arXiv:2603.11494v1 [cs.DB] 12 Mar 2026

PRMB：长时程认知行为疗法咨询对话中的奖励模型基准测试

Yougen Zhou、Qin Chen、Ningning Zhou、Jie Zhou 和 Liang He

摘要—**大语言模型（LLM）在心理健康医疗应用中具有潜力，尤其是在基于认知行为疗法（CBT）的咨询中；在这一领域，奖励模型对于使大语言模型与期望的治疗行为保持对齐起着关键作用。然而，由于过程导向数据集覆盖有限，以及评估目标与心理对齐目标不一致，现有的奖励模型评估往往无法反映其在长时程干预中的对齐效果。为解决这些局限，我们提出 PRMB，这是一个专为评估多会话认知行为疗法咨询中的奖励模型而设计的综合基准。PRMB 涵盖 6 次会谈和 21 种不同的负面情景，并纳入成对比较和 Best-of-N 偏好评估。我们证明，该基准与下游咨询对话表现呈正相关。基于该基准，我们对最先进的奖励模型进行了广泛分析，揭示了先前基准未能发现的泛化缺陷，并凸显了生成式奖励模型的潜力。此外，我们深入考察了推理时策略在奖励模型评估中的有效性，并分析了生成式奖励模型的影响因素。本研究通过建立心理健康对话中奖励模型评估框架，推动了面向个性化医疗的智能信息学发展。评估代码和数据集已公开发布于 https://github.com/YouKenChaw/PRMB。**

索引词—**大语言模型、奖励模型、基准、认知行为疗法、心理咨询对话。**

I. 引言

LARGE 语言模型（大语言模型）正越来越多地应用于心理健康相关场景，包括情感支持对话、认知重构和心理治疗对话 [1]–[4]。在既有的治疗范式中，认知行为疗法（CBT）是结构最为严谨且循证基础最为充分的方法之一，因此自然成为基于大语言模型的模拟咨询系统的目标 [5]。有效的认知行为疗法咨询不仅要求语言流畅且具有帮助性，还要求安全性、治疗适切性，以及跨会谈的一致性。然而，评估基于大语言模型的模拟咨询师是否满足这些要求，仍是一项尚未解决的挑战。近期研究探索了使用基于大语言模型的评判智能体，依据预定义标准评估回应 [6], [7]。这些评判模型通常会进一步调整为奖励模型，以指导强化学习，训练模拟咨询师智能体[8], [9]。尽管奖励模型的应用日益广泛，但其在咨询场景中的可靠性仍未得到充分验证。

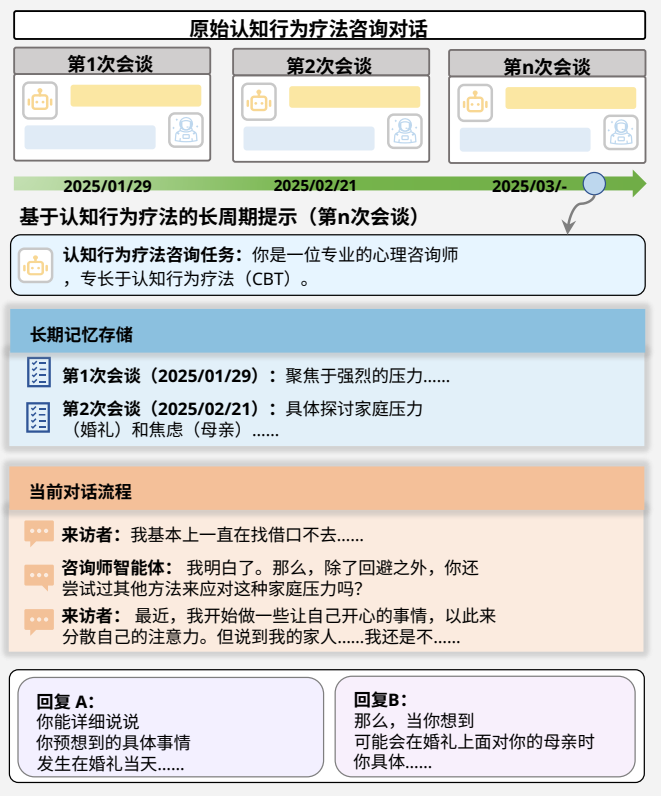

图1. 基于认知行为疗法的咨询对话需要会话级连贯性、跨多个治疗会话的长期一致性，以及对结构化治疗进程的遵循。

现有的大多数奖励模型基准侧重于短上下文的通用对话场景，通常涉及单轮或简短的多轮交互，强调孤立的回复质量 [10]–[12]。相比之下，基于认知行为疗法的咨询对话需要会话级连贯性、跨多个治疗会话的长期一致性，以及对治疗进程的遵循，如图1所示。因此，从短程基准中得出的奖励模型信号可能无法捕捉仅在延长的咨询轨迹中才会出现的关键过程违规，从而在验证用于长程咨询任务的奖励模型方面留下了关键空白。

为弥补这一空白，我们推出了**PRMB**，这是一个用于评估长时程、多会谈认知行为疗法咨询中奖励模型的基准。PRMB 基于公开可用的

Yougen Zhou就职于华东师范大学上海人工智能教育研究院，中国上海（电子邮箱：zyg@stu.ecnu.edu.cn）。Ningning Zhou就职于华东师范大学心理与认知科学学院，中国上海（电子邮箱：nnzhou@psy.ecnu.edu.cn）。Qin Chen、Jie Zhou和Liang He就职于华东师范大学计算机科学与技术学院，中国上海（电子邮箱：qchen@cs.ecnu.edu.cn）。Qin Chen为通讯作者。

认知行为疗法咨询案例构建，并采用精心设计的模拟来访者角色，仅用于研究和评估。对于每个咨询案例，我们采用渐进式总结框架，逐步整合历史信息与当前会话上下文，从而在不暴露原始对话历史的情况下进行逼真的长上下文评估。基于这一设置，我们整理了超过 1.2 万个成对偏好实例，以及由十种具有代表性的大语言模型生成的 Best-of-N 样本。

借助 PRMB，我们研究了三个关键维度。首先，我们考察现有奖励模型能否在不同的 CBT咨询场景和长期交互上下文中实现泛化。其次，我们通过 Best-of-N 采样，考察该基准对下游性能的预测能力。第三，我们通过对照分析，研究不同的推理时策略如何影响长时程认知行为疗法咨询评估。通过广泛的基准测试与分析，我们的结果揭示了当前奖励模型在认知行为疗法咨询中的显著局限，尤其体现在一致性、稳健性和下游对齐方面。我们认为，PRMB 为系统评估面向过程的对话任务提供了必要的基础。本文的主要贡献如下：

- 我们提出了 **PRMB**，这是一个旨在评估长时程、多会话认知行为疗法咨询对话中奖励模型的基准。
- 我们对具有代表性的判别式奖励模型和以大语言模型为评判者的奖励模型进行了基准测试，揭示了它们在面向咨询的评估中存在显著的一致性和稳健性局限。
- 我们分析了不同奖励模型的推理时策略，以提供实证见解，帮助提升奖励模型在长时程咨询对话场景中的可靠性。

本文其余部分安排如下。第二节回顾基于大语言模型的模拟心理对话、奖励模型以及用于奖励模型评估的现有基准等相关工作。第三节详细介绍我们基准的构建过程及统计数据。第四节展示对最先进奖励模型的基准评测结果。第五节评估我们的基准与下游咨询对话之间的相关性。第六节讨论推理时策略的有效性，并分析奖励模型的影响因素。最后，第七节总结全文并概述未来研究方向。

II. 相关工作

A. 模拟心理对话

大语言模型（LLMs）通过实现更连贯且更具情感支持性的互动，推动了治疗导向型对话的发展 [13]–[16]。后续研究纳入咨询知识和结构化治疗策略，以提升心理支持质量 [17]–[19], ，着重于症状识别和策略条件化的回复生成。近年来，强化学习（RL）被引入，以使对话智能体与治疗目标保持对齐 [9], [20]。研究者利用情感感知奖励信号，优化共情行为和感知支持质量。这使优化重点从文本相似度转向更高层次的心理结果，也让奖励模型成为治疗对话系统中实现行为对齐的关键组成部分。然而，奖励模型通常被视为特定任务的模块，其有效性以及与心理治疗原则的一致性鲜少得到检验。尤其是，几乎没有研究系统评估奖励模型是否能够捕捉长时程治疗目标。这一空白凸显了在基于认知行为疗法的长时程心理咨询场景中，针对奖励模型开展专项基准评测的必要性。

B. 奖励模型

奖励模型（RM）是大语言模型（LLM）对齐的核心组成部分，可提供偏好信号，引导模型生成符合人类偏好的回答 [21]–[23]。奖励模型大致分为判别式和生成式两类 [11], [24]。判别式奖励模型通常基于人类偏好数据集进行训练，输出反映回复质量相对高低的标量分数，并常作为基于人类反馈的强化学习（RLHF）流程中的优化目标 [10], [21], [25]。生成式奖励模型（通常称为大语言模型评审） [26]–[28], 则借助强大的预训练大语言模型，通过提示直接评估回答。通过提供明确的评估标准，这些模型无需额外微调即可给出偏好判断或详细的评估理由。近期研究表明，在通用领域的偏好任务中，能力强大的闭源模型可媲美传统判别式奖励模型 [29],[30]。然而，这些模型在特定领域和长时程评估场景中的可靠性仍未得到充分了解。

C. 奖励模型基准测试

已有多项基准被提出，用于系统评估奖励模型的可靠性。早期工作（如 RewardBench [10], ）为跨不同领域测试奖励模型提供了统一框架，涵盖开放式聊天到推理等任务，并推动奖励建模成为一个研究领域。RMB [11] 提出将最佳N项（BoN）评估作为评估奖励模型的新基准范式。后续研究将评估扩展到更具挑战性或更专业的场景，例如多语言性[31], 、数学推理 [32],和智能体行为 [33]。RoleRMBench [34] 推出了首个针对角色扮演对话奖励建模的基准；在这类对话中，人类偏好细腻、多面且依赖上下文。尽管现有基准涵盖面广，但它们主要关注短上下文或简短的多轮交互，对长时程偏好建模的探索仍然不足。

III. 数据构建

我们的目标是构建一个基准，用于评估奖励模型（RM）能否在长时程咨询场景中可靠地代表认知行为疗法（CBT）咨询师偏好，以及能否为下游任务提供有效的奖励信号。遵循先前的工作，我们构建了两种偏序：一个成对偏好集，其中包含

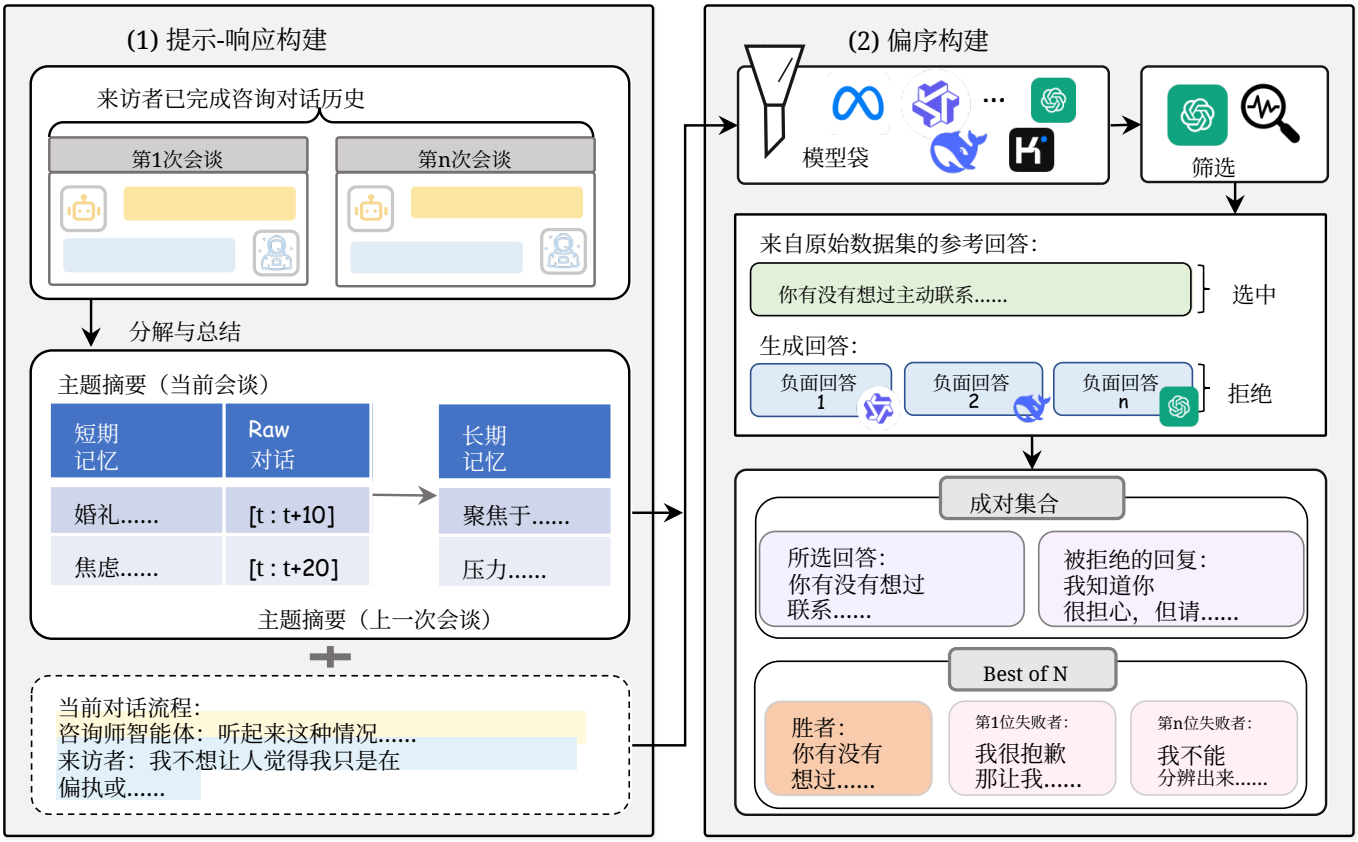

图2. 数据构建流程概览 流程：（1）将多轮会话对话采样为提示，并为其获取多个回复。（2）将它们整理成成对数据或 Best-of-N 列表。

（选中，拒绝）响应对；以及一个最佳N项（BoN）集，其中包含（查询，胜者，落败者）元组，在该元组中，奖励模型从多个候选项中选出最佳响应。图 2 展示了构建流程的整体概览。

A. 提示-响应构建

1）咨询案例来源：我们从公开资源中收集认知行为疗法咨询案例，包括来自美国心理学会（APA）网站的真实案例[1]，认知行为疗法教科书，以及来自DiaCBT的模拟多会话案例 [35]，一个在认知行为疗法指导约束下创建的大规模合成数据集。所有案例仅用于研究和评估。每个案例均经过筛查，以确保覆盖核心认知行为疗法阶段，包括问题构建、自动思维识别、认知重构和行为干预。经过筛选和验证后，我们保留了118个已验证案例，每个案例包含六次治疗会话。

2）提示设计：为模拟长时程认知行为疗法咨询，每条提示均采用统一格式，包括：（1）定义咨询师角色的任务指令；（2）概括跨会谈来访者状态的长期摘要；（3）概括当前会话截至目前情况的短期摘要；以及（4）最近的对话轮次。我们采用渐进式摘要策略[36]，，逐步将此前会谈的信息整合到当前上下文中。这种方法既能保留关键的治疗轨迹（例如核心信念和既往策略），又能将上下文长度控制在适当范围内。所用提示见附录 A-1。我们总共构建了逾 1.3 万条提示，涵盖多种长时程情境。

3）响应生成：对于每个提示，我们使用十个先进的大语言模型生成咨询师响应。这些模型因在主流排行榜[2]上表现出色，并被近期研究广泛用作基线模型而入选。模型涵盖了多种架构和训练范式。完整列表及生成设置见附录 A-2。为构建稳健的偏好基准，我们针对每个提示生成两类响应：**正向响应**直接取自来源认知行为疗法案例中的原始咨询师话语。这些响应是具有临床依据的选定响应，体现了成熟的认知行为疗法实践，可作为偏好信号稳定且循证的锚点。尽管真实的治疗响应可能有所不同，但这些响应以专业标准为根基，可提供可靠的参照点 [37]。**负向响应**则通过针对同一批大语言模型的提示工程生成。我们借鉴了 [38], 的定性元分析，该分析归纳出来访者报告的 21 种负面体验元类别，分属四个类别（治疗师的不当行为、妨碍关系的方面、治疗匹配度差以及负面治疗影响）。负面

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

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

变体以正向基线提示为基础，但会引入受控且细微的偏差，以体现一个或多个类别。这些偏差经过设计，意在隐蔽（只有专家仔细审查时才能察觉），从而模拟心理治疗中真实、累积的微小伤害，而不表现出明显的有害性。

4) 响应过滤：我们采用多阶段筛选流程，仅保留高质量且具有可比性的回复。对于不完整、偏离主题、过度重复或含有不安全内容的候选回复，予以剔除。我们还通过基于规则的检查，排除与真实认知行为疗法案例中的参考回答相比长度偏差过大的回复。这一筛选流程可减轻冗长程度带来的偏差，确保偏好判断反映的是咨询质量，而非表面上的长度差异。

B. 偏序构建

1) 成对偏好构建：对于每个提示，我们将来源案例中咨询师的原始回复视为选中回复，并将一个由模型生成的负面回复视为被拒绝 回复，以此构建偏好对。为提高难度和区分度，我们仅保留质量差距适中的偏好对，并剔除质量明显低劣的回复。被拒绝的回复从多个大语言模型中抽样，以平衡模型分布。

2) 与人类偏好的一致性：为验证我们在复杂的长时程认知行为疗法领域构建的偏好对的质量，我们随机抽取了200对，并邀请三位独立的人工标注员（持证认知行为疗法从业者）分别选出更受偏好的咨询回复（不允许选择平局）。标注员与我们构建的偏好对表现出高度一致，分别在86%、92%和94%的偏好对中选中了选中回复。此外，三位标注员在84%的偏好对上达成完全一致，多数投票结果在96%的案例中与专家标签一致。这些结果表明，在大多数情况下，偏好信号明确且一致，证实了我们基准的高可靠性。

3）最佳 *N* 项集构建：此外，我们使用最佳 4 项配置创建了一个 BoN 评估集。对于每个查询，参考回答被指定为胜者，而四个模型生成的回答（不同于成对集合中使用的回答）则作为败者。这一设计用于测试奖励模型能够在竞争性设置中可靠地生成符合临床依据的回答，而非穷尽所有可能的高质量策略。

C. 基准的详细统计

该基准包含从 118 个认知行为疗法案例（每个案例包含 6 次治疗会话）中衍生的 13,000 多条提示，由此生成大量成对偏好对和最佳N项（BoN）查询。下文将展示这些数据在 6 次治疗会话、14 种核心认知行为疗法策略和 21 种负面体验类别中的分布，凸显该数据集的规模与均衡性。

1）各会谈的分布：该基准包含 13893 条提示，分布于每个案例的六次治疗会话中。如图 3 所示，提示数量从第一会话到第六会话逐渐增加，其中第六会话所占比例最高（19.61%）。

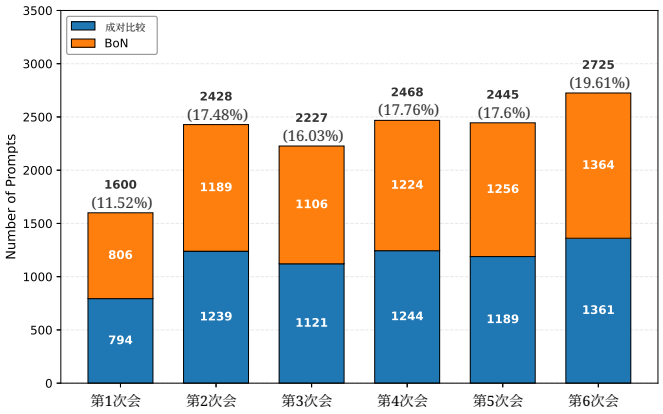

图 3. 各治疗会话中的提示分布（每个柱形上方标注百分比）、成对偏好对数量和最佳N项查询数量。成对比较和 BoN 的数量标注在相应的堆叠区段内。会话主题：（1）信息收集、目标设定和案例概念化；（2）识别并质疑自动化思维；（3）讨论案例概念化；（4）调整中间信念和核心信念；（5）复发预防；（6）巩固。

表 I 提示、成对比较与最佳N项查询在各治疗策略中的分布。

策略	提示（%）	成对比较	BoN
信息收集	811 (5.84)	404	407
制定议程	580 (4.17)	278	302
每周回顾	411 (2.96)	231	180
制定治疗目标	165 (1.19)	85	80
心理教育	1156 (8.32)	571	585
处理自动思维	3973 (28.60)	1990	1983
动机增强	1905 (13.71)	968	937
处理中间信念和核心信念	799 (5.75)	400	399
行为技术	1087 (7.82)	547	540
复发预防	553 (3.98)	257	296
家庭作业	823 (5.92)	419	404
征求反馈	457 (3.29)	227	230
总结	481 (3.46)	239	242
其他	692（4.98）	332	360
总计	13893（100.00）	6948	6945

成对偏好对（6948 对）和 BoN 查询（6945 条）在各会谈中的分配相对均衡，从而确保能够对不同治疗阶段进行稳健评估。

2）核心认知行为疗法策略的分布：表I总结了基准中14种核心认知行为疗法策略的覆盖情况。出现频率最高的策略是“处理自动思维”（28.60%），其次是“动机增强”（13.71%）和“心理教育”（8.32%），这与认知行为疗法的核心认知和准备性要素相符。出现频率较低的策略，如“制定治疗目标”（1.19%）和“复发预防”（3.98%），往往出现在治疗早期或后期阶段，这与典型的治疗进程一致。这些分布确保基准既能捕捉长时程的时间动态，也能体现特定认知行为疗法技术的内容多样性，从而为评估奖励模型在心理治疗中识别细微偏好信号的能力提供具有临床代表性的测试平台。

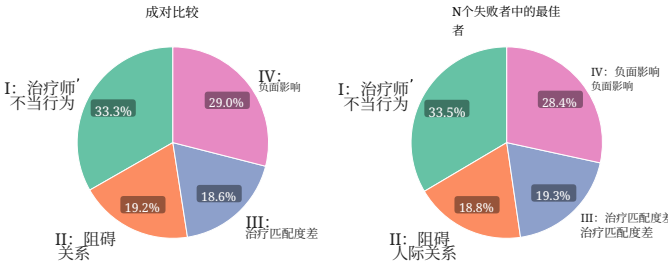

图4. 成对比较和 BoN 失败者集合中负面体验类别的百分比分布，按四类分类体系分组。

表II 主要子类别的分布。百分比相对于每个集合中的负面体验总数计算（而非集合内类别）。

子类别	成对比较（%）	BoN（%）
簇 I：治疗师的不当行为		
治疗师不倾听	4.63%	4.96%
治疗师未理解	4.76%	4.67%
治疗师被认为不称职	4.59%	4.74%
治疗师贬低来访者	4.72%	4.86%
治疗师评判	4.88%	4.76%
治疗师利用来访者谋取私利	4.92%	4.73%
其他不当言语反应	4.78%	4.81%
第二类：阻碍关系		
疏远和/或缺乏共情	4.94%	4.61%
不安全感或不信任	4.56%	4.72%
困惑或不确定	5.04%	4.81%
与治疗师的人际关系不匹配	4.63%	4.62%
第三类：治疗匹配度差		
对实际方面的负面评价 4.43%		4.71%
期望未得到满足	4.56%	4.81%
与干预措施不匹配	4.88%	4.76%
对治疗结束不满意	4.71%	4.99%
簇 IV：治疗的负面影响		
没有变化或变化不足	4.88%	4.75%
治疗后问题加重	5.38%	4.68%
对治疗过程感到恐惧	4.99%	4.73%
动机或希望丧失	4.53%	4.75%
治疗期间的不愉快感受	4.69%	4.75%
治疗中唤起的负面认知	4.48%	4.77%

3）负面体验类别的分布：图4按簇汇总了负面体验类别的分布，采用了 [38]提出的四簇分类体系。在成对集合和 BoN 集合中，治疗师的不当行为（簇 I）占比最高（分别为33.3%和33.5%），其次是治疗的负面影响（簇 IV，分别为29.0%和28.4%）。簇 II 和簇 III（妨碍关系的方面和治疗匹配度差）的占比相对较低，但仍相当可观。表II进一步详细列出了各簇中较突出的子类别。成对集合中最常见的子类别是治疗后问题加重（5.38%）、困惑或不确定（5.04%）和对治疗过程感到恐惧（4.99%）。BoN 集合中最常见的三项则是治疗师不倾听（4.96%）、治疗师贬低来访者（4.86%）和其他不当言语反应（4.81%）。四个簇中的分布均衡且具有临床意义，各子类别的占比也相对均匀，这确保了基准——mark充分捕捉了来访者所报告的负面体验的广度，从而能够稳健评估奖励模型对细微治疗伤害的敏感性。

D. 与现有基准的比较

PRMB在任务结构和评估重点上均不同于现有奖励模型基准，如表III所示。RewardBench、RMB和RewardBench2主要在短上下文设置下评估偏好，针对事实正确性等通用技能。尽管这些基准对广泛的对话评估有效，但它们未能捕捉咨询互动的过程导向性质。同样，RoleRMBench关注单次互动中的角色扮演一致性，但并未建模会话级进程偏好依赖关系。

相比之下，PRMB专为长时程、多会谈认知行为疗法咨询而设计。PRMB中的偏好依赖于贯穿咨询进展的会话级连贯性。此外，渐进式摘要框架通过保留关键历史信息、同时避免完整对话记录，使PRMB有别于先前基准，从而能够在不超出上下文限制的情况下大规模开展真实的长上下文评估。

IV. 奖励模型基准测试

我们在 PRMB 上评估了多种最先进的奖励模型（RM）。本节介绍评估设置和主要结果。

A. 评估设置

我们在大语言模型评审范式下评估判别式奖励模型和生成式模型，考察它们在长期认知行为疗法咨询中捕捉与咨询师一致的偏好的能力。判别式奖励模型按照其在基于人类反馈的强化学习流程中的标准作用，为每个提示—响应对分配一个标量奖励。根据先前采用大语言模型评审的评估方式，生成式奖励模型会收到提示，直接从候选项中选出首选回应 [11]。

对于每个成对实例 i，基准会提供一个选中回答和一个被拒绝的回复（x_i^+，x_i^-）。若判别式奖励模型为选中回答赋予更高分数，则视为判断正确。若生成式奖励模型明确选择选中回答，则视为判断正确。成对准确率$\mathcal{A}_{\text{pw}}$ 的计算方式为：

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

其中，如果奖励模型更偏好选中的回复而非被拒绝的回复，则 $g(\cdot)$ 为 1；否则为 0。

对于每个最佳N项实例 i，基准都会提供一个胜者$x_i^\star$和一组败者 $\{x_{ij}^-\}_{j=1}^{P_i}$。奖励模型只有在实例 i中对胜者的偏好超过所有败者时，才算成功。最佳样本准确率$\mathcal{A}_{\text{BoN}}$ 因此为：

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

表III 代表性奖励模型基准比较。

基准	Best-of-N（N ＞ 2）	多会话	未见提示	目标技能
RewardBench [10]	✗	✗	✗	通用
RMB [11]	✓	✗	✗	通用
RewardBench2 [12]	✓	✗	✓	通用
RoleRMBench [34]	✗	✗	✗	角色扮演
PRMB（我们的方法）	✓	✓	✓	认知行为疗法咨询

表 IV PRMB 排行榜，按成对比较和最佳样本准确率的平均值排序。 ⋆ 表示大语言模型评审模型。

奖励模型	成对准确率	N选优准确率	总体准确率
Llama-3.1-8B-Instruct-RM-RB2	0.8646	0.6700	0.7673
deepseek-v3.2-exp⋆	0.7796	0.7344	0.7570
gpt-4o-mini⋆	0.8368	0.6266	0.7317
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
Qwen2.5-7B-Instruct⋆	0.7860	0.5584	0.6722
Gemma3-4B-IT⋆	0.7533	0.5006	0.6270
Skywork-Reward-V2-Qwen3-0.6B	0.7596	0.4825	0.6211
gpt-oss-20b⋆	0.7500	0.4862	0.6181
Mistral-7B-Instruct-v0.3⋆	0.7306	0.4937	0.6122
Skywork-Reward-V2-Llama-3.1-8B	0.7306	0.4556	0.5931
internlm2-1.8b-reward	0.7047	0.3950	0.5499
LLaMA-3.2-3B-Instruct⋆	0.6513	0.4001	0.5257

我们采用这些基于准确率的指标，直接衡量相对偏好判断的正确性。这是奖励模型还原有依据的排序并提供可靠训练信号的核心要求。

B. 评估结果

1) 奖励模型比较：表 IV 报告了所有受评奖励模型的性能，并按成对比较准确率和 BoN 准确率的平均值排序。排行榜显示，这些模型在 PRMB 上呈现出明显的性能梯度：最强模型的成对比较准确率达到 80% 的中高段，但 BoN 准确率很少超过 70%，因此总体得分介于 52.57% 至 76.73% 之间，显著低于通用奖励建模基准通常可达到的上限。

表现最佳的模型 LLAMA-3.1-8B-INSTRUCT- RM-RB2 达到 86.46% 的成对准确率和 67.00% 的最佳样本准确率（总体为 76.73%）。排名第二的模型 DEEPSEEK-V3.2-EXP（一个 LLM 评判模型）紧随其后，成对准确率为 77.96%，最佳样本准确率为 73.44%（总体为 75.70%）。SKYWORK-REWARD-V2 系列表现出很强的一致性和单调扩展特征：随着容量增加，性能稳定提升（8B ＞ 4B ＞ 1.7B＞ 0.6B），成对准确率范围为 75.96% 至 83.65%，最佳样本准确率范围为 45.25% 至 60.92%。INTERNLM2-REWARD 系列的扩展趋势虽非单调，但总体呈正向：成对准确率从 70.47%（1.8B）提升至 79.74%（7B），再提升至 81.22%（20B）；总体准确率则从 54.99% 提升至 68.25%，再提升至 69.40%，但从 7B 到 20B 的增益明显减弱。这一模式表明尽管更大的容量通常有助于在该领域实现更好的对齐，但由于长时程认知行为疗法偏好的专业性和过程敏感性，超过中等规模模型后，规模收益会变得十分有限。

生成式 LLM 评判模型的迁移能力表现不一，但总体有限。GPT-4O-MINI（成对比较准确率为 83.68%，BoN 准确率为 62.66%）和 QWEN2.5-7B-INSTRUCT（成对比较准确率为 78.60%，BoN 准确率为 55.84%）等高容量评判模型在成对比较中表现尚可，但在 BoN 上的表现明显下降，往往接近或低于 60%。这表明，通用的评价推理难以轻易泛化到长期心理治疗所需的细腻且注重一致性的偏好信号。

总体而言，PRMB 对判别式和生成式奖励模型都提出了挑战，凸显出它们在长时间交互中以及面对多个细微负向干扰项时维持治疗偏好对齐能力方面的不足。

2）成对比较与 *BoN* 的比较：在几乎所有模型中，成对比较和 BoN 评估协议之间都存在显著且一致的性能差距。

成对准确率介于 65.13%（LLAMA-3.2- 3B-INSTRUCT）和 86.46%（LLAMA-3.1-8B-INSTRUCT-RM- RB2）之间，大多数模型集中在 73% 至 84% 之间。相比之下，最佳样本准确率的范围较低，介于 39.50%（INTERNLM2-1.8B-REWARD）和 73.44%（DEEPSEEK-V3.2- EXP）之间，大多数模型介于 45% 至 67% 之间。所有模型的平均成对比较与 BoN 差距超过 20 个百分点，许多情况下接近 30 个百分点。

这种在 BoN 设置下显著的性能下降表明，成对比较可能会高估奖励模型在

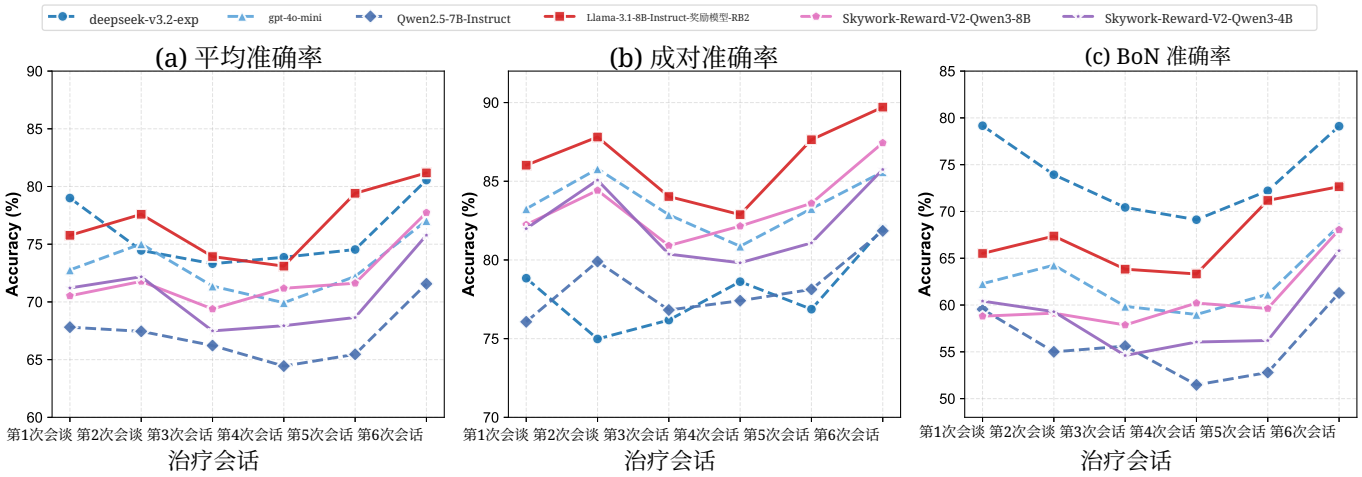

图5. 六个治疗会话中的 (a) 平均准确率（成对比较 + BoN）、(b) 成对准确率和 (c) 最佳 N 项准确率。生成式 LLM 评判模型（虚线）与判别式奖励模型（实线）通过颜色和标记样式加以区分。实线表示判别式奖励模型；虚线表示生成式 LLM 评判模型。

真实治疗情境中的对齐能力。虽然模型通常能够在直接比较中可靠地区分正向的参考回答与单个负向变体，但当需要从四个相互竞争的负向响应中进行排序时，它们往往无法将参考回答排在首位；这些负向响应中有许多包含源自心理治疗实践的微妙伤害。生成式 LLM 评判模型的差距尤其明显，其 BoN 得分经常徘徊在 60% 左右或以下，这表明它们在存在多个干扰项时维持偏好一致性的能力有限。

C. 各治疗会话中的表现

为考察奖励模型在认知行为疗法不同治疗阶段的表现，我们针对按六个治疗会话划分的提示，报告六个代表性模型的成对比较准确率、最佳样本准确率及其平均值。所选模型包括表现最好的三个生成式大语言模型评审模型（DEEPSEEK-V3.2-EXP、GPT-4O-MINI、QWEN2.5-7B-INSTRUCT），以及表现最好的三个判别式奖励模型（LLAMA-3.1-8B-INSTRUCT-RM-RB2、SKYWORK-REWARD-V2-QWEN3-8B、SKYWORK-REWARD-V2-QWEN3-4B）。

图 5 展示了这些模型在六次会谈中的平均准确率（成对比较 + 最佳样本准确率）。不同治疗阶段的表现有所差异，早期会谈（1–2）通常取得较高的最佳样本准确率，这可能是因为评估、建立治疗关系和议程设定等内容较为简单。中后期会谈（3–5）中，大多数模型的最佳样本表现下降更明显，反映出认知重构、行为实验、家庭作业回顾和巩固阶段的复杂性不断增加。

判别式奖励模型（实线）在所有治疗会话中始终优于生成式 LLM 评判模型（虚线）。LLAMA-3.1-8B-INSTRUCT-RM-RB2 的整体表现最佳，平均准确率从早期会谈（75.8%）提升至后期会谈（81.2%），表明其在后期阶段能够更好地处理累积的治疗语境和过程导向偏好。SKYWORK-REWARD-V2 的各变体保持了相对稳定的表现，在后期会谈中略有提升。

相比之下，生成式评判器的成对比较与 BoN 准确率差距更大（平均为 15–20 个百分点），且在中后期会谈中的准确率下降更为明显。例如，GPT-4O-MINI 的 BoN 准确率从 62.3%（第1次会谈）降至第4至第6次会谈的 58.9%–68.5%，这表明随着治疗进程推进，当存在多个负面变体时，其稳健性有限。

这些阶段性模式凸显了随着认知行为疗法从初期的问题构建转向行为巩固和复发预防，治疗决策的难度也在不断增加。结果表明，当前的奖励模型在粗略偏好判别方面仍比在长时程场景中稳健地找回治疗效果最优的回应更有效；尤其是在后期会谈中，细微且累积的伤害更可能显现。

D. 不同认知行为疗法策略的表现

为评估不同奖励模型在认知行为疗法会谈中引导大语言模型生成回应的有效性，我们比较了其在 14 种核心认知行为疗法策略上的偏好准确率。所选模型包括此前评估阶段确定的表现最佳的 3 种生成式奖励模型和 3 种判别式奖励模型。

图6展示了各模型在正确选择每种认知行为疗法策略标注实例的首选回应时的平均准确率（%）。总体而言，判别式模型的总体平均准确率（72.39%）略高于生成式模型（71.50%），但无论是在模型类型内部还是不同模型类型之间，都存在显著差异。

总体表现最佳的模型是 LLAMA-3.1-8B-INSTRUCT-RM-RB2（平均 = 76.66%），它在十四种策略中的八种策略上取得了最高准确率，尤其在家庭作业（82.01%）、征求反馈（83.22%）和行为技术（81.74%）方面表现出色。DEEPSEEK-V3.2-EXP 模型排名第二（平均 = 75.23%），并在

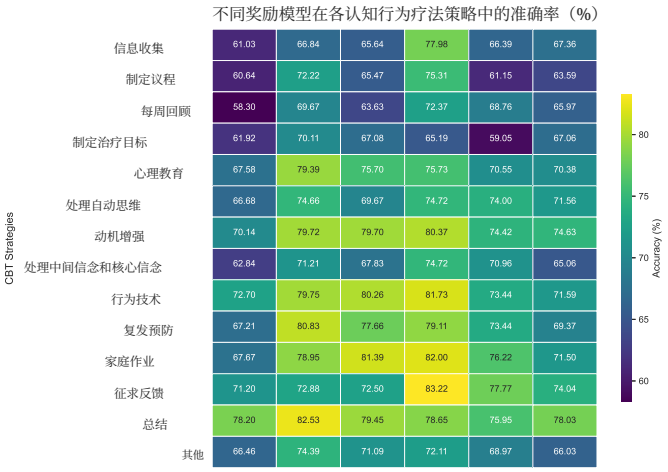

图6. 各认知行为疗法策略下，表现最佳的生成式和判别式奖励模型的准确率（%）。数值表示各策略的平均正确判断比例。

总结（82.54%）、复发预防（80.83%）和心理教育（79.39%）方面表现突出。

值得注意的是，各认知行为疗法策略的表现差异显著。总结、行为技术、动机增强和家庭作业是表现最佳的类别之一（总体平均值为 $> 76\%$），这表明在这些结构化、以行动为导向的领域中，奖励模型能够更可靠地区分高质量的回复。相比之下，制定治疗目标、制定议程和每周回顾等策略的表现相对较低且波动较大（总体平均值为 $\approx 65\text{–}66\%$），这可能反映出这些会谈初期或议程设定环节具有更强的主观性或情境依赖性。

这些发现表明，与生成式奖励方法相比，判别式奖励模型，尤其是采用基于排序的目标进行微调的模型，在认知行为疗法特定的偏好建模方面可能具有虽小但稳定的优势。然而，某些生成式模型在特定高层级策略上的出色表现，凸显了生成式奖励模型在未来治疗对话系统奖励建模中的潜在价值。

E. 细粒度负面体验上的表现

为评估奖励模型区分心理治疗来访者所报告的细微负面体验的能力，我们将 21 个元类别归并为四个主要簇，并计算每个簇中所有实例的平均成对比较准确率和 BoN 准确率。

图 7 展示了排名前三的生成式和判别式奖励模型在这四个簇中的平均成对准确率和 BoN 准确率。六个评测模型的各类别详细成对准确率和 BoN 准确率见附录 B。成对准确率在各簇之间始终保持相对较高且稳定（81–87%），其中在簇 I 上表现最佳（平均 85.8%），

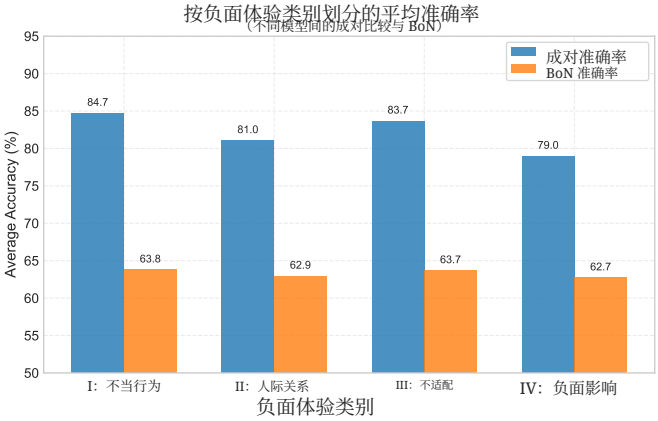

图 7. 负面体验四个主要簇的平均成对比较准确率（浅色柱）和 BoN 准确率（深色柱）。数值为六个受评奖励模型的平均值。簇 I 的表现最佳，而簇 IV 的成对比较与 BoN 准确率差距最大，凸显了检测累积性体验伤害的挑战。

而在簇 IV 上表现最低（平均 79.4%）。BoN 准确率始终较低（65–68%），其中簇 I 的下降幅度最大（平均 20.9%）。与细微的、以过程为导向或累积性的伤害相比，模型对明显且即时的治疗师错误更为敏感。例如，“治疗师评判”“治疗师贬低来访者”和“治疗师不倾听”等类别通常能取得更高的 BoN 准确率，而“治疗后问题加重”“动机或希望丧失”和“治疗期间的不愉快感受”则是最具挑战性的类别之一，即使对于表现最佳的模型，BoN 得分也常低于 65%。

这些发现表明，当前奖励模型对显著、可观察的不当行为更为敏感，而对随时间累积、且是心理治疗中来访者不满核心来源的体验性伤害，则不够敏感。各簇中持续存在的 BoN 性能下降，凸显了奖励建模方法需要明确纳入长时程推理，并能敏锐识别负面治疗体验的完整谱系。

V. 下游评估

任何奖励模型基准测试的一项核心要求，是能够预测下游对齐表现，而不只是反映孤立的偏好判断。本节评估 PRMB 得出的奖励模型排名，是否与 Best-of-N （ BoN ）推理下的现实效果相关；BoN 推理是一种广泛使用的测试时策略。

A. 实验设置

我们使用一组留出的咨询提示来评估下游性能。对于每个提示，我们从一个温度为 0.8 的策略模型中采样 16 条候选回复。随后，每个奖励模型都会为这些候选回复打分并排序，并选取得分最高的回复作为最终输出。为了自动评估生成质量，我们将选出的回复与一条认知行为疗法咨询师回复进行比较

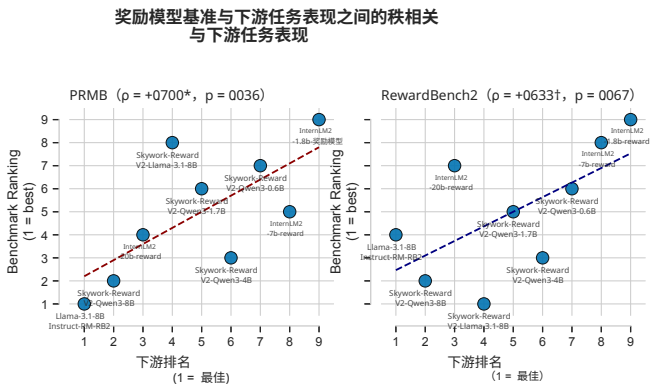

图8. 基准排名（PRMB 和 RewardBench2）与下游性能之间的秩相关（根据四个策略模型生成的 Best-of-N 回复的平均 BERTScore 排名； 1 = best）。PRMB： $\rho = 0.700$ ($p \approx 0.036$)；RewardBench2： $\rho = 0.633$ ($p \approx 0.067$)。

，并使用 BERTScore [39] 进行评估。虽然 BERTScore 是一种表层指标，无法全面衡量治疗适切性，但它能为相对质量差异提供稳定且可复现的近似指标，足以用于评估不同奖励模型之间排名的一致性。随后，我们计算由 PRMB 得出的奖励模型排名与基于 BERTScore 的下游排名之间的 Spearman 秩相关系数。相关性通过汇总四个策略模型[3]的下游 BoN 性能计算得出。这项评估关注相对排名的一致性，而非绝对生成质量。

B. 结果

图8展示了两种自动化奖励模型基准得分与四种不同策略模型的下游 BoN 推理采样性能之间的秩相关。下游排名根据四种策略模型所选 BoN 回复的平均 BERTScore 得出，其中 1 表示平均 BERTScore 最高。PRMB 与下游性能呈现强正向单调关系（Spearman $\rho = 0.700$, $p \approx 0.036$），表明 PRMB 排名靠前的奖励模型始终能选出与参考咨询师回复在词汇对齐方面表现更佳（以 BERTScore 衡量）的回复。RewardBench2 则呈现中等偏强的相关性（$\rho = 0.633$, $p \approx 0.067$），方向为正，但仅达到边缘显著水平。这些发现表明，在这一长时程治疗领域任务中，与覆盖范围更广、面向通用用途的 RewardBench2 相比，PRMB 能为推理时对齐效果提供更可靠的预测信号。

VI. 讨论

A. 推理时策略的影响

表 V 展示了若干常见推理时策略应用于不同骨干架构的生成式奖励模型时的效果。结果揭示了显著的

[3]Qwen2.5-7B-Instruct、LLaMA-3.2-3B-Instruct、Mistral-7B-Instruct-v0.3 和 Gemma3-4B-IT。

表 V 不同生成式奖励模型骨干网络上推理时策略对成对比较和 Best-of-N（BoN）准确率的影响。

方法	成对准确率	N选优准确率
gpt-4o-mini	0.8368	0.6266
+少样本（2样本）	0.8271	0.6000
+ 检索增强生成	**0.8582**	**0.6584**
+ 自我改进	0.8472	0.6288
+ 思维链	0.8276	0.5783
Qwen2.5-7B-Instruct	0.7860	0.5584
+少样本（2样本）	0.7706	0.5379
+ 检索增强生成	**0.8044**	**0.6029**
+ 自我改进	0.7700	0.5479
+ 思维链	0.7884	0.5054
LLaMA-3.2-3B-Instruct	0.6513	0.4001
+少样本（2样本）	0.6220	0.3865
+ 检索增强生成	**0.6950**	**0.4813**
+ 自我改进	0.4800	0.2071
+ 思维链	0.5025	0.2535

不同模型系列与评估协议（成对比较与 Best-of-N）之间的性能差异，且除检索增强生成（RAG）外，没有任何策略能带来一致的改进。

少样本提示（2样本）在全部三个骨干模型以及两种评估协议上均持续降低性能。成对准确率在 gpt-4o-mini 上下降 0.97%，在 Qwen2.5-7B-Instruct 上下降 1.54%，在 LLaMA-3.2-3B-Instruct 上下降 2.93%，相应的 BoN 降幅分别为 2.66%、2.05% 和 1.36%。这种一致的性能下降表明，提供少量上下文偏好示例会引入噪声，或无法与长时程治疗对话中复杂的偏好分布对齐，且与模型规模或架构无关。

Self-Refine 与思维链（CoT）提示表现出最不利的影响，尤其是在更具挑战性的 Best-of-N 设置中。在 gpt-4o-mini 和 Qwen2.5-7B-Instruct 上，两种方法均使成对准确率出现小幅至中等幅度下降，并使最佳样本准确率出现更明显的下降。对 LLaMA-3.2-3B-Instruct 而言，性能下降尤为严重：Self-Refine 将成对准确率降至 48.00%（−17.13%），将 BoN 降至 20.71%（−19.30%）；而思维链将成对准确率降至 50.25%（−14.88%），将 BoN 降至 25.35%（−14.66%）。这些急剧下滑表明，迭代式自我修正和冗长的逐步推理可能会放大初始错误或强化表层启发式，而无法解决多轮治疗互动中典型的、细微的轨迹层面偏好错位。

检索增强生成（RAG）是唯一一种在所有三种骨干模型和两种评估协议中都能持续提升性能的策略。在 LLAMA-3.2-3B-INSTRUCT 上，它带来的相对增益最大（成对比较 +4.37%，BoN +8.12%），其次是 QWEN2.5-7B-INSTRUCT（成对比较 +1.84%，BoN +4.45%）和 GPT- 4O-MINI（成对比较 +2.14%，BoN +3.18%）。这些持续的正向效果表明，注入外部认知行为疗法相关知识能够有效弥补知识缺口，并提高与治疗效果最优回应的对齐程度。

RAG 带来的可靠提升凸显了一个有前景的方向：在推理过程中纳入高质量的外部参考反馈信号，或许能帮助生成式奖励模型更好地捕捉咨询对话中的过程导向偏好。与往往会引入噪声或放大弱点的启发式方法（少样本、Self-Refine、思维链）不同，RAG 能提供与任务要求更为契合的领域专属指导。尽管绝对增益仍然有限，尚未完全缩小开源模型与闭源模型之间的性能差距，但这些结果促使我们进一步探索参考增强方法，例如偏好对检索和会话上下文感知检索增强生成。

总体而言，实验表明，流行的训练后推理时策略无法为长时程、过程导向的咨询任务中的生成式奖励建模带来稳健或可靠的收益。在所测试的方法中，只有检索增强生成能够持续带来正向（尽管有限）的增益，而少样本提示、Self-Refine 和思维链则经常降低性能，尤其是在 Best-of-N 设置下以及资源受限的模型上。这些发现凸显了基于启发式的推理增强的局限性，并强调需要训练时干预、更精细的参考机制，以及能够更好捕捉会谈级偏好动态与长期治疗连贯性的评估基准。

VII. 结论

本文提出 PRMB，这是一个用于评估长时程、多次会谈的认知行为疗法咨询中奖励模型的基准。大量评估表明，判别式与生成式奖励模型在该场景下都难以可靠地捕捉与咨询师一致的偏好。我们还发现，仅靠推理时启发式方法不足以在过程导向的对话领域中对齐奖励模型。我们希望 PRMB 能够为未来明确考虑长时程结构、相对判断和治疗轨迹的奖励建模方法研究奠定基础。

附录 A 数据构建补充材料

1）渐进式摘要生成：为了在保持提示长度可控的同时支持长程认知行为疗法咨询评估，我们采用渐进式摘要生

.....框架来构建咨询提示。历史信息会被逐步压缩为结构化摘要，并在各次治疗会话中持续传递。具体而言，对于由多次会谈组成的每个咨询案例，我们会维护两类摘要：

- **短期摘要**，用于记录当前会谈截至当前轮次的关键进展，例如近期的情绪状态和新表达的想法，如图 9 所示。

- **长期摘要**，用于记录有关来访者的稳定信息和跨会谈信息，包括主诉问题、核心信念、反复出现的自动思维、此前识别出的认知扭曲，以及已引入的治疗策略，如图 10 所示。

短期记忆提示（第一阶段）

```
你是一名专业的认知行为疗法（认知行为疗法）治疗师。请根据以下对话撰写一份
**150-250词**的摘要。

要求：- 识别关键事件和主题 - 来访者的核心情绪、担忧和困扰 -
可能的自动思维或核心信念 - 治疗师与来访者互动的重点及初步目
标


对话内容：
{dialogue_text}

请输出：一份清晰、结构化的 150–250 字摘要。
```

短期记忆提示（第二阶段）

```
你是一位专业的认知行为治疗师（CBT）咨询师。

以下是你对对话的上一版总结（请不要丢弃
任何有用的信息）：
---------------
{previous_summary}
---------------

现在，你将获得一段新增的对话内容。请：
1. 阅读新增内容，并提取与
咨询进展相关的任何新信息（包括情绪、自动思维、
人际事件、话题变化、新出现的问题等）。
2. **补充、修订并完善**先前的摘要。
3. 输出一份**连贯一致的更新摘要，长度为 150–250
字。**

新的对话内容如下：
---------------
{dialogue_text}
---------------

请输出一份全面的摘要（150–250 字），整合此前对话的所有内容，并纳入本轮
的新信息，而不只是总结新内容。
```

图 9. 短期记忆生成中的提示用法。

在第一次会谈开始时，这两类摘要均初始化为空。每次会谈结束后，系统会使用结构化总结提示，将短期摘要合并到长期摘要中，并重置短期摘要，以供下一次会谈使用。此过程确保治疗轨迹中的关键信息能够在多次会谈之间得到保留，同时不会超出上下文长度。重要的是，摘要仅根据此前的对话内容生成，不会纳入或引用用于偏好评估的候选咨询师回复。这可防止信息泄漏，并确保奖励模型的评估仅基于真实长时程咨询场景中可获取的信息。

2）候选回复生成中模型列表的使用：用于候选回复生成的模型包括开放权重模型和闭源系统，以及经过指令微调和聊天优化的变体。所有模型均采用零样本生成设置，并使用固定的解码参数，以确保不同候选回复之间具有可比性。所有模型生成回复时的温度均设为 0。8，最大输出长度则足以完成一个完整的咨询轮次。未对任何模型进行特定的提示调优或后处理。每个提示都与多个由不同模型独立生成的回复配对，从而构建具有挑战性的偏好对和 Best-of-N 实例，以反映真实的模型 di-

长期记忆提示

您是一位专业的认知行为治疗师（CBT）咨询师。以下是本次会谈的若干分段摘要，按时间顺序排列。请综合这些摘要，为本次咨询生成一份结构化的会谈级摘要，并确保格式与示例完全一致：

会谈摘要重点：- 主要讨论主题及引发情绪的情境 - 本次会谈讨论的关键问题（例如认知重构、情绪调节、回避行为、行为实验等）进展：- 来访者在情绪识别、认知重构和行为调整方面的具体进展 - 新的领悟、已完成的练习或有所提升的技能挫折／挑战：- 本次会谈中尚未解决的困难 - 反复出现的自动思维、行为模式及阻碍因素 - 技能应用中的瓶颈或现实情境中的挑战下一步：- 建议的下一阶段认知行为疗法干预方向（认知重评、行为实验、暴露疗法、情绪监测、家庭作业等） - 具体且可执行的后续行动建议要求：1. 语言应客观、中立、专业，表达清晰具体。2. 每个部分列出 2–4 项，并使用项目符号（-）。3. 所有内容必须依据输入摘要，不得凭空编造。4. 不要输出解释、提示或其他额外文本。输入内容如下：{summaries_text} 请根据以上要求生成会谈层面的摘要。

图10. 长期记忆生成中的提示使用 n.

多样性。用于候选回复生成的模型完整列表见表VI。

表 VI 用于咨询师回复候选生成的模型。

模型名称	Type
GLM-4.6	闭源
Qwen3-235B-A22B-Thinking-2507	开放权重
Qwen3-30B-A3B-Thinking-2507	开放权重
Qwen3-32B	开放权重
DeepSeek-V3.2-Exp	开放权重
DeepSeek-R1-0528	开放权重
gpt-4o-mini	闭源
gpt-oss-20b	开放权重
MiniMax-M2	开放权重
Kimi-K2	开放权重

附录 B
负面体验的详细表现

本附录提供了六个已评估奖励模型在21个负面体验元类别上的详细成对比较准确率与最佳样本准确率。各类别按其原始四簇分类体系分组，以供参考。图11提供了各模型与各类别最佳样本准确率的热力图可视化，颜色越深表示性能越高。

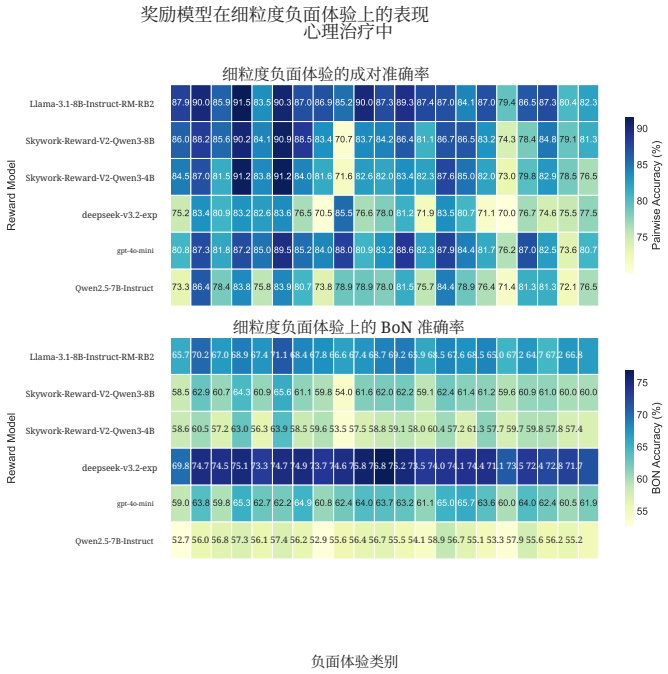

图11. 六个已评估模型在各类别上的详细成对比较准确率与最佳样本准确率。

参考文献

[1] H. Qiu、H. He、S. Zhang、A. Li 和 Z. Lan，“SMILE：通过 ChatGPT 将单轮包容性语言扩展为多轮对话，以提供心理健康支持”，载于计算语言学协会自然语言处理实证方法会议论文集：*EMNLP 2024*。计算语言学协会，2024 年 11 月，第 615–636 页。[2] Y. Chen、X. Xing、J. Lin、H. Zheng、Z. Wang、Q. Liu 和 X. Xu，“SoulChat：通过使用多轮共情对话进行微调，提升大语言模型的共情、倾听和安慰能力”，载于计算语言学协会自然语言处理实证方法会议论文集：*EMNLP 2023*。计算语言学协会，2023 年 12 月，第 1170–1183 页。[3] Y. He、Y. Pan、W. Li、J. You、J. Deng 和 F. Ren，“ECC：用于共情回应的情绪—原因对话数据集”，载于*2025* 年自然语言处理实证方法会议论文集。计算语言学协会，2025 年 11 月，第 6011–6028 页。[4] A. Sharma、K. Rushton、I. W. Lin、T. Nguyen 和 T. Althoff，“通过人类—语言模型交互促进自我引导式心理健康干预：认知重构案例研究”，载于*2024* 年 *CHI* 人机交互系统中的人为因素会议论文集，丛书：CHI ’24。计算机协会，2024 年。[5] S. Lee、S. Kim、M. Kim、D. Kang、D. Yang、H. Kim、M. Kang、D. Jung、M. H. Kim、S. Lee、K.-M. Chung、Y. Yu、D. Lee 和 J. Yeo，“Cactus：运用认知行为理论开展心理咨询对话”，载于计算语言学协会自然语言处理实证方法会议论文集：*EMNLP 2024*。计算语言学协会，2024 年 11 月，第 14 245–14 274 页。[6] H. Xie、Y. Chen、X. Xing、J. Lin 和 X. Xu，“PsyDT：利用大语言模型构建具有个性化咨询风格的心理咨询师数字孪生，用于心理咨询”，载于计算语言学协会第 *63* 届年会论文集（第 *1* 卷：长篇论文）。计算语言学协会，2025 年 7 月，第 1081–1115 页。[7] C. Zhang、R. Li、M. Tan、M. Yang、J. Zhu、D. Yang、J. Zhao、G. Ye、C. Li 和 X. Hu，“CPsyCoun：面向中文心理咨询的基于报告的多轮对话重构与评估框架”，载于计算语言学协会 *ACL 2024* 论文集。计算语言学协会，2024 年 8 月，第 13 947–13 966 页。[8] B. Zhang、R. Ma、Q. Jiang、P. Wang、J. Chen、Z. Xie、X. Chen、Y. Wang、F. Ye、J. Li、Y. Yang、Z. Tu 和 X. Li，“将 Sentient agent 作为评判者：评估大语言模型中的高阶社会认知

ArXiv，第 abs/2505.02847 卷，2025 年。 [在线]。可访问：https://api.semanticscholar.org/CorpusID:278338890
模型，
[9] P. Wang、R. Ma、B. Zhang、X. Chen、Z. He、K. Luo、Q. Lv、Q. Jiang、Z. Xie、S. Wang、Y. Li、F. Ye、J. Li、Y. Yang、Z. Tu 和 X. Li，《Rlver：面向共情代理的可验证情绪奖励强化学习》，*ArXiv*，第 abs/2507.03112 卷，2025 年。[10] N. Lambert、V. Pyatkin、J. Morrison、L. Miranda、B. Y. Lin、K. Chandu、N. Dziri、S. Kumar、T. Zick、Y. Choi、N. A. Smith 和 H. Hajishirzi，《RewardBench：语言建模奖励模型评估》，载于计算语言学协会：*NAACL 2025* 论文集。计算语言学协会，2025 年 4 月，第 1755–1797 页。[11] E. Zhou、G. Zheng、B. Wang、Z. Xi、S. Dou、R. Bao、W. Shen、L. Xiong、J. Fan、Y. Mou、R. Zheng、T. Gui、Q. Zhang 和 X. Huang，《RMB：全面基准测试大语言模型对齐中的奖励模型》，*ArXiv*，第 abs/2410.09893 卷，2024 年。 [在线]。可访问：https://api.semanticscholar.org/CorpusID:273346208
[12] S. Malik、V. Pyatkin、S. Land、J. D. Morrison、N. A. Smith、H. Hajishirzi 和 N. Lambert，《Rewardbench 2：推进奖励模型评估》，*ArXiv*，第 abs/2506.01937 卷，2025 年。 [在线]。可访问：https://api.semanticscholar.org/CorpusID:279419102
[13] J. M. Liu、D. Li、H. Cao、T. Ren、Z. Liao 和 J. Wu，《Chatcounselor：用于心理健康支持的大语言模型》，*arXiv* 预印本 *arXiv:2309.15461*，2023 年。[14] Y. Chen、X. Xing、J. Lin、H. Zheng、Z. Wang、Q. Liu 和 X. Xu，《Soulchat：通过使用多轮共情对话进行微调，提升大语言模型的共情、倾听和安慰能力》，载于计算语言学协会：*EMNLP 2023* 论文集，2023 年，第 1170–1183 页。[15] E. Jo、D. A. Epstein、H. Jung 和 Y.-H. Kim，《理解部署利用大语言模型的对话式人工智能进行公共卫生干预的益处与挑战》，载于*2023* 年 *CHI* 人机交互系统中的人为因素会议论文集，2023 年，第 1–16 页。[16] J. Wei、S. Kim、H. Jung 和 Y.-H. Kim，《利用大语言模型驱动聊天机器人收集用户自我报告数据》，*ACM* 人机交互论文集，第 8 卷，CSCW1 期，第 1–35 页，2024 年。[17] S.-L. Hsu、R. S. Shah、P. Senthil、Z. Ashktorab、C. Dugan、W. Geyer 和 D. Yang，《帮助助人者：通过 AI 赋能的实践与反馈为同伴咨询师提供支持》，*arXiv* 预印本 *arXiv:2305 08982*，2023 年。。[18] S. Chen、M. Wu、K. Q. Zhu、K. Lan、Z. Zhang 和 L. Cui，《面向精神科医生和患者模拟的大语言模型赋能聊天机器人：应用与评估》，*arXiv* 预印本 *arXiv:2305.13614*，2023 年。[19] H. Na，《CBT-LLM：一种用于基于认知行为疗法的心理健康问答的中文大语言模型》，载于*2024* 年计算语言学、语言资源与评估联合国际会议论文集（*LREC-COLING 2024*），2024 年，第 2930–2940 页。[20] N. Zhang、R. Sun、R. Su、S. Ma、S. Zhang、X. Weng、X. Zhang、Y. Zhan、Y. Xu、Z. Chen 等，《Echo-n1：情感强化学习前沿》，*arXiv* 预印本 *arXiv:2512.00344*，2025 年。[21] L. Ouyang、J. Wu、X. Jiang、D. Almeida、C. L. Wainwright、P. Mishkin、C. Zhang、S. Agarwal、K. Slama、A. Ray、J. Schulman、J. Hilton、F. Kelton、L. Miller、M. Simens、A. Askell、P. Welinder、P. Christiano、J. Leike 和 R. Lowe，《通过人类反馈训练语言模型遵循指令》，载于第*36*届神经信息处理系统国际会议论文集，丛书：NIPS '22。Curran Associates Inc.，2022 年。[22] A. Askell、Y. Bai、A. Chen、D. Drain、D. Ganguli、T. J. Henighan、A. Jones、N. Joseph、B. Mann、N. Dassarma、N. Elhage、Z. Hatfield-Dodds、D. Hernandez、J. Kernion、K. Ndousse、C. Olsson、D. Amodei、T. B. Brown、J. Clark、S. McCandlish、C. Olah 和 J. Kaplan，《作为对齐实验室的通用语言助手》，*ArXiv*，第 abs/2112.00861 卷，2021 年。[23] Y. Bai、A. Jones、K. Ndousse、A. Askell、A. Chen、N. DasSarma、D. Drain、S. Fort、D. Ganguli、T. J. Henighan、N. Joseph、S. Kadavath、J. Kernion、T. Conerly、S. El-Showk、N. Elhage、Z. Hatfield-Dodds、D. Hernandez、T. Hume、S. Johnston、S. Kravec、L. Lovitt、N. Nanda、C. Olsson、D. Amodei、T. B. Brown、J. Clark、S. McCandlish、C. Olah、B. Mann 和 J. Kaplan，《通过基于人类反馈的强化学习训练有帮助且无害的助手》，*ArXiv*，第 abs/2204.05862 卷，2022 年。[24] L. Zhang、A. Hosseini、H. Bansal、M. Kazemi、A. Kumar 和 R. Agarwal，《生成式验证器：将奖励建模作为下一个词元预测》，ArXiv，第 abs/2408.15240 卷，2024 年。 [在线]。获取地址：https://api.semanticscholar.org/CorpusID:271963324
[25] P. F. Christiano、J. Leike、T. B. Brown、M. Martic、S. Legg 和 D. Amodei，“从人类偏好中进行深度强化学习”，载于《第 31 届神经信息处理系统国际会议论文集》，NIPS'17 系列。Curran Associates Inc.，2017 年，第 4302–4310 页。[26] L. Zhu、X. Wang 和 X. Wang，“JudgeLM：微调后的大语言模型是可扩展的评审者”，ArXiv，第 abs/2310.17631 卷，2023 年。[27] L. Zheng、W.-L. Chiang、Y. Sheng、S. Zhuang、Z. Wu、Y. Zhuang、Z. Lin、Z. Li、D. Li、E. Xing、H. Zhang、J. E. Gonzalez 和 I. Stoica，“用 MT-Bench 和聊天机器人竞技场评判大语言模型评审”，载于《神经信息处理系统进展》，A. Oh、T. Naumann、A. Globerson、K. Saenko、M. Hardt 和 S. Levine 编，第 36 卷。Curran Associates, Inc.，2023 年，第 46 595–46 623 页。[28] S. Kim、J. Suk、S. Longpre、B. Y. Lin、J. Shin、S. Welleck、G. Neubig、M. Lee、K. Lee 和 M. Seo，“Prometheus 2：专用于评估其他语言模型的开源语言模型”，载于《2024 年自然语言处理实证方法会议论文集》。计算语言学协会，2024 年 11 月，第 4334–4353 页。[29] W. Yuan、R. Y. Pang、K. Cho、X. Li、S. Sukhbaatar、J. Xu 和 J. Weston，“自奖励语言模型”，载于《第 41 届机器学习国际会议论文集》，ICML'24 系列。JMLR.org，2024 年。[30] Z. Sun、Y. Shen、H. Zhang、Q. Zhou、Z. Chen、D. D. Cox、Y. Yang 和 C. Gan，“SALMON：借助可指令奖励模型实现自对齐”，载于《学习表征国际会议》，2023 年。 [在线]。获取地址：https://api.semanticscholar.org/CorpusID:263831033
[31] S. Gureja、L. J. V. Miranda、S. B. Islam、R. Maheshwary、D. Sharma、G. T. Winata、N. Lambert、S. Ruder、S. Hooker 和 M. Fadaee，“M-RewardBench：评估多语言场景下的奖励模型”，载于《计算语言学协会第 63 届年会论文集（第一卷：长篇论文）》。计算语言学协会，2025 年 7 月，第 43–58 页。[32] S. Kim、D. Kang、T. Kwon、H. Chae、J. Won、D. Lee 和 J. Yeo，“评估奖励模型在数学推理方面的鲁棒性”，ArXiv，第 abs/2410.01729 卷，2024 年。 [在线]。获取地址：https://api.semanticscholar.org/CorpusID:273025999
[33] X. H. Lu、A. Kazemnejad、N. Meade、A. Patel、D. Shin、A. Zambrano、K. Sta´nczak、P. Shaw、C. Pal 和 S. Reddy，“AgentRewardBench：评估网页智能体轨迹的自动评估”，ArXiv，第 abs/2504.08942 卷，2025 年。 [在线]。获取地址：https://api.semanticscholar.org/CorpusID:277780411
[34] H. Ding、Q. Feng、D. Liu、Q. Zhao、T. Yao、S. Wang、D. Chen、J. Li、Z. Gan、J. Zhang、C. Wang 和 Y. Wang，“RoleRMBench 与 RoleRM：面向对话系统中基于人设的角色扮演的奖励建模”，2025 年。 [在线]。获取地址：https://api.semanticscholar.org/ CorpusID:283737256
[35] Y. Zhou、N. Zhou、Q. Chen、J. Zhou、A. Zhou 和 L. He，“DiaCBT：由认知概念化图指导的基于认知行为疗法的心理咨询长周期对话语料库”，ArXiv，第 abs/2509.02999 卷，2025 年。 [在线]。获取地址：https://api.semanticscholar.org/CorpusID:281091992
[36] Y. Lv、H. Pan、Z. Wang、J. Liang、Y. Liu、R. Fu、M. Liu、Z. Wang 和 B. Qin，“CogGPT：释放认知动力学在大语言模型上的潜力”，载于《计算语言学协会论文集：EMNLP 2024 研究成果》。计算语言学协会，2024 年 11 月，第 6074–6091 页。[37] Y. Y. Chiu、A. Sharma、I. W. Lin 和 T. Althoff，“大语言模型治疗师行为评估的计算框架”，ArXiv，第 abs/2401.00820 卷，2024 年。 [在线]。获取地址：https://api.semanticscholar.org/CorpusID:266693148
[38] Z. Vybıral、B. M. Ogles、T. ˇRih´aˇcek、B. Urbancov´a 和 V. Gociekov´a，“从来访者视角看心理治疗中的负面体验：定性元分析”。Psychotherapy Research，2024 年。[39] T. Zhang*、V. Kishore*、F. Wu*、K. Q. Weinberger 和 Y. Artzi，“BERTScore：使用 BERT 评估文本生成”，载于《学习表征国际会议》，2020 年。 [在线]。获取地址：https://openreview.net/forum?id=SkeHuCVFDr