, and (4) the most re-

[1] https://www.apa.org/pubs/databases/psychtherapy/

cent dialogue turns. Central to our design is a progressive summarization strategy (Lv et al., 2024), which recursively compresses historical session data into the subsequent context. This approach ensures that essential therapeutic anchors, such as previously identified cognitive distortions and established coping strategies, remain accessible to the model. The prompts used for trajectory preservation are provided in the Appendix A.1. This design enables the evaluation of RMs beyond isolated responses, testing their ability to reason over compressed historical context and maintain consistency across sessions. By balancing global context with local dialogue length, we construct diverse prompts that provide a robust foundation for evaluating trajectory-level counseling dynamics.

3.2 Response Generation and Filtering

Response Generation For each prompt, we generate counselor responses using ten state-of-the-art LLMs selected for their strong performance and widespread use as baselines on mainstream leaderboards[2]. The complete list and generation settings are provided in the Appendix A.2.

To construct a challenging preference benchmark, we generate two types of responses for each prompt. **Positive responses** are directly adopted from the original counselor utterances in the sourced CBT cases. These serve as clinically grounded chosen responses, reflecting established CBT practice and providing evidence-based anchors for preference signals (Chiu et al., 2024). **Negative responses** are generated via targeted prompt engineering on the same LLMs. We draw upon the qualitative meta-analysis by (Vybíral et al., 2024), which identifies 21 client-reported negative experience meta-categories across four clusters (*therapists' misbehavior*, *hindering relationship aspects*, *poor treatment fit*, and *negative treatment impacts*). Negative variants start from a positive baseline prompt but introduce controlled, subtle deviations to embody one or more categories. Importantly, these negative variants are designed to be contextually plausible and difficult to distinguish, requiring RMs to detect nuanced affective deviations rather than obvious errors. This makes the resulting preference pairs more challenging than those in conventional benchmarks.

Response Filtering We implement a multi-stage refinement pipeline to ensure the retention of high-quality and comparable response candidates. Specifically, we exclude responses that are semantically incomplete, irrelevant, excessively redundant, or violate safety protocols. Furthermore, to mitigate known biases, we apply rule-based constraints to filter out candidates with significant length deviations relative to authentic CBT reference cases. This step ensures that preference signals are driven by affective and therapeutic quality rather than superficial artifacts such as length or redundancy, which are known to bias RMs.

3.3 Partial Order Construction

Pairwise Preference Construction We form preference pairs by treating the original counselor response from the sourced case as the *chosen* response and one model-generated negative response as the *rejected* response. To increase difficulty and discriminative value, we retain only pairs with moderate quality gaps, discarding trivially poor responses. Rejected responses are sampled across multiple LLMs to balance model distributions.

Agreement with Human Preference To validate the quality of our constructed preference pairs in the complex CBT setting, we randomly sampled 200 pairs and invited three independent human annotators (licensed CBT practitioners) to independently select the preferred counseling response (ties were not allowed). The annotators showed strong alignment with our constructed pairs, selecting the chosen response in 86%, 92%, and 94%, respectively. Inter-annotator agreement is substantial (Fleiss' $\kappa = 0.82$, average pairwise agreement 89%), and the majority vote agreed with our expert labels in 96% of cases. These results confirm that PRMB provides reliable and non-trivial preference signals in a challenging, domain-specific setting.

Best-of-N Set Construction We additionally create a BoN evaluation set using a Best-of-4 configuration. For each query, the reference response is designated as the *winner*, with four model-generated responses (distinct from those used in the pairwise set) as *losers*. This setting better reflects usage scenarios, where RMs must select among multiple plausible candidates rather than evaluate isolated response pairs.

3.4 Data Statistics of our Benchmark

PRMB comprises 13,893 samples derived from 118 CBT cases, forming a large set of pairwise preference pairs and BoN queries. We analyze

[2] https://opencompass.org.cn/home

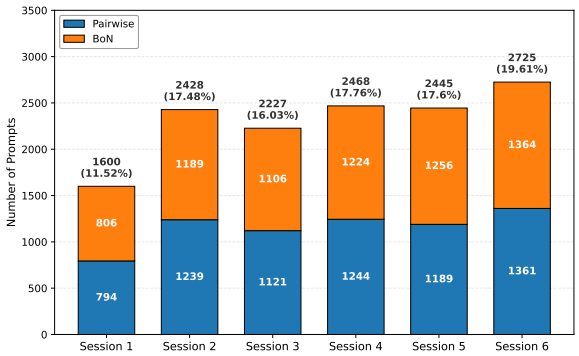

Figure 3: Distribution of samples, pairwise preference pairs, and BoN queries across therapeutic sessions.

Strategy	Samples (%)	Pairwise	BoN
Information Gathering	811 (5.84)	404	407
Setting the Agenda	580 (4.17)	278	302
Weekly Review	411 (2.96)	231	180
Defining Therapeutic Objectives	165 (1.19)	85	80
Psychoeducation	1156 (8.32)	571	585
Working with Automatic Thoughts	3973 (28.60)	1990	1983
Motivational Enhancement	1905 (13.71)	968	937
Working with Intermediate and Core Beliefs	799 (5.75)	400	399
Behavioral Techniques	1087 (7.82)	547	540
Relapse Prevention	553 (3.98)	257	296
Homework Assignments	823 (5.92)	419	404
Requesting Feedback	457 (3.29)	227	230
Summarization	481 (3.46)	239	242
Other	692 (4.98)	332	360
Total	13893 (100.00)	6948	6945

Table 1: Distribution of samples, pairwise pairs, and BoN queries across therapeutic strategies.

its distribution across therapeutic sessions, CBT strategies, and negative experience categories to demonstrate its coverage, diversity, and suitability for evaluating RMs in complex counseling scenarios. More information about CBT is provided in Appendix B.1.

Distribution across Sessions Figure 3 shows the distribution of samples across six sessions, with a gradual increase from Session 1 to Session 6 (highest proportion 19.61%). This pattern reflects the natural accumulation of contextual information and interaction complexity in CBT, where later sessions involve richer histories and more nuanced therapeutic reasoning. Importantly, both pairwise preference pairs (6,948) and BoN queries (6,945) are evenly distributed across sessions, ensuring that evaluation is not biased toward specific stages. This design enables consistent assessment of RMs across different phases of counseling, from early problem formulation to later-stage cognitive restructuring and intervention.

Distribution across CBT Strategies Table 1 summarizes the coverage of 14 core CBT strategies. High-frequency strategies such as "*Working with Automatic Thoughts*" (28.60%) and "*Motivational Enhancement*" (13.71%) align with the central cognitive mechanisms of CBT, while lower-frequency strategies (e.g., "*Defining Therapeutic Objectives* and "*Relapse Prevention*") correspond to specific stages in the therapeutic process. This distribution reflects realistic therapeutic practice and ensures that PRMB captures both common and stage-specific counseling behaviors. As a result, RMs are evaluated across diverse therapeutic functions rather than overfitting to a narrow subset of dialogue patterns.

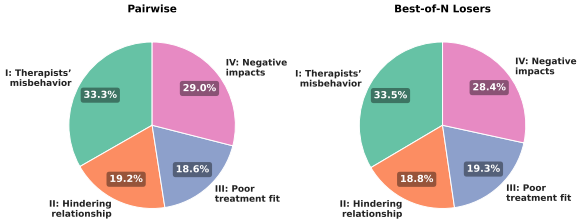

Figure 4: Percentage distribution of negative experience categories in the pairwise and BoN loser sets, grouped by the four-cluster taxonomy.

Distribution of Negative Experiences Figure 4 presents the distribution of negative experience categories based on the four-cluster taxonomy (Vybíral et al., 2024). In both pairwise and BoN sets, *Therapists' misbehavior* (Cluster I) and *Negative treatment impacts* (Cluster IV) are the most prevalent, while *hindering relationship aspects* and *poor treatment fit* also maintain substantial coverage. More details of prominent sub-categories are provided in Appendix B.2. This balanced distribution ensures that PRMB captures a broad spectrum of subtle therapeutic risks. Importantly, many of these negative experiences are context-dependent and not trivially identifiable, making them particularly challenging for RMs to detect. This further reinforces PRMB's role as a benchmark for evaluating nuanced affective alignment.

3.5 Comparison with Existing Benchmarks

PRMB differs from existing RM benchmarks in both task structure and evaluation focus, as summarized in Table 2. RewardBench (Lambert et al., 2025), RMB (Zhou et al., 2024), and RewardBench2 (Malik et al., 2025) primarily evaluate short-context preferences in general-domain settings, focusing on isolated response quality or factual correctness. While effective for broad evaluation, these benchmarks do not capture the trajectory-dependent and affective aspects of alignment required in counseling scenarios. Rol-

Benchmark	Best-of-N	Multi-session	Unseen Prompts	Target Skill
RewardBench (Lambert et al., 2025)	✗	✗	✗	General
RMB (Zhou et al., 2024)	✓	✗	✗	General
RewardBench2 (Malik et al., 2025)	✓	✗	✓	General
RoleRMBench (Ding et al., 2025)	✗	✗	✗	Role-playing
PRMB (ours)	✓	✓	✓	CBT Counseling

Table 2: Comparison of representative reward model benchmarks.

eRMBench (Ding et al., 2025) introduces role-playing consistency, but remains limited to single-interaction settings without modeling session-level progression or evolving user states.

In contrast, PRMB is explicitly designed for multi-session CBT counseling, where preferences depend on coherence across sessions and sensitivity to subtle emotional cues. In addition, PRMB incorporates a progressive summarization framework to preserve essential historical context without relying on full dialogue transcripts, enabling scalable evaluation of long-horizon interactions. Together, these features position PRMB as a complementary benchmark that targets aspects of alignment not covered by existing evaluation protocols.

4 Benchmarking Reward Models

We evaluate a broad range of state-of-the-art reward models (RMs) on PRMB. This section presents the evaluation setup and key findings.

4.1 Evaluation Setup

We evaluate both discriminative and generative RMs. Discriminative RMs assign scalar rewards to responses, while generative RMs directly select human-preferred responses following LLM-as-a-judge paradigms (Zhou et al., 2024). For each pairwise instance i, the benchmark provides a chosen and rejected response (x_i^+, x_i^-). Accuracy measures whether the RM assigns higher preference to the chosen response. The pairwise accuracy $\mathcal{A}_{\text{pw}}$ is computed as:

$$\mathcal{A}_{\text{pw}} = \frac{1}{N}\sum_{i=1}^{N} g(x_i^+, x_i^-), \quad (1)$$

where $g(\cdot)$ is 1 if the RM prefers chosen over the rejected response, otherwise it is 0.

For each BoN instance i, the benchmark provides a winner $x_i^\star$ and a set of losers $\{x_{ij}^-\}_{j=1}^{P_i}$. An RM succeeds on instance i only if it prefers the winner over all losers. The BoN accuracy $\mathcal{A}_{\text{BoN}}$ is:

$$\mathcal{A}_{\text{BoN}} = \frac{1}{N}\sum_{i=1}^{N}\prod_{j=1}^{P_i} g(x_i^\star, x_{ij}^-). \quad (2)$$

We adopt accuracy-based metrics to directly measure the correctness of relative preference judgments, which is essential for RMs to provide reliable supervision signals.

4.2 Evaluation Results

Overall Performance Table 3 summarizes the performance of all evaluated RMs. While models achieve relatively strong performance in pairwise settings (up to 86.46%), their BoN accuracy remains substantially lower (mostly below 70%), resulting in overall scores between 52.57% and 76.73%. Compared to typical results reported on general-domain benchmarks, these results indicate that RM performance is notably reduced in counseling scenarios involving subtle affective distinctions. This result reveals that existing evaluation protocols may provide an optimistic estimate of alignment quality when applied to process-oriented tasks.

Pairwise vs. Best-of-N Evaluation A consistent and substantial gap emerges between pairwise and BoN performance across nearly all models. While pairwise accuracy typically falls between 73% and 84%, BoN accuracy drops to 45%–67%, with an average gap exceeding 20 percentage points. This gap indicates that pairwise evaluation overestimates RM capability in realistic settings. Although RMs can distinguish a correct response from a single negative alternative, they often fail to identify it when competing against multiple plausible distractors. This limitation is particularly pronounced in LLM-as-a-judge RMs, which exhibit strong pairwise performance but significantly lower BoN robustness. These findings highlight a key limitation that local preference comparisons do not adequately reflect RMs' ability to perform global ranking under realistic inference conditions.

4.3 Performance across Negative Experiences

We further analyze performance across different categories of negative experiences. As shown in Figure 5, pairwise accuracy remains relatively stable across clusters, whereas BoN accuracy exhibits

Reward Model	Pairwise Acc.	BoN Acc.	Overall Acc.
Discriminative Reward Models			
Llama-3.1-8B-Instruct-RM-RB2	0.8646	0.6700	0.7673
Skywork-Reward-V2-Qwen3-8B	0.8365	0.6092	0.7229
Skywork-Reward-V2-Qwen3-4B	0.8247	0.5885	0.7066
internlm2-20b-reward	0.8122	0.5758	0.6940
internlm2-7b-reward	0.7974	0.5675	0.6825
Skywork-Reward-V2-Qwen3-1.7B	0.8051	0.5454	0.6753
Skywork-Reward-V2-Qwen3-0.6B	0.7596	0.4825	0.6211
Skywork-Reward-V2-Llama-3.1-8B	0.7306	0.4556	0.5931
internlm2-1.8b-reward	0.7047	0.3950	0.5499
LLM-as-a-judge Reward Models			
deepseek-v3.2-exp⋆	0.7796	0.7344	0.7570
gpt-4o-mini⋆	0.8368	0.6266	0.7317
Qwen2.5-7B-Instruct⋆	0.7860	0.5584	0.6722
Gemma3-4B-IT⋆	0.7533	0.5006	0.6270
gpt-oss-20b⋆	0.7500	0.4862	0.6181
Mistral-7B-Instruct-v0.3⋆	0.7306	0.4937	0.6122
LLaMA-3.2-3B-Instruct⋆	0.6513	0.4001	0.5257

Table 3: The leaderboard of PRMB, ranked by the average of pairwise and BoN accuracy. ⋆ denotes LLM-as-a-judge reward models.

significant variation. Models perform well on overt and easily identifiable errors (e.g., *Therapist Judging*, *Not Listening*), but struggle with subtle and cumulative harms (e.g., *Loss of Hope*, *Increased Problems after Therapy*). In these cases, BoN accuracy often drops below 65%, indicating difficulty in capturing nuanced affective signals. Detailed per-category results are provided in Appendix C.3. This pattern suggests that current RMs are more sensitive to explicit misbehavior than to implicit or long-term emotional impacts. Such limitations are particularly critical in psychotherapy, where harmful effects often emerge gradually rather than through single responses. These results demonstrate that PRMB exposes systematic weaknesses in RM evaluation, particularly in handling subtle, context-dependent alignment.

5 Further Analysis

A key requirement of any reward model (RM) benchmark is its ability to predict downstream alignment performance. In this section, we evaluate whether PRMB rankings are predictive of real-world response quality, and further analyze how common inference-time strategies affect RM performance in this setting. Together, these analyses assess both the validity of PRMB and the robustness of current reward modeling practices. More analysis is provided in Appendix C.

5.1 Downstream Evaluations

Experimental Setup We assess downstream performance on a held-out set of 2,000 counseling prompts. For each prompt, we sample 16 candidate responses from a policy LLM at temperature 0.8. Each RM scores and ranks the candidates, and the highest-scoring response is selected as the final output. We evaluate generation quality using both automatic and human metrics. Automatic evaluation is conducted via BERTScore (Zhang* et al., 2020) against reference CBT counselor responses. For human evaluation, we sample 150 prompts and ask three licensed CBT practitioners to rate responses on two 5-point Likert scales: **therapeutic appropriateness** and **therapeutic alliance building**. Annotators are blinded to model identities. We compute Spearman's rank correlation coefficient between the PRMB rankings and the downstream rankings across four policy models[3].

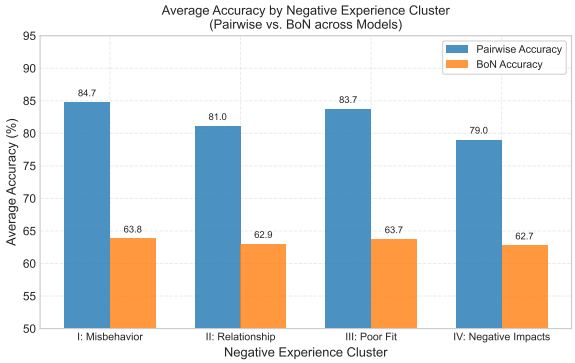

Figure 5: Average pairwise and BoN accuracy across the four main clusters of negative experiences.

Results As shown in Figure 6, PRMB rankings exhibit strong and significant correlation with

[3]Qwen2.5-7B-Instruct, LLaMA-3.2-3B-Instruct, Mistral-7B-Instruct-v0.3, and Gemma3-4B-IT.

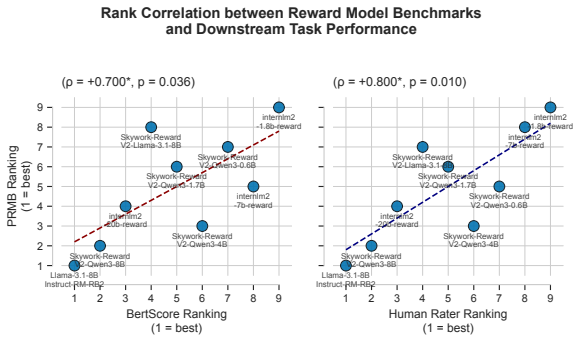

Figure 6: Rank correlation between benchmark rankings and downstream BoN performance. PRMB achieves a strong positive monotonic relationships with BertScore ($\rho = 0.700$, $p \approx 0.036$) and human rater ($\rho = 0.800$, $p \approx 0.010$).

Method	Pairwise Acc.	BoN Acc.
gpt-4o-mini	0.8368	0.6266
+ Few shot (2-shot)	0.8271	0.6000
+ RAG	**0.8582**	**0.6584**
+ Self Refine	0.8472	0.6288
+ CoT	0.8276	0.5783
Qwen2.5-7B-Instruct	0.7860	0.5584
+ Few shot (2-shot)	0.7706	0.5379
+ RAG	**0.8044**	**0.6029**
+ Self Refine	0.7700	0.5479
+ CoT	0.7884	0.5054
LLaMA-3.2-3B-Instruct	0.6513	0.4001
+ Few shot (2-shot)	0.6220	0.3865
+ RAG	**0.6950**	**0.4813**
+ Self Refine	0.4800	0.2071
+ CoT	0.5025	0.2535

Table 4: Effects of inference-time strategies on pairwise and BoN accuracy across different generative reward model backbones.

downstream performance. The correlation with BERTScore reaches $\rho = 0.700$ ($p = 0.036$), while correlation with human evaluation is even higher at $\rho = 0.800$ ($p <= 0.01$). These results indicate that RMs ranked highly on PRMB consistently select responses with better lexical alignment and superior clinical quality, including stronger adherence to CBT principles and improved therapeutic alliance. The stronger correlation with human ratings suggests that PRMB captures aspects of alignment that are not fully reflected by surface-level automatic metrics. This provides evidence that PRMB serves as a meaningful proxy for real-world alignment quality in counseling scenarios.

5.2 Effects of Inference-Time Strategies

Experimental Setup We evaluate four common inference-time strategies on three generative RMs. **Few-shot (2-shot)** provides two high-quality preference pairs as in-context examples. **Chain-of-Thought (CoT)** instructs step-by-step reasoning before the final judgment. **Self-Refine** generates an initial judgment followed by two rounds of self-critique. **Retrieval-Augmented Generation (RAG)** prepends the top-5 most relevant high-quality responses retrieved from an external knowledge base. All settings use temperature 0.0 and identical decoding parameters.

Results Table 4 reveals that most heuristic prompting strategies fail to improve and often degrade performance in PRMB. Few-shot prompting consistently reduces accuracy across all models, suggesting that in-context examples may introduce noise in trajectory-dependent preference judgments. Both CoT and self-refine lead to performance drops, particularly in the BoN setting. For smaller models, these strategies can cause severe degradation (e.g., BoN accuracy drops below 25%), indicating that additional reasoning steps may amplify biases rather than improve decision quality. In contrast, RAG consistently improves performance across all models, with the largest gains observed in smaller backbones. This suggests that retrieving domain-specific therapeutic knowledge is more effective than relying on implicit reasoning strategies. These findings highlight that common inference-time strategies are not universally beneficial for reward modeling. In process-oriented counseling scenarios, external knowledge grounding appears to be more reliable than self-generated reasoning for improving alignment judgments.

6 Conclusions

In this paper, we identify a key limitation: while reward models perform strongly under standard pairwise settings, their effectiveness is reduced in process-oriented counseling scenarios that require distinguishing among multiple plausible responses and handling subtle negative triggers. We introduce PRMB, a benchmark designed to capture trajectory-dependent and negative experience of alignment in multi-session CBT dialogues. Our results show that PRMB reveals consistent performance gaps across models and correlates strongly with downstream response quality, suggesting it provides a more realistic estimate of alignment in human-centered interactions. More broadly, our findings highlight the need for evaluation frameworks that move beyond local preference comparisons toward process-oriented and context-sensitive assessments.

Limitations

As far as we know, it is the first attempt to construct a benchmark for evaluating reward models in long-horizon, multi-session CBT-based counseling. Although the benchmark is constructed from both real and simulated CBT counseling cases, it does not fully capture the duration, depth, and complexity of real-world psychotherapy. CBT counseling in real scenarios typically spans more sessions and involves richer interpersonal dynamics. Thus, bridging the gap with real-world data while ensuring ethical safety remains a significant challenge.

Ethical Considerations

Data Privacy and Usage

The 118 CBT cases in PRMB are constructed from publicly available resources (e.g., CBT textbooks and educational examples) and expert-curated synthetic counseling dialogues designed for research purposes. No real patient data or personally identifiable information is included in the dataset.

Negative emotional triggers are intentionally introduced as controlled perturbations to evaluate long-horizon safety and preference modeling in counseling dialogue systems. Human preference annotations were conducted by three licensed CBT practitioners under informed consent. Annotators were instructed to perform preference selection independently.

PRMB is designed strictly as a research benchmark for evaluating reward models in counseling dialogue settings. It is not intended for clinical use or for training or deploying end-to-end therapeutic systems. Due to the presence of intentionally constructed unsafe or negative response patterns, direct use of the dataset for model training may lead to undesirable behaviors and is therefore discouraged.

References

Amanda Askell, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, T. J. Henighan, Andy Jones, Nicholas Joseph, Benjamin Mann, Nova Dassarma, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, John Kernion, Kamal Ndousse, Catherine Olsson, Dario Amodei, Tom B. Brown, Jack Clark, Sam McCandlish, Chris Olah, and Jared Kaplan. 2021. A general language assistant as a laboratory for alignment. *ArXiv*, abs/2112.00861.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova Dassarma, Dawn Drain, Stanislav Fort, Deep Ganguli, T. J. Henighan, Nicholas Joseph, Saurav Kadavath, John Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom B. Brown, Jack Clark, Sam McCandlish, Chris Olah, Benjamin Mann, and Jared Kaplan. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *ArXiv*, abs/2204.05862.

Judith S Beck. 2020. *Cognitive behavior therapy: Basics and beyond*. Guilford Publications.

Siyuan Chen, Mengyue Wu, Kenny Q Zhu, Kunyao Lan, Zhiling Zhang, and Lyuchun Cui. 2023a. Llm-empowered chatbots for psychiatrist and patient simulation: application and evaluation. *arXiv preprint arXiv:2305.13614*.

Yirong Chen, Xiaofen Xing, Jingkai Lin, Huimin Zheng, Zhenyu Wang, Qi Liu, and Xiangmin Xu. 2023b. SoulChat: Improving LLMs' empathy, listening, and comfort abilities through fine-tuning with multi-turn empathy conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1170–1183. Association for Computational Linguistics.

Yirong Chen, Xiaofen Xing, Jingkai Lin, Huimin Zheng, Zhenyu Wang, Qi Liu, and Xiangmin Xu. 2023c. Soulchat: Improving llms' empathy, listening, and comfort abilities through fine-tuning with multi-turn empathy conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1170–1183.

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

Paul F. Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 4302–4310. Curran Associates Inc.

Hang Ding, Qiming Feng, Dongqi Liu, Qi Zhao, Tao Yao, Shuo Wang, Dongsheng Chen, Jian Li, Zhenye Gan, Jiangning Zhang, Chengjie Wang, and Yabiao Wang. 2025. Rolermbench&rolerm: Towards reward modeling for profile-based role play in dialogue systems.

Srishti Gureja, Lester James Validad Miranda, Shayekh Bin Islam, Rishabh Maheshwary, Drishti Sharma, Gusti Triandi Winata, Nathan Lambert, Sebastian Ruder, Sara Hooker, and Marzieh Fadaee. 2025. M-RewardBench: Evaluating reward models in multilingual settings. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 43–58. Association for Computational Linguistics.

Steven C Hayes and Stefan G Hofmann. 2021. "third-wave" cognitive and behavioral therapies and the emergence of a process-based approach to intervention in psychiatry. *World psychiatry*, 20(3):363–375.

Yuanyuan He, Yongsen Pan, Wei Li, Jiali You, Jiawen Deng, and Fuji Ren. 2025. ECC: An emotion-cause conversation dataset for empathy response. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 6011–6028. Association for Computational Linguistics.

Shang-Ling Hsu, Raj Sanjay Shah, Prathik Senthil, Zahra Ashktorab, Casey Dugan, Werner Geyer, and Diyi Yang. 2023. Helping the helper: Supporting peer counselors via ai-empowered practice and feedback. *arXiv preprint arXiv:2305.08982*.

Eunkyung Jo, Daniel A Epstein, Hyunhoon Jung, and Young-Ho Kim. 2023. Understanding the benefits and challenges of deploying conversational ai leveraging large language models for public health intervention. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–16.

Seungone Kim, Juyoung Suk, Shayne Longpre, Bill Yuchen Lin, Jamin Shin, Sean Welleck, Graham Neubig, Moontae Lee, Kyungjae Lee, and Minjoon Seo. 2024a. Prometheus 2: An open source language model specialized in evaluating other language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4334–4353. Association for Computational Linguistics.

Sunghwan Kim, Dongjin Kang, Taeyoon Kwon, Hyungjoo Chae, Jungsoo Won, Dongha Lee, and Jinyoung Yeo. 2024b. Evaluating robustness of reward models for mathematical reasoning. *ArXiv*, abs/2410.01729.

Nathan Lambert, Valentina Pyatkin, Jacob Morrison, LJ Miranda, Bill Yuchen Lin, Khyathi Chandu, Nouha Dziri, Sachin Kumar, Tom Zick, Yejin Choi, Noah A. Smith, and Hannaneh Hajishirzi. 2025. RewardBench: Evaluating reward models for language modeling. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 1755–1797. Association for Computational Linguistics.

Suhyun Lee, Palakorn Achananuparp, Neemesh Yadav, Ee-Peng Lim, and Yang Deng. 2026. Mhsafeeval: Role-aware interaction-level evaluation of mental health safety in large language models. *arXiv preprint arXiv:2604.17730*.

Suyeon Lee, Sunghwan Kim, Minju Kim, Dongjin Kang, Dongil Yang, Harim Kim, Minseok Kang, Dayi Jung, Min Hee Kim, Seungbeen Lee, Kyong-Mee Chung, Youngjae Yu, Dongha Lee, and Jinyoung Yeo. 2024. Cactus: Towards psychological counseling conversations using cognitive behavioral theory. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14245–14274. Association for Computational Linguistics.

June M Liu, Donghao Li, He Cao, Tianhe Ren, Zeyi Liao, and Jiamin Wu. 2023. Chatcounselor: A large language models for mental health support. *arXiv preprint arXiv:2309.15461*.

Xing Han Lù, Amirhossein Kazemnejad, Nicholas Meade, Arkil Patel, Dongchan Shin, Alejandra Zambrano, Karolina Stańczak, Peter Shaw, Christopher Pal, and Siva Reddy. 2025. Agentrewardbench: Evaluating automatic evaluations of web agent trajectories. *ArXiv*, abs/2504.08942.

Yaojia Lv, Haojie Pan, Zekun Wang, Jiafeng Liang, Yuanxing Liu, Ruiji Fu, Ming Liu, Zhongyuan Wang, and Bing Qin. 2024. CogGPT: Unleashing the power of cognitive dynamics on large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 6074–6091. Association for Computational Linguistics.

Saumya Malik, Valentina Pyatkin, Sander Land, Jacob Daniel Morrison, Noah A. Smith, Hanna Hajishirzi, and Nathan Lambert. 2025. Rewardbench 2: Advancing reward model evaluation. *ArXiv*, abs/2506.01937.

Hongbin Na. 2024. Cbt-llm: A chinese large language model for cognitive behavioral therapy-based mental health question answering. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2930–2940.

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 Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22. Curran Associates Inc.

Jacqueline B Persons. 2012. *The case formulation approach to cognitive-behavior therapy*. Guilford Press.

Huachuan Qiu, Hongliang He, Shuai Zhang, Anqi Li, and Zhenzhong Lan. 2024. SMILE: Single-turn to multi-turn inclusive language expansion via ChatGPT for mental health support. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 615–636. Association for Computational Linguistics.

Ashish Sharma, Kevin Rushton, Inna Wanyin Lin, Theresa Nguyen, and Tim Althoff. 2024. Facilitating self-guided mental health interventions through human-language model interaction: A case study of cognitive restructuring. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24. Association for Computing Machinery.

Zhiqing Sun, Yikang Shen, Hongxin Zhang, Qinhong Zhou, Zhenfang Chen, David D. Cox, Yiming Yang, and Chuang Gan. 2023. Salmon: Self-alignment with instructable reward models. In *International Conference on Learning Representations*.

Ningjing Tang, Alice Qian, Qiaosi Wang, Esther Howe, Blake Bullwinkel, Paola Pedrelli, Jina Suh, Hoda Heidari, and Hong Shen. 2026. Beyond the single turn: Reframing refusals as dynamic experiences embedded in the context of mental health support interactions with llms. *arXiv preprint arXiv:2602.01694*.

Zbyněk Vybíral, Benjamin M Ogles, Tomáš Řiháček, Barbora Urbancová, and Veronika Gocieková. 2024. Negative experiences in psychotherapy from clients' perspective: A qualitative meta-analysis. *Psychotherapy Research*.

Peisong Wang, Ruotian Ma, Bang Zhang, Xingyu Chen, Zhiwei He, Kang Luo, Qingsong Lv, Qingxuan Jiang, Zheng Xie, Shanyi Wang, Yuan Li, Fanghua Ye, Jian Li, Yifan Yang, Zhaopeng Tu, and Xiaolong Li. 2025. Rlver: Reinforcement learning with verifiable emotion rewards for empathetic agents. *ArXiv*, abs/2507.03112.

Jing Wei, Sungdong Kim, Hyunhoon Jung, and Young-Ho Kim. 2024. Leveraging large language models to power chatbots for collecting user self-reported data. *Proceedings of the ACM on Human-Computer Interaction*, 8(CSCW1):1–35.

Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Xian Li, Sainbayar Sukhbaatar, Jing Xu, and Jason Weston. 2024. Self-rewarding language models. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Bang Zhang, Ruotian Ma, Qingxuan Jiang, Peisong Wang, Jiaqi Chen, Zheng Xie, Xingyu Chen, Yue Wang, Fanghua Ye, Jian Li, Yifan Yang, Zhaopeng Tu, and Xiaolong Li. 2025a. Sentient agent as a judge: Evaluating higher-order social cognition in large language models. *ArXiv*, abs/2505.02847.

Lunjun Zhang, Arian Hosseini, Hritik Bansal, Mehran Kazemi, Aviral Kumar, and Rishabh Agarwal. 2024. Generative verifiers: Reward modeling as next-token prediction. *ArXiv*, abs/2408.15240.

Naifan Zhang, Ruihan Sun, Ruixi Su, Shiqi Ma, Shiya Zhang, Xianna Weng, Xiaofan Zhang, Yuhan Zhan, Yuyang Xu, Zhaohan Chen, et al. 2025b. Echo-n1: Affective rl frontier. *arXiv preprint arXiv:2512.00344*.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Advances in Neural Information Processing Systems*, volume 36, pages 46595–46623. Curran Associates, Inc.

Enyu Zhou, Guodong Zheng, Bing Wang, Zhiheng Xi, Shihan Dou, Rong Bao, Wei Shen, Limao Xiong, Jessica Fan, Yurong Mou, Rui Zheng, Tao Gui, Qi Zhang, and Xuanjing Huang. 2024. Rmb: Comprehensively benchmarking reward models in llm alignment. *ArXiv*, abs/2410.09893.

Yougen Zhou, Ningning Zhou, Qin Chen, Jie Zhou, Aimin Zhou, and Liang He. 2025. Diacbt: A long-periodic dialogue corpus guided by cognitive conceptualization diagram for cbt-based psychological counseling. *ArXiv*, abs/2509.02999.

Lianghui Zhu, Xinggang Wang, and Xinlong Wang. 2023. Judgelm: Fine-tuned large language models are scalable judges. *ArXiv*, abs/2310.17631.

A Additional Materials on Data Construction

A.1 Progressive Summary Generation

To support long-horizon CBT counseling evaluation while keeping prompt length manageable, we adopt a progressive summary generation framework to construct counseling prompts. Historical information is incrementally compressed into structured summaries that are carried forward across sessions. Specifically, for each counseling case consisting of multiple sessions, we maintain two types of summaries:

- **Short-term summary**, which captures the key developments within the current session up to the present turn, such as recent emotional states and newly expressed thoughts, as show in Figure 7.
- **Long-term summary**, which captures stable and cross-session information about the client, including presenting problems, core beliefs, recurring automatic thoughts, previously identified cognitive distortions, and therapeutic strategies that have been introduced, as shown in Figure 8.

At the beginning of the first session, both summaries are initialized as empty. After each session concludes, the short-term summary is merged into the long-term summary using a structured summarization prompt, and the short-term summary is reset for the next session. This process ensures that essential therapeutic trajectories are preserved across sessions without exceeding context length.

Prompt for short-term memory (stage one)

You are a professional cognitive behavioral therapy (CBT) therapist. Please write a **150-250 word** summary based on the following dialogue.

Requirements:
- Identify key events and themes
- The client's core emotions, concerns, and struggles
- Possible automatic thoughts or core beliefs
- The focus of the therapist-client interaction and initial goals

Dialogue content:
{dialogue_text}

Please output: A clear, structured summary of 150-250 words.

Prompt for short-term memory (stage two)

You are a professional Cognitive Behavioral Therapist (CBT) counselor.

Below is your previous summary of the dialogue (please do not discard any useful information):

{previous_summary}

Now, you are given new additional dialogue content. Please:
1. Read the new content and extract any new information relevant to the counseling progress (including emotions, automatic thoughts, interpersonal events, topic changes, newly emerging issues, etc.).
2. **Supplement, revise, and enhance** the previous summary.
3. Output a **coherent and consistent updated summary of 150–250 words.**

The new dialogue content is as follows:

{dialogue_text}

Please output: a comprehensive summary (150-250 words) that integrates all previous dialogue content and includes the new information from this round, not just a summary of the new content.

Figure 7: Prompt usage in short-term memory generation.

Importantly, summaries are constructed solely from prior dialogue content and do not incorporate or reference any candidate counselor responses used for preference evaluation. This prevents information leakage and ensures that reward models are evaluated only on information that would be available in realistic long-horizon counseling settings.

A.2 Model List Usage in Response Candidate Generation

The models used for response candidate generation include both open-weight and closed-source systems, and span instruction-tuned and chat-optimized variants. All models are used in a zero-shot generation setting with fixed decoding parameters to ensure comparability across candidates. Across all models, responses are generated with temperature set to 0.8 and a maximum output length sufficient to complete a full counseling turn. No model-specific prompt tuning or post-processing is applied. Each prompt is paired with multiple independently generated responses sampled from different models, enabling the construction of challenging preference pairs and Best-of-N instances that reflect realistic model diversity. A complete list of models used for response candidate generation is shown in Table 5.

Prompt for long-term memory

You are a professional Cognitive Behavioral Therapist (CBT) counselor. Below are several segmented summaries of this session, arranged chronologically. Please synthesize these summaries to generate a structured session-level summary of this consultation, ensuring the format is identical to the example:

Session Summary
Focus:
- Main discussion topics and triggering situations
- Key issues discussed in this session (e.g., cognitive restructuring, emotion regulation, avoidance behavior, behavioral experiments, etc.)
Progress:
- The client's specific progress in emotion recognition, cognitive restructuring, and behavioral adjustment
- New insights, completed exercises, or improved skills
Setbacks / Challenges:
- Unresolved difficulties from this session
- Recurring automatic thoughts, behavioral patterns, and hindering factors
- Bottlenecks in skill application or challenges in real-world situations
Next Steps:
- Recommended next-stage CBT intervention directions (cognitive reappraisal, behavioral experiments, exposure, emotion monitoring, homework assignments, etc.)
- Specific and actionable follow-up action suggestions

Requirements:
1. Language should be objective, neutral, and professional, using clear and specific expressions.
2. List 2–4 items in each section using bullet points (-).
3. All content must be based on the input summaries and cannot be fabricated.
4. Do not output explanations, prompts, or additional text.

Input content is as follows:
{summaries_text}

Please generate a session-level summary according to the above requirements.

Figure 8: Prompt usage in long-term memory generation.

Model Name	Type
GLM-4.6	Closed-source
Qwen3-235B-A22B-Thinking-2507	Open-weight
Qwen3-30B-A3B-Thinking-2507	Open-weight
Qwen3-32B	Open-weight
DeepSeek-V3.2-Exp	Open-weight
DeepSeek-R1-0528	Open-weight
gpt-4o-mini	Closed-source
gpt-oss-20b	Open-weight
MiniMax-M2	Open-weight
Kimi-K2	Open-weight

Table 5: Models used for response candidate generation.

B Details of PRMB

B.1 Cognitive Behavioral Therapy

Cognitive Behavioral Therapy (CBT) initially targeted the treatment of depression. Its core principle is the interconnectivity of thoughts, emotions, behaviors, and physiological responses within a unified system. Changes in one component can significantly influence the others. For instance, when an unexpected and alarming event occurs, such as a television explosion, a person might experience an adrenaline surge (physiological response), an immediate behavioral reaction (grabbing a fire blanket), anxiety (emotional response), and a thought like *The house is on fire, and I might die* (cognitive

response). This scenario illustrates CBT's recognition of the interconnected and holistic nature of these processes.

CBT emphasizes that individuals respond uniquely to situations, shaped by their thoughts. Feelings, rather than being directly caused by situations, largely stem from how situations are perceived and interpreted. Central to CBT is its focus on two aspects of thinking: **automatic thoughts**, which refers to thoughts that arise unconsciously in the stream of consciousness, and **intermediate beliefs**, which are assumptions that generate thoughts. These insights form the foundation for CBT to understanding and modifying thought patterns to improve emotional and behavioral outcomes.

The technical label system is developed by CBT experts drawing from their consulting experience and referring to a related book named *Practical Techniques of CBT*. The descriptions for the 14 selected CBT techniques can be found in Table 7. Session themes: (1) Information gathering, goal setting, and case conceptualization. (2) Identifying and challenging automated thoughts. (3) Discussion case conceptualization. (4) Adjusting intermediate beliefs and core beliefs. (5) Relapse prevention. (6) Consolidation.

B.2 Distribution of Sub-Categories in Negative Experiences

Table 6 presents the distribution of prominent sub-categories identified in our analysis of negative client experiences with therapy. Percentages are reported relative to the total negative experiences in pairwise and Best-of-N, not within-cluster. Sub-categories are grouped by the four main clusters. The most frequent sub-categories in the pairwise set are *Increased Problems after Therapy* (5.38%), *Confusion or Uncertainty* (5.04%), and *Feeling Fear of the Therapy Process* (4.99%). In the BoN set, the top three are *Therapist Not Listening* (4.96%), *Therapist Devaluing the Client* (4.86%), and *Other Inappropriate Verbal Reactions* (4.81%).

Sub-category	Pairwise (%)	BoN (%)
Cluster I: therapists' misbehavior		
Therapist Not Listening	4.63%	4.96%
Therapist Not Understanding	4.76%	4.67%
Therapist Perceived as Incompetent	4.59%	4.74%
Therapist Devaluing the Client	4.72%	4.86%
Therapist Judging	4.88%	4.76%
Therapist Using Client for Own Benefit	4.92%	4.73%
Other Inappropriate Verbal Reactions	4.78%	4.81%
Cluster II: hindering relationship		
Distance and/or Lack of Empathy	4.94%	4.61%
Insecurity or Distrust	4.56%	4.72%
Confusion or Uncertainty	5.04%	4.81%
Poor Interpersonal Match with the Therapist	4.63%	4.62%
Cluster III: poor treatment fit		
Negative Evaluation of Practical Aspects	4.43%	4.71%
Unmet Expectations	4.56%	4.81%
Lack of Fit with the Intervention	4.88%	4.76%
Dissatisfaction with Therapy Ending	4.71%	4.99%
Cluster IV: negative impacts of treatment		
No Change or Insufficient Change	4.88%	4.75%
Increased Problems after Therapy	5.38%	4.68%
Feeling Fear of the Therapy Process	4.99%	4.73%
Loss of Motivation or Hope	4.53%	4.75%
Unpleasant Feelings During Therapy	4.69%	4.75%
Negative Cognitions Aroused in Therapy	4.48%	4.77%

Table 6: Distribution of prominent sub-categories. Percentages are relative to the total negative experiences in each set (not within-cluster).

C Further Benchmarking Reward Models

C.1 Performance across Sessions

We evaluate pairwise accuracy, BoN accuracy, and their average for the top three discriminative and top three LLM-as-a-judge RMs across six CBT sessions as shown in Figure 9. Early sessions (1–2) yield higher BoN accuracies due to simpler content, while mid-to-late sessions (3–6) show clear degradation, reflecting increased complexity in cognitive restructuring and consolidation. Discriminative RMs consistently outperform LLM-as-a-judge RMs across all sessions. LLAMA-3.1-8B-INSTRUCT-RM-RB2 achieves the highest performance, improving from 75.8% (early) to 81.2% (late sessions). The SKYWORK-REWARD-V2 family remains relatively stable with modest gains later. Generative RMs exhibit larger pairwise–BoN gaps (15–20 points) and sharper drops in mid-to-late sessions. For example, GPT-4O-MINI BoN accuracy falls from 62.3% (Session 1) to 58.9%–68.5% (Sessions 4–6). These results highlight the growing challenge of robust preference alignment in later CBT stages, where subtle cumulative harms emerge. Current RMs excel more at coarse discrimination than at recovering optimal responses in long-horizon scenarios.

C.2 Performance across CBT Strategies

We evaluate the top-3 generative and top-3 discriminative RMs across 14 CBT strategies as shown in Figure 10. Discriminative RMs show a modest but consistent advantage in CBT preference modeling. LLAMA-3.1-8B-INSTRUCT-RM-RB2 achieves the highest mean accuracy, topping 8 of 14 strategies, particularly *homework assignments*, *requesting feedback*, and *behavioral*

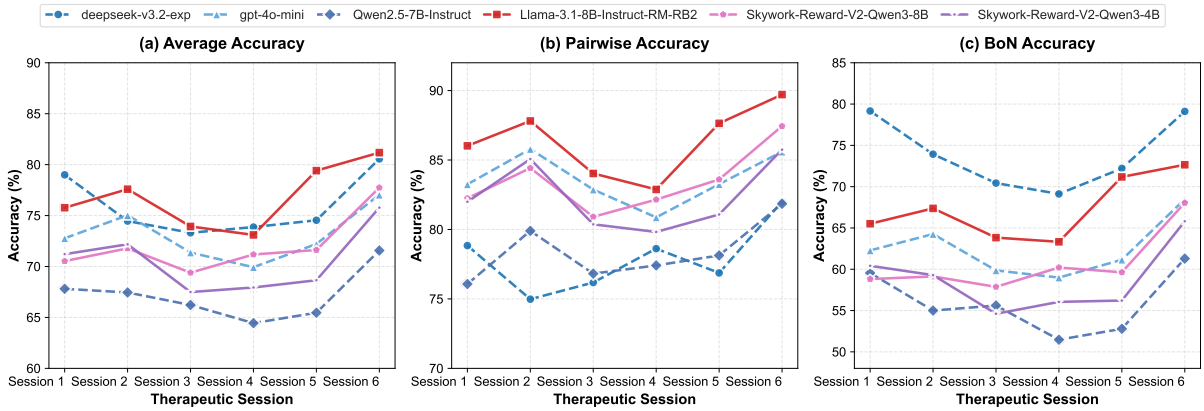

Figure 9: (a) Average accuracy, (b) Pairwise accuracy, and (c) Best-of-N accuracy across the six therapeutic sessions. Solid lines represent discriminative reward models; dashed lines represent LLM-as-a-judge models.

Figure 10: Accuracy of top generative and discriminative reward models across CBT strategies.

techniques. DEEPSEEK-V3.2-EXP ranks second. Action-oriented strategies (*summarization*, *behavioral techniques*, *motivational enhancement*, *homework assignments*) yield high accuracy, while agenda-setting ones (*defining therapeutic objectives*, *setting the agenda*, *weekly review*) perform lower.

C.3 Detailed Performance on Negative Experience

This appendix provides detailed pairwise and BoN accuracy for each of the 21 negative experience meta-categories, across the six evaluated reward models. Categories are grouped by their original four-cluster taxonomy for reference. Figure 11 provides a heatmap visualization of BoN accuracy across models and categories, with darker shades indicating higher performance.

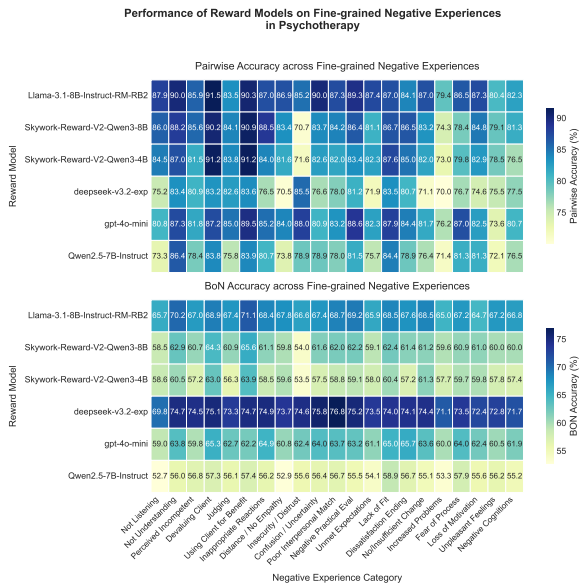

Figure 11: Detailed per-category pairwise and BoN accuracies for the six evaluated models.

CBT Strategy	Description
Information Gathering	The therapist understands key information such as the client's basic situation, main complaints, growth experience, current living conditions, etc.
Setting the Agenda	The therapist sets the topic to be discussed today and gives the reasons while confirming that the patient agrees with the topic raised by the therapist.
Weekly Review	The therapist begins by reviewing the patient's assessment from the previous evaluation and keeps the initial conceptualization and treatment plan in mind during the session.
Defining Therapeutic Objectives	The therapist focuses on the identification of specific problems. As a logical extension, the therapist can help the patient transform these problems into goals and work on them in therapy.
Psychoeducation	When clients express distress about their problems or disorders, this distress is explored and explained in a timely manner during therapy.
Working with Automatic Thoughts	When discussing a specific problem with a patient, the therapist can elicit the automatic thoughts associated with the problem and teach the patient to identify the automatic thoughts.
Motivational Enhancement	The therapist increases the client's motivation to change the status quo, participate actively, and continue to work hard.
Working with Intermediate and Core Beliefs	Therapists guides patients in constructing conceptualizations of their automatic thoughts, confirming, falsifying, or modifying each part of the conceptualization.
Behavioral Techniques	Therapists select behavioral techniques based on their overall case conceptualization and the goals of a particular session.
Relapse Prevention	As the patient's symptoms improve and he or she masters basic skills, the therapist gradually reduces the number of treatments after obtaining the patient's consent and cooperation.
Homework Assignments	The therapist assigns homework, increasing the likelihood that the patient will successfully complete the homework, conceptualize the problem, and review the completed homework.
Requesting Feedback	After the final summary, the therapist asks the patient for feedback on the session.
Summarization	The therapist focuses the patient's attention on the most important points of the session in a positive way, and as the patient progresses, the patient can be asked to summarize the most important points.
Other	The class of 'Other' encompasses less psychologically oriented topics, like financial support and caregivers.

Table 7: Explanations of CBT strategies.