PRMB: Comprehensively Benchmarking Reward Models in Long-Horizon CBT-based Psychological Counseling

Anonymous ACL submission

Abstract

Reward models (RMs) are widely used to align large language models (LLMs), yet their reliability in long-horizon conversational settings remains poorly understood. In cognitive behavioral therapy (CBT)-based counseling, preference judgments depend on session-level coherence, long-term consistency, and therapeutic process fidelity, posing challenges beyond short-context evaluation. We introduce **PRMB**, a benchmark for evaluating reward models in long-horizon, multi-session CBT-based counseling. PRMB is constructed from a combination of real-world and simulated counseling cases using a progressive summarization framework, and comprises over 15k pairwise and Best-of-N preference instances. Evaluating both discriminative and LLM-as-a-judge reward models, we find that state-of-the-art RMs exhibit low accuracy, session-wise degradation, and systematic over-empathizing biases. Moreover, PRMB rankings positively correlate with downstream Best-of-N inference performance across multiple policy models. PRMB provides a foundation for reward modeling in process-oriented conversational domains.

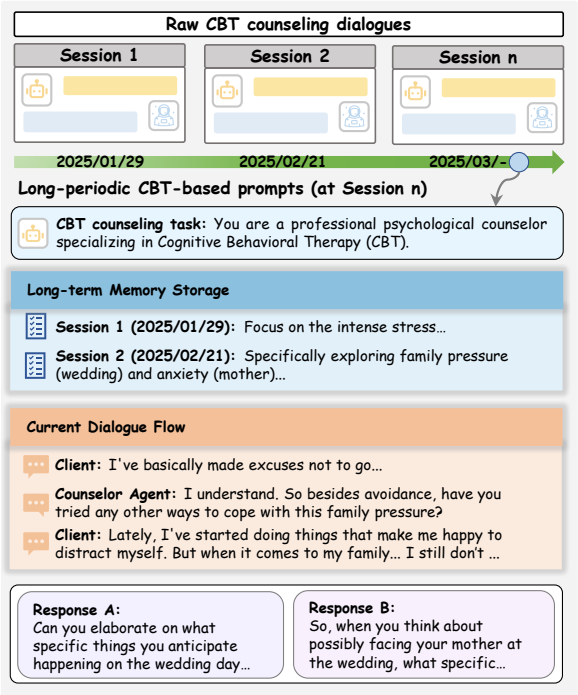

Figure 1: CBT-based counseling dialogues require session-level coherence, long-term consistency across multiple sessions, and adherence to structured therapeutic progression.

1 Introduction

Large language models (LLMs) are increasingly being adopted in mental health-related scenarios, including emotional support conversation, cognitive restructuring, and simulated psychotherapy dialogues (Qiu et al., 2024; Chen et al., 2023; He et al., 2025; Sharma et al., 2024). Among established therapeutic paradigms, Cognitive Behavioral Therapy (CBT) is one of the most structured and evidence-based approaches, making it a natural target for LLM-based simulated counseling systems (Lee et al., 2024). Effective CBT-based counseling requires not only linguistic fluency and helpfulness, but also safety, therapeutic appropriateness, and consistency across sessions. However, evaluating whether an LLM-based simulated counselor satisfies these requirements remains an open challenge.

Recent work has explored using LLM-based judge agents to evaluate responses under predefined criteria (Xie et al., 2025; Zhang et al., 2024a). These judge models are often further adapted as reward models (RMs) to guide reinforcement learning in training simulated counselor agents (Zhang et al., 2025; Wang et al., 2025). Despite their growing use, the reliability of RMs in counseling-oriented settings remains insufficiently validated. Most existing RM benchmarks focus on short-context general dialogue settings, typically involving single-turn or brief multi-turn exchanges that emphasize isolated response quality (Lambert et al., 2025; Zhou et al., 2024; Malik et al., 2025).

In contrast, CBT-based counseling dialogues require session-level coherence, long-term consistency across multiple sessions, and adherence to therapeutic progression, as illustrated in Figure 1. As a result, RM signals derived from short-horizon benchmarks may fail to capture critical process violations that only emerge over extended counseling trajectories, leaving a critical gap in validating RMs for long-horizon counseling tasks.

To address this gap, we introduce **PRMB**, a benchmark for evaluating RMs in long-horizon, multi-session CBT-based counseling. PRMB is constructed using publicly available CBT counseling cases and carefully crafted simulated client personas solely for research and evaluation. For each counseling case, we adopt a progressive summarization framework that incrementally integrates historical information with the current session context, enabling realistic long-context evaluation without exposing raw dialogue histories. Based on this setup, we curate over 15k pairwise preference instances and Best-of-N samples generated by ten representative LLMs.

Using PRMB, we investigate three key dimensions. First, we examine whether existing RMs can generalize across diverse CBT counseling scenarios and long-term interaction contexts. Second, we examine the predictive power of our benchmark for downstream performance by performing Best-of-N sampling. Third, we conduct controlled analyses to examine how different inference-time strategies affect RM in long-horizon CBT counseling evaluation. Through extensive benchmarking and analysis, our results reveal substantial limitations of current RMs in CBT counseling, particularly in terms of consistency, robustness, and downstream alignment. We believe that PRMB provides a necessary foundation for systematically evaluating RMs in process-oriented conversational tasks. The main contributions of this paper are as follows:

- We introduce **PRMB**, a benchmark designed to evaluate reward models in long-horizon, multi-session CBT-based counseling[1].
- We benchmark representative discriminative and LLM-as-judge reward models, revealing significant limitations in their consistency and robustness for counseling-oriented evaluation.
- We conduct analyses of common inference-time strategies across reward models to provide empirical insights on improving RM reliability in long-horizon counseling settings.

[1]All data and code will be released upon acceptance.

2 Related Work

2.1 Reward Model

Reward models (RMs) are a core component of large language model (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 divided into discriminative and generative RMs (Zhang et al., 2024b; Zhou et al., 2024). Discriminative RMs are typically trained on human preference datasets to output scalar scores reflecting relative response quality, and are commonly used as optimization targets 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 (Zhu et al., 2023; Zheng et al., 2023; Kim et al., 2024a), leverage powerful pretrained LLMs to directly evaluate responses via prompting. By providing explicit evaluation criteria, these models can produce preference judgments or detailed rationales without additional fine-tuning. Recent work has demonstrated that strong closed-source models (e.g., GPT-4) can rival traditional discriminative RMs on general-domain preference tasks (Yuan et al., 2024; Sun et al., 2023). However, their reliability in domain-specific and long-horizon evaluation settings remains less understood.

2.2 Benchmarks of Reward Models

Several benchmarks have been introduced to systematically evaluate RM reliability. Early efforts, such as RewardBench (Lambert et al., 2025), provided a unified infrastructure for testing reward models across diverse domains, from open-ended chat to reasoning, and helped establish reward modeling as a research field. RMB (Zhou et al., 2024) propose the Best-of-N (BoN) evaluation as a new benchmark paradigm for assessing RMs. Subsequent efforts have expanded evaluation to more challenging or 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) introduce the first benchmark for reward modeling in role-playing dialogue, where human preferences are nuanced, multi-faceted, and context-dependent. Despite their comprehensiveness, existing bench-

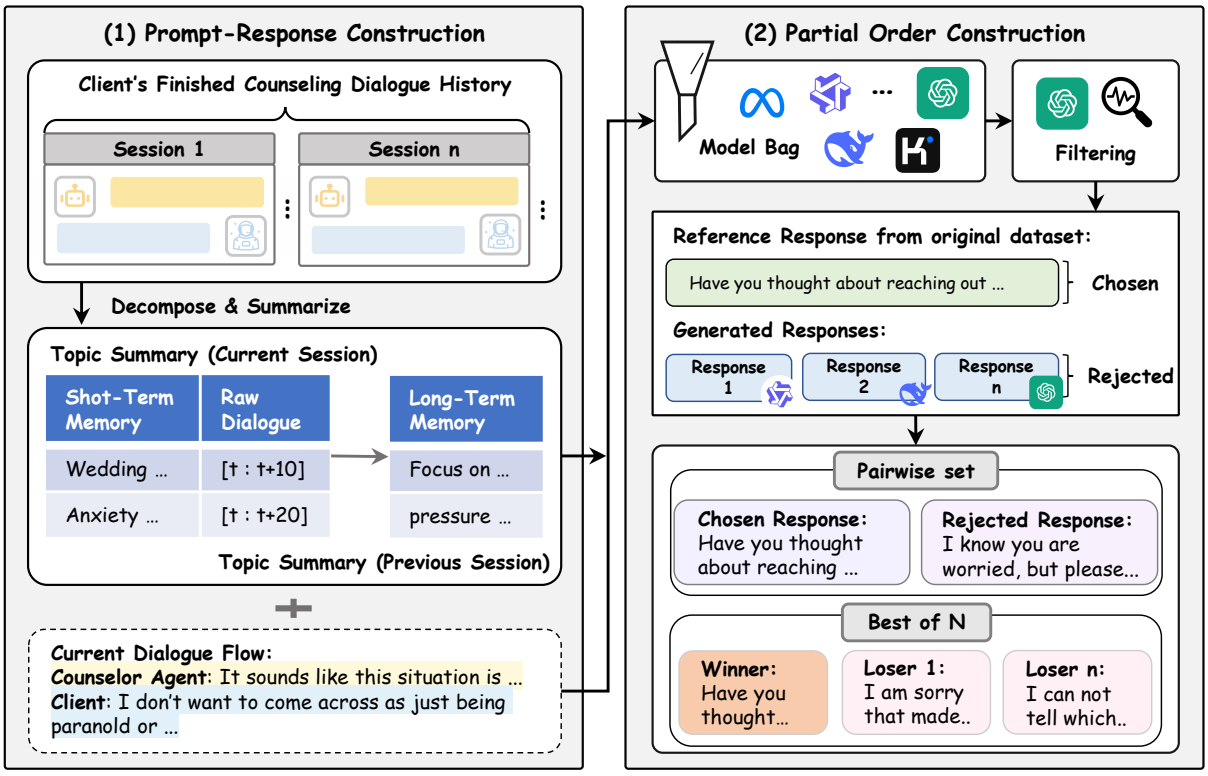

Figure 2: An overview of data construction process: (1) Sampling multi-session dialogues into prompts and obtaining multiple responses for them. (2) Organizing them into pairs or best-of-N lists.

marks predominantly focus on short-context or brief multi-turn interactions, leaving long-horizon preference modeling largely underexplored.

3 Data Construction

Our goal is to construct a benchmark that evaluates whether reward models (RMs) can serve as reliable proxies for CBT counselor preferences in long-horizon counseling scenarios, and whether they can provide effective reward signals for downstream tasks. Following prior work, we construct two types of partial ordering: a pairwise preference set with (*chosen*, *rejected*) response pairs, and a Best-of-N (BoN) set with (*query*, *winner*, *losers*) tuples, where an RM selects the best response among multiple candidates. Figure 2 provide an overall of the construction pipeline.

3.1 Prompt-Response Construction

Counseling case sourcing. We collect CBT counseling cases from publicly available resources, including real-world examples from the American Psychological Association (APA) website[2], 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 12 real-world cases and 106 validated simulated cases, each comprising six sessions.

Prompt design. To simulate long-horizon CBT counseling, each prompt follows a unified format consisting of: (1) a task instruction defining the counselor role, (2) a long-term summary capturing cross-session client states, (3) a short-term summary of the current session so far, and (4) the most recent dialogue turns. We adopt a progressive summarization strategy (Lv et al., 2024), which incrementally integrates prior session information into the current context. This approach preserves key therapeutic trajectories (e.g., core beliefs and prior strategies) while maintaining manageable context length. Prompt templates are provided in the Appendix A. In total, we construct over 15k prompts spanning diverse long-horizon conditions.

Response generation. For each prompt, we generate counselor responses using ten state-of-the-art

[2] https://www.apa.org/pubs/databases/psyctherapy/

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 1: Comparison of representative reward model benchmarks.

LLMs selected for their strong performance on mainstream leaderboards[3] and widespread use as baselines in recent studies. The models span diverse architectures and training paradigms. The complete list and generation settings are provided in the Appendix B.

3.2 Partial Order Construction

Response filtering. We apply a multi-stage filtering process to retain only high-quality, comparable responses. Candidates are removed if they are incomplete, off-topic, excessively repetitive, or contain unsafe content. We further exclude responses with extreme length deviations relative to reference responses from real CBT cases using rule-based checks. This filtering aims to mitigate verbosity-related bias, ensuring that preference judgments reflect counseling quality rather than superficial length differences.

Pairwise preference construction. For each prompt, we form preference pairs by treating the original CBT counselor response from as the chosen response and one model-generated response as the rejected response. We emphasize that reference responses serve as clinically grounded anchors rather than optimal exemplars, acknowledging that their role is to provide stable preference signals rooted in established CBT practice (Chiu et al., 2024). 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 long-horizon CBT domain, we randomly sampled 200 pairs and invited three independent human annotators 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% of the pairs, respectively. Moreover, all three annotators reached full consensus in 84% of the pairs, and the majority vote agreed with our expert labels in 96% of cases. These results demonstrate that, in the majority of instances, clear and consistent preference signals exist, confirming the high reliability of our benchmark. A detailed breakdown of annotation statistics, annotator backgrounds, and the full evaluation protocol is provided in the Appendix C.

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 design tests whether RMs can reliably recover clinically grounded responses in competitive settings, rather than exhaustively capturing all possible high-quality strategies. Examples are provided in the Appendix D.

3.3 Comparison with Existing Benchmarks.

PRMB differs from existing RM benchmarks in both task structure and evaluation focus, as summarized in Table 1. RewardBench, RMB, and RewardBench2 primarily evaluate preferences under short-context settings, targeting general-purpose skills such as factual correctness. Although effective for broad dialogue evaluation, these benchmarks do not capture the process-oriented nature of counseling interactions. Similarly, RoleRMBench focuses on role-playing consistency within a single interaction, but does not model session-level progression preference dependencies.

In contrast, PRMB is designed for long-horizon, multi-session CBT counseling. Preferences in PRMB depend on session-level coherence across counseling progression. Moreover, the progressive summarization framework distinguishes PRMB from prior benchmarks by preserving essential historical information while avoiding full dialogue transcripts, enabling realistic long-context evaluation at scale without exceeding context limits.

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

Reward Model	Pairwise Acc.	BoN Acc.	Overall Acc.
internlm2-1.8b-reward	0.5183	0.2183	0.3683
internlm2-7b-reward	0.4109	0.1643	0.2876
internlm2-20b-reward	0.3713	0.1431	0.2572
Skywork-Reward-V2-Qwen3-8B	0.3125	0.1053	0.2089
Skywork-Reward-V2-Llama-3.1-8B	0.3045	0.0885	0.1965
deepseek-v3.2-exp⋆	0.3046	0.0845	0.1946
Skywork-Reward-V2-Qwen3-1.7B	0.3033	0.0807	0.1920
Skywork-Reward-V2-Qwen3-0.6B	0.2998	0.0739	0.1869
llama-3.2-3b-instruct⋆	0.3165	0.0485	0.1825
Skywork-Reward-V2-Qwen3-4B	0.2750	0.0871	0.1811
Llama-3.1-8B-Instruct-RM-RB2	0.2791	0.0494	0.1643
Qwen3-8B⋆	0.2567	0.0529	0.1548
gpt-4o-mini⋆	0.2141	0.0508	0.1325
gpt-oss-20b⋆	0.1901	0.0445	0.1173

Table 2: The leaderboard of PRMB, ranked by the average of pairwise and BoN accuracy. The benchmark is challenging for even top existing reward models. ⋆ denotes LLM-as-a-judge models.

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 the main result.

4.1 Evaluation Setup

We evaluate both discriminative RMs and generative models under the LLM-as-a-Judge paradigm, assessing their ability to capture counselor-aligned preferences in long-horizon CBT counseling. Discriminative RMs assign a scalar reward to each prompt–response pair, following their standard role in RLHF pipelines. Generative RMs are prompted to directly select the preferred response among candidates following prior LLM-as-a-judge evaluations Zhou et al. (2024).

For each pairwise instance i, the benchmark provides a chosen and rejected response (x_i^+, x_i^-). A discriminative RM is correct when it assigns a higher score to the chosen response. A generative RM is considered correct when it explicitly selects the chosen response. The pairwise accuracy $\mathcal{A}_{\text{pw}}$ is computed as:

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

where $g(\cdot)$ is 1 if the RM prefers chosen over rejected response, otherwise 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 therefore:

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

We adopt these accuracy-based metrics to directly measure the correctness of relative preference judgments, which is the core requirement for RMs to recover grounded rankings and serve as reliable training signals.

4.2 Evaluation Results

Table 2 reports the performance of all evaluated RMs, ranked by the average of pairwise and BoN accuracy.

Comparison across reward models. The highest performing model, INTERNLM2-1.8B-REWARD, achieves 51.83% pairwise accuracy and 21.83% BoN accuracy, substantially lower than performance on conventional RM benchmarks. Notably, larger models within the same family (INTERNLM2-7B and 20B) perform worse, suggesting that standard scaling laws do not directly translate to alignment with long-term counseling preferences. The Skywork-Reward models exhibit relatively consistent but modest performance across base models and scales.

Generative LLM-as-a-judge models also struggle on PRMB. Despite their strong performance on general-purpose evaluation tasks, models such as GPT-4O-MINI and QWEN3-8B achieve pairwise

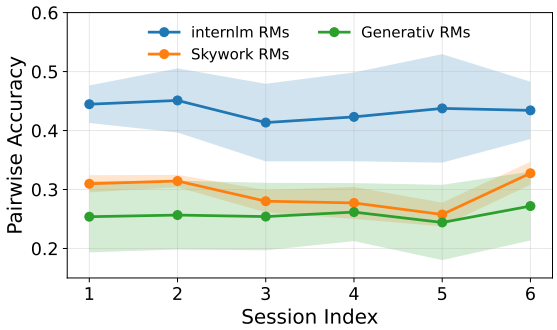

Figure 3: Session-wise pairwise accuracy of reward models on PRMB. Results are aggregated at the model-family level for readability.

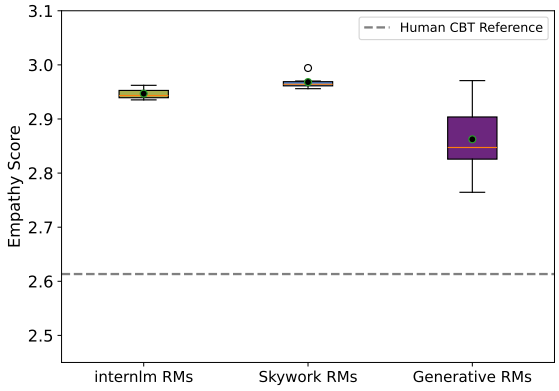

Figure 4: Empathy score distributions of RMs, with the reference CBT responses shown as a reference line.

accuracies below 26% and BoN accuracies close to chance. This indicates limited transfer from generic judgment capabilities to the process-oriented preferences required in CBT counseling.

Overall, PRMB reveals a preference regime that remains challenging for both discriminative and generative reward models, highlighting the need for RMs that explicitly account for long-horizon reasoning and therapeutic consistency.

4.3 Further Analysis

Comparison across counseling sessions. To examine whether RMs maintain consistent preference judgments over extended counseling sessions, we analyze pairwise accuracy cross counseling sessions. As shown in Figure 3, most models exhibit degrading or unstable performance as sessions progress. This trend is particularly pronounced for larger discriminative reward models and for generative LLM-as-judge models, whose accuracy drops substantially in later sessions. In contrast, smaller reward models (e.g., INTERNLM2-1.8B-REWARD) exhibit more stable behavior across sessions. Session-wise results for individual models are reported in Appendix E. These findings indicate that scaling alone does not resolve long-horizon preference modeling and may amplify reliance on superficial cues that degrade as contextual dependencies accumulate.

Impact of Empathy Score We further analyze whether RMs exhibit systematic bias toward empathetic responses. Following prior work (Qian et al., 2023), we use GPT-4 as an automatic empathy evaluator. Figure 4 compares the empathy score distributions of responses preferred by different RMs against reference CBT responses. Across all models, we observe a consistent upward shift in empathy scores relative to the reference, indicating a tendency to favor more empathetic responses. This bias is strongest and most stable for Skywork-based discriminative RMs, while generative models exhibit larger variance. The similarity of this effect across architectures suggests that over-emphasis on surface-level empathetic cues is a general failure mode of current RMs, rather than a consequence of model scale or paradigm. The empathy scores for each individual RM are reported in Appendix F.

5 Downstream Evaluations

A core requirement of any RM benchmark is its ability to predict downstream alignment performance, rather than merely reflecting isolated preference judgments. In this section, we assess whether RM rankings induced by PRMB correlate with real-world effectiveness under Best-of-N (BoN) inference, a widely used test-time strategy.

5.1 Experimental Setup

We assess downstream performance using a held-out set of counseling prompts. For each prompt, we sample 16 candidate responses from a policy model with temperature 0.8. Each RM then scores and ranks the candidates, and the highest-scoring response is selected as the final output. To automatically evaluate generation quality, we compare the selected responses against a reference CBT counselor response using BERTScore (Zhang* et al., 2020). Although BERTScore is a surface-level metric and does not fully capture therapeutic appropriateness, it provides a stable and reproducible proxy for relative quality differences, which is sufficient for evaluating ranking consistency across RMs. We then compute Spearman's rank correlation coeffi-

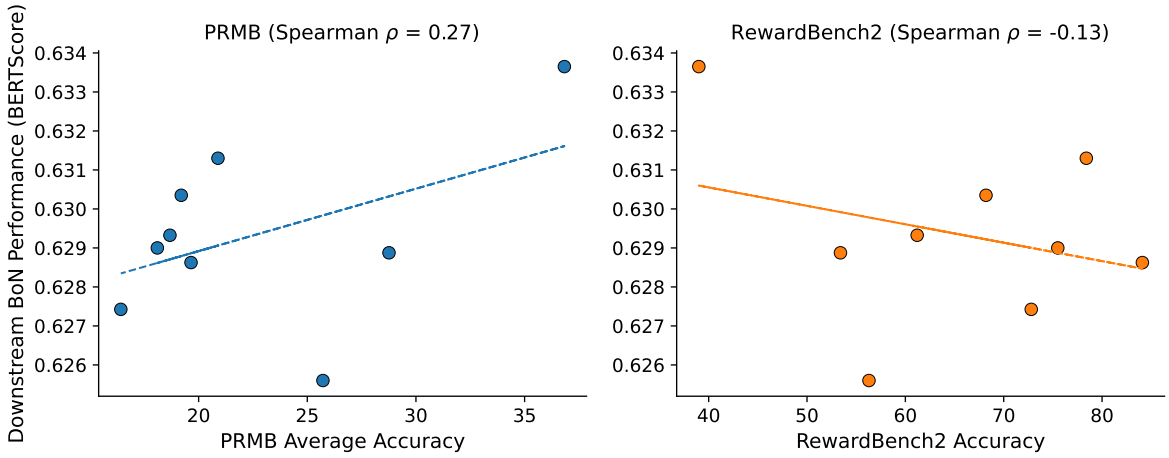

Figure 5: Spearman rank correlation between PRMB rankings and downstream Best-of-N performance, where downstream performance is averaged across four policy models.

cient between RM rankings induced by PRMB and downstream rankings based on BERTScore. Correlations are computed by aggregating downstream BoN performance across four policy models[4]. This evaluation focuses on relative ranking consistency, rather than absolute generation quality.

5.2 Results

Figure 5 shows the correlation between benchmark scores and downstream BoN performance across four diverse policy models, and we additionally report per-policy downstream results in Appendix G. PRMB achieves a positive Spearman rank correlation of $\rho = 0.27$, indicating that RMs with higher PRMB scores tend to select better responses during inference-time BoN sampling. Although the correlation is moderate, this result is notable given the difficulty of long-horizon counseling tasks. The positive correlation suggests that PRMB captures RM behaviors that are broadly predictive of downstream alignment effectiveness, rather than being tied to a specific policy model.

In contrast, RewardBench2 exhibits a negative Spearman correlation ($\rho = -0.13$), indicates that RM rankings induced by RewardBench2 are poorly aligned with the counseling domain, and in some cases inversely aligned. This discrepancy can be attributed to that RewardBench2 focuses on short-turn, general-purpose preference judgments.

Overall, these results indicate that PRMB better reflects RM behaviors that are predictive of inference-time alignment in long-horizon conversational settings.

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

5.3 Further Analysis with CTRS

We further analyze RM behavior using the Cognitive Therapy Rating Scale (CTRS), following prior work showing that GPT-4 can serve as a scalable proxy evaluator for counseling-related dimensions (Lee et al., 2024). We evaluate responses selected under Best-of-N sampling across four policy models, scoring six CTRS dimensions: *collaboration*, *focus*, *guided discovery*, *interpersonal effectiveness*, *strategy*, and *understanding*.

Across all settings, GPT-4 assigns uniformly high CTRS scores, with nearly all responses falling in the 4-5 range. This reflects a clear ceiling effect when clinician-designed rating scales are applied using LLM evaluators. To examine this effect more closely, we randomly sample 200 instances from PRMB and obtain CTRS scores from both a human CBT-trained evaluator and GPT-4. For each instance, we compare the CTRS scores of reference responses (*Chosen*) and their corresponding lower-ranked candidates (*Rejected*). The average scores are shown in Figure 6.

Notably, GPT-4 exhibits a systematic bias, assigning higher CTRS scores overall and even favoring rejected responses over chosen ones. This pattern suggests that LLM-based evaluators tend to overweight surface-level therapeutic signals that are easily captured by language, while failing to reliably assess long-horizon CBT reasoning, session-level coherence, and process fidelity. This miscalibration provides insight into why generative reward models, which rely on similar evaluation mechanisms, perform poorly on PRMB despite appearing clinically strong under static rating schemes.

Figure 6: Average CTRS scores for chosen and rejected responses, evaluated by a human CBT-trained expert and GPT-4 on 200 sampled instances.

Method	Pairwise Acc.	BoN Acc.
gpt-4o-mini	0.2141	0.0508
+ Few-shot (2-shot)	0.1816	0.0252
+ RAG	0.1864	0.0221
+ Self-Refine	0.2063	0.0431
+ CoT	0.1825	0.0353
Qwen3-8B	0.2567	0.0529
+ Few-shot (2-shot)	0.2459	0.0483
+ RAG	0.3379	0.0743
+ Self-Refine	0.2673	0.0435
+ CoT	0.2123	0.0453
llama-3.2-3b-instruct	0.3165	0.0485
+ Few-shot (2-shot)	0.3583	0.0783
+ RAG	0.3799	0.0675
+ Self-Refine	0.3213	0.0431
+ CoT	0.3125	0.0435

Table 3: Effects of inference-time strategies on pairwise and Best-of-N (BoN) accuracy across different generative reward model backbones.

Overall, this analysis demonstrates that absolute CTRS-style ratings approximated by LLM evaluators, lack sufficient resolution to distinguish fine-grained preference differences among strong counseling responses. These findings motivate the design of PRMB as a relative, preference-based benchmark that emphasizes session-level consistency and comparative judgment rather than standalone clinical scores.

6 Discussion

Effects of Inference-Time Strategies. Table 3 reports the impact of several inference-time strategies on generative reward models across different backbones. Overall, the results indicate that such methods exhibit highly inconsistent effects and fail to provide a reliable improvement across models or evaluation settings.

Few-shot prompting yields mixed outcomes. While it improves pairwise and BoN accuracy for LLAMA-3.2-3B-INSTRUCT, it degrades performance on GPT-4O-MINI and provides only marginal gains for QWEN3-8B. This sensitivity suggests that few-shot demonstrations do not generalize well across architectures or preference distributions, particularly in long-horizon counseling scenarios.

Retrieval-augmented generation (RAG) improves performance on some open models, notably QWEN3-8B and LLAMA-3.2-3B-INSTRUCT, but fails to benefit GPT-4O-MINI. Although external CBT knowledge appears helpful for models with limited internalized domain knowledge, RAG alone does not resolve preference misalignment, as improvements remain modest and inconsistent between pairwise and BoN settings.

Self-refinement and explicit chain-of-thought (CoT) reasoning do not consistently improve performance and often lead to degradation, particularly in BoN accuracy. This suggests that iterative or verbose reasoning may amplify surface-level heuristics or reinforce initial misjudgments rather than correcting long-horizon preference errors.

Taken together, these results demonstrate that inference-time heuristics are insufficient for robust reward modeling in process-oriented counseling, reinforcing the need for benchmarks and training objectives that explicitly model session-level dependencies and therapeutic trajectories.

7 Conclusion

In this paper, we propose PRMB, a benchmark designed to evaluate reward models in long-horizon, multi-session CBT-based counseling. Extensive evaluations show that both discriminative and generative reward models struggle to reliably capture counselor-aligned preferences in this setting. We additionally find that inference-time heuristics alone are insufficient for aligning reward models in process-oriented conversational domains. We hope PRMB will serve as a foundation for future research on reward modeling methods that explicitly account for long-horizon structure, relative judgment, and therapeutic trajectories.

Limitations

As far as we known, 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.

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.

Yirong Chen, Xiaofen Xing, Jingkai Lin, Huimin Zheng, Zhenyu Wang, Qi Liu, and Xiangmin Xu. 2023. 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, Singapore. Association for Computational Linguistics.

Yu Ying Chiu, Ashish Sharma, Inna Wanyin Lin, and Tim Althoff. 2024. A computational framework for behavioral assessment of llm therapists. *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, Red Hook, NY, USA. 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, Vienna, Austria. Association for Computational Linguistics.

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, Suzhou, China. Association for Computational Linguistics.

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, Miami, Florida, USA. 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, Albuquerque, New Mexico. Association for Computational Linguistics.

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, Miami, Florida, USA. Association for Computational Linguistics.

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, Miami, Florida, USA. 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.

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, Red Hook, NY, USA. Curran Associates Inc.

Yushan Qian, Weinan Zhang, and Ting Liu. 2023. Harnessing the power of large language models for empathetic response generation: Empirical investigations and improvements. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6516–6528, Singapore. Association for Computational Linguistics.

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, Miami, Florida, USA. 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, New York, NY, USA. 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*.

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.

Haojie Xie, Yirong Chen, Xiaofen Xing, Jingkai Lin, and Xiangmin Xu. 2025. PsyDT: Using LLMs to construct the digital twin of psychological counselor with personalized counseling style for psychological counseling. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1081–1115, Vienna, Austria. Association for Computational Linguistics.

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. 2025. Sentient agent as a judge: Evaluating higher-order social cognition in large language models. *ArXiv*, abs/2505.02847.

Chenhao Zhang, Renhao Li, Minghuan Tan, Min Yang, Jingwei Zhu, Di Yang, Jiahao Zhao, Guancheng Ye, Chengming Li, and Xiping Hu. 2024a. CPsyCoun: A report-based multi-turn dialogue reconstruction and evaluation framework for Chinese psychological counseling. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 13947–13966, Bangkok, Thailand. Association for Computational Linguistics.

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

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 con-

ceptualization 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 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. Instead of exposing full dialogue histories, 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, newly expressed thoughts, and intermediate therapeutic steps, 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.

Summaries are generated using a fixed LLM-based summarization prompt with deterministic decoding. 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.

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.

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.

B Model List Usage in Response Candidate Generation

To generate diverse counselor response candidates for preference construction, we employ a set of ten representative large language models that are widely used in recent alignment and dialogue studies. These models are selected to cover a range of architectures, parameter scales, and training paradigms.

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 4.

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	Open-weight
gpt-oss-20b	Open-weight
MiniMax-M2	Open-weight
Kimi-K2	Open-weight

Table 4: Models used for counselor response candidate generation.

C Agreement with human preference

To validate that the constructed preference pairs reflect human-aligned judgments in long-horizon CBT counseling, we conduct a human preference agreement study on a randomly sampled subset of the dataset.

Annotators. We recruit three independent annotators with formal training in psychology and familiarity with CBT principles. Annotators are not involved in dataset construction and are blinded to the reference labels during annotation. The task is framed as a preference judgment rather than a clinical diagnosis, and annotators are instructed to assess responses based on therapeutic appropriateness, coherence with prior context, and alignment with CBT processes.

Annotation protocol. Each annotator is presented with a counseling prompt and two candidate counselor responses corresponding to a constructed (chosen, rejected) pair. Annotators are asked to select the response they judge to be more appropriate as a CBT counselor response. Ties are not allowed. Annotators do not receive feedback during the annotation process.

D Examples

This section presents representative examples from PRMB to illustrate the structure of evaluation prompts, response candidates, and preference relationships in long-horizon CBT counseling. All examples are simplified and anonymized for clarity, and are provided solely for research illustration purposes.

D.1 Example of Progressive Summary and Evaluation Prompt

We first show an example of an evaluation prompt constructed using the progressive summarization framework. The prompt consists of a task instruction, a long-term summary accumulated from previous sessions, a short-term summary of the current session, and the most recent dialogue context.

Task Instruction.

You are a CBT-based counselor. Your goal is to provide therapeutically appropriate, safe, and context-consistent responses that follow CBT principles.

Long-term Summary (after Session 2).

> Session 1 Summary:
>
> Focus:
>
> - The main discussion topic and triggering situation was the client's emotional distress related to divorce, including anxiety, loneliness, and fear of judgment.
>
> - The key issues in this session included cognitive restructuring and emotion regulation, helping the client identify and challenge negative automatic thoughts.
>
> Progress:

- The client made progress in identifying automatic thoughts, for example, recognizing the contradiction between being harsh on herself and being tolerant of others.

- She began recording her emotional and thought patterns and participated in small experiments to explore the possibility of sharing her divorce experience in social situations.

- The client tried to initially improve her social anxiety by attending dance classes and taking her son to the park alone.

Setbacks / Challenges:

- The client continues to face shame related to divorce and anxiety about being judged by others, especially struggling to cope with emotions in social situations.

- Her concerns about loneliness, adapting to parent-child relationships, and financial pressure remain unresolved issues.

- The recurring automatic and catastrophic thoughts have created some obstacles to her emotion regulation.

Next Steps:

- Recommended next-stage CBT intervention includes further cognitive reappraisal and behavioral experiments to help her gradually face the challenges of social situations.

- Suggest that the client share her thoughts about divorce with close friends in a safe environment, while also rebuilding parent-child activities with her son, such as crafts, to strengthen emotional connection.

- Ask the client to continue recording emotional changes and thought patterns for in-depth analysis and discussion in the next session.

Session 2 Summary:

Focus:

- Primarily discussed anxiety and fear of abandonment caused by divorce and loneliness.

- Key issues included emotion recognition, identification and adjustment of automatic thoughts, and reconstruction of social skills.

Progress:

- The client, Mia, gradually realized the importance of naming emotions and began to identify and record her automatic thoughts.

- Her courage to participate in social activities has increased; she bravely tried to take her son out alone and communicate with friends.

- Mia began using a "courage diary" to record moments of success, demonstrating positive reflection and growth in self-awareness.

Setbacks / Challenges:

- Mia still faces strong self-protective behaviors and a fear of judgment in social situations, often feeling scrutinized at parties.

- Recurring automatic thoughts, such as "I'm a failure" and "Everyone is judging me," continue to trouble her emotions.

- She lacks confidence in future social activities, especially feeling nervous about participating in dance classes alone, but exhibits avoidance tendencies.

Next Steps:

- We recommend that Mia continue recording her automatic thoughts and successful experiences, and try relaxation techniques such as deep breathing to control her anxiety.

- We encourage her to persist in participating in social activities and to schedule alone time after class to promote self-care and help alleviate negative emotions.

- Through small behavioral experiments, we will further enhance her self-efficacy and gradually rebuild her social skills.

Short-term Summary (current session).

In a recent consultation, the client continued to face social anxiety stemming from her divorce and child custody, particularly her self-doubt. She sought support from family and friends, such as her

cousin, demonstrating a growing awareness of the importance of building a support network. She recalled experiences traveling alone with her son Brady and successfully handling school registration; despite the anxiety, these experiences made her realize her growing capabilities. Her feelings of helplessness lessened, and her adaptability, especially when handling urgent work tasks, indicated a positive approach to challenges. While still fearful of handling everything alone in the future, her self-efficacy gradually improved, and she now believes she can solve problems independently to a degree of 50%. She also began proactively contacting old friends, despite lingering concerns about losing support due to the divorce. Overall, she demonstrated a positive attitude throughout the process and, with the therapist's assistance, gradually shifted from negative automatic thoughts to greater self-confidence. Documenting her successes helped solidify this transformation.

Recent Dialogue (15–20 turns).

> **Counselor**: Hmm, hmm. And you also started taking dance classes? You've been very consistent these past four months.
> **Client**: Yes... although it's tiring sometimes, dancing really improves my mood.
> **Counselor**: That's great. I remember when we first met, you felt like you couldn't do anything right, but now you've made so much progress in so many areas.
> **Client**: Yeah... although it hasn't been easy.
> **Counselor**:

This prompt is provided to multiple models to generate counselor response candidates, which are then used to construct pairwise and Best-of-N evaluation instances.

D.2 Pairwise Preference Example

We next present a pairwise preference example constructed from the above prompt.

Chosen Response.

> Yes. So how do you explain it? What enabled you to make these advancements?

Rejected Response.

> You did a fantastic job. Every step forward is worth celebrating. Let's look at how this dance has specifically helped you.

Although both responses are fluent and supportive, the chosen response is preferred because it engages in guided discovery by prompting the client to reflect on the underlying factors that contributed to her progress. In contrast, the rejected response primarily offers reassurance and praise without explicitly advancing the CBT process.

D.3 Best-of-N (BoN) Example

We further illustrate a Best-of-N instance constructed from the same prompt. In this simplified Best-of-4 setting, the task is to identify the most therapeutically appropriate response among multiple candidates.

Winner (Reference Response).

> Yes. So how do you explain it? What enabled you to make these advancements?

Loser 1.

> It sounds like your efforts are starting to bring about positive changes. What specific emotional benefits do you feel dancing has brought you?

Loser 2.

> Dancing may be tiring, but it improves your mood. When you feel it's not easy, what thoughts often come to mind?

Loser 3.

> It sounds like you've put in a lot of effort. Could you elaborate on how this progress came about?

This Best-of-N setup evaluates whether a reward model can consistently prioritize responses that maintain therapeutic focus, align with accumulated session context, and promote CBT-specific reasoning, rather than selecting responses based solely on surface-level empathy.

D.4 Discussion of Example

Together, these examples illustrate the core challenges targeted by PRMB. Responses that appear empathetic and supportive in isolation may not effectively advance long-term therapeutic goals. Correct preference judgments require integrating session history, recognizing therapeutic intent, and prioritizing process-level consistency over immediate emotional validation. This highlights why short-context benchmarks and static clinical rating schemes struggle to differentiate response quality in long-horizon CBT counseling, and motivates the design of PRMB as a relative, preference-based evaluation benchmark.

E Session-wise Accuracy Analysis

This appendix reports per-session accuracy for individual reward models (RMs) on PRMB. Accuracy is computed separately for each counseling session, reflecting the reward model's ability to correctly identify the preferred response at different stages of long-horizon CBT counseling.

Evaluation setting. All results are obtained under the same evaluation protocol as described in the main paper. Each session corresponds to a distinct stage in a six-session CBT counseling trajectory, with progressively richer historical context and therapeutic dependencies. Session-wise accuracy is measured independently and then reported without aggregation, in order to reveal performance variation across counseling stages.

Session-wise results. Table 5 presents the accuracy of each reward model on Sessions 1 through 6.

Discussion. Overall, reward model performance varies substantially across counseling sessions. While some models exhibit relatively stable accuracy across sessions, many show noticeable degradation or fluctuation in later sessions, where longer histories and stronger cross-session dependencies are required. This session-wise breakdown complements the aggregate results reported in the main paper and highlights the challenges of long-horizon preference modeling in CBT counseling.

F Empathy Score Statistics

This section reports detailed empathy score statistics for individual reward models (RMs), supplementing the aggregate analysis presented in the main paper. The goal of this analysis is to examine whether different RMs exhibit systematic biases toward empathetic signals when selecting counseling responses.

Empathy evaluation protocol. Following prior work on empathetic response assessment (Qian et al., 2023), we employ GPT-4 as an automatic empathy evaluator. Given a counseling prompt and a selected response, GPT-4 is prompted to assign an empathy score on a five-point Likert scale, where higher values indicate stronger perceived empathy. The evaluator is instructed to focus on expressed emotional understanding, validation, and supportive language, without assessing overall therapeutic quality or CBT correctness.

To ensure consistency, the same evaluation prompt and decoding configuration are applied across all models. Empathy scores are computed for responses selected by each RM, and averaged across all evaluation instances.

Per-model empathy statistics. Table 6 reports the mean and standard deviation of empathy scores for each individual RM. For reference, we also include the average empathy score of the human CBT reference responses (*Chosen*).

Discussion. Across all evaluated models, responses selected by reward models consistently receive higher empathy scores than the human CBT reference responses. This trend holds across discriminative and generative RMs, as well as across different model scales. Several models assign substantially higher empathy scores despite exhibiting low accuracy on PRMB, indicating that strong empathetic expression alone is insufficient for capturing counselor-aligned preferences in long-horizon CBT counseling.

G Per-policy Downstream Evaluation

This appendix reports per-policy downstream evaluation results using Best-of-N (BoN) sampling guided by different reward models. To simplify comparison, we use the averaged PRMB accuracy as a single aggregated benchmark score for each reward model.

Evaluation setting. For each reward model, we conduct BoN inference using four representative open-source policy models: Qwen2.5-7B-Instruct, LLaMA-3.2-3B-Instruct, Mistral-7B-Instruct-v0.3,

Reward Model	S1	S2	S3	S4	S5	S6
internlm2-1.8b-reward	0.4897	0.5243	0.5016	0.5293	0.5571	0.5006
internlm2-7b-reward	0.4218	0.4347	0.3954	0.3774	0.4222	0.4165
internlm2-20b-reward	0.4224	0.3941	0.3435	0.3627	0.3333	0.3854
Skywork-Reward-V2-Qwen3-8B	0.3274	0.3348	0.2999	0.3089	0.2567	0.3443
Skywork-Reward-V2-Llama-3.1-8B	0.3074	0.3070	0.2880	0.2938	0.2709	0.3506
deepseek-v3.2-exp	0.2560	0.3054	0.3057	0.3079	0.2911	0.3356
Skywork-Reward-V2-Qwen3-1.7B	0.3050	0.3136	0.2971	0.2882	0.2759	0.3336
Skywork-Reward-V2-Qwen3-0.6B	0.2802	0.3224	0.2813	0.2941	0.2854	0.3223
Skywork-Reward-V2-Qwen3-4B	0.2861	0.3081	0.2639	0.2635	0.2195	0.3048
Llama-3.1-8B-Instruct-RM-RB2	0.3227	0.3085	0.2520	0.2315	0.2651	0.3045
llama-3.2-3b-instruct (Generative RM)	0.3493	0.3191	0.3150	0.3137	0.3218	0.3242
Qwen3-8B (Generative RM)	0.2450	0.2695	0.2663	0.3001	0.2356	0.2956
gpt-4o-mini (Generative RM)	0.2242	0.2252	0.2116	0.2156	0.1931	0.2209
gpt-oss (Generative RM)	0.1859	0.1769	0.1840	0.2096	0.1704	0.2069

Table 5: Session-wise accuracy of individual reward models on PRMB.

Reward Model	Mean Empathy	Std.
Human CBT Reference (Chosen)	2.6134	0.42
internlm2-1.8b-reward	2.9353	0.31
internlm2-7b-reward	2.9433	0.29
internlm2-20b-reward	2.9622	0.27
Skywork-Reward-V2-Qwen3-8B	2.9560	0.18
Skywork-Reward-V2-Llama-3.1-8B	2.9942	0.19
deepseek-v3.2-exp	2.8258	0.34
Skywork-Reward-V2-Qwen3-1.7B	2.9654	0.21
Skywork-Reward-V2-Qwen3-0.6B	2.9607	0.23
Skywork-Reward-V2-Qwen3-4B	2.9701	0.20
Llama-3.1-8B-Instruct-RM-RB2	2.9629	0.28
llama-3.2-3b-instruct (Generative RM)	2.7645	0.37
Qwen3-8B (Generative RM)	2.8473	0.37
gpt-4o-mini (Generative RM)	2.9709	0.34
gpt-oss (Generative RM)	2.9037	0.36

Table 6: Empathy score statistics for individual reward models and human CBT reference responses.

and Gemma-3-4B-IT. For each policy, multiple candidate responses are sampled and ranked according to the reward model, and the top-ranked response is selected. The selected responses are evaluated using the same automatic evaluation protocol described in the main paper.

Results. Table 7 presents downstream performance across different policy models, together with the final PRMB score (averaged over pairwise and BoN evaluations) and RewardBench2 results.

Discussion. Using a single aggregated PRMB score simplifies comparison across reward models and reveals a clear trend: reward models with higher PRMB scores consistently yield stronger downstream performance across different policy models. In contrast, RewardBench2 scores exhibit weaker and less consistent alignment with downstream BoN results in CBT counseling scenarios. This further supports the suitability of PRMB as an evaluation benchmark for long-horizon, process-oriented conversational alignment.

Reward Model	Qwen2.5-7B	LLaMA-3.2-3B	Mistral-7B	Gemma-3-4B	PRMB Avg Acc.	RB2
Skywork-Reward-V2-Qwen3-0.6B	0.6404	0.6335	0.6172	0.6262	0.1869	61.2
Skywork-Reward-V2-Qwen3-1.7B	0.6422	0.6335	0.6204	0.6253	0.1920	68.2
Skywork-Reward-V2-Qwen3-4B	0.6408	0.6336	0.6174	0.6242	0.1811	75.5
Skywork-Reward-V2-Qwen3-8B	0.6421	0.6369	0.6208	0.6254	0.2089	78.4
Skywork-Reward-V2-Llama-3.1-8B	0.6405	0.6322	0.6185	0.6233	0.1965	84.1
internlm2-1.8b-reward	0.6431	0.6388	0.6286	0.6241	0.3683	39.0
internlm2-7b-reward	0.6429	0.6377	0.6114	0.6235	0.2876	53.4
internlm2-20b-reward	0.6425	0.6371	0.6126	0.6292	0.2572	56.3
LLaMA-3.1-8B-Instruct-RM-RB2	0.6379	0.6319	0.6161	0.6238	0.1643	72.8

Table 7: Per-policy downstream Best-of-N performance guided by different reward models. PRMB Avg Acc. denotes the final benchmark score averaged over pairwise and BoN evaluations.