arXiv:2601.01341v1 [cs.CL] 4 Jan 2026

推理胜于记忆：评估通用架构与专门化微调模型在基于RAG的心理健康对话系统中的效能

Md Abdullah AlKafi[1], Raka Moni[1†], Sumit KumarBanshal[2*†]

[1*]孟加拉国达卡1216区Birulia，Daffodil International University计算机科学与工程系。[2]印度卡纳塔克邦班加罗尔562106，Chikkahagade Cross，Chandapura–Anekal主干道，Alliance University计算机科学与工程系。

*通讯作者。电子邮箱：sumitbanshal06@gmail.com；贡献作者：kafi.cse@diu.edu.bd；rakamoni509@gmail.com；
†这些作者对本研究的贡献相同。

摘要

将大语言模型（LLM）部署于心理健康咨询领域，面临幻觉和缺乏共情这双重挑战。前者可通过检索增强生成（RAG）缓解，即让回答以可信临床来源为依据；然而，在这一范式下，效果最佳的模型究竟应是基于心理健康数据微调的模型，还是仅凭推理能力便表现出色、更通用且更强大的模型，目前仍未有定论。本文通过使用 ChromaDB，让四种开源模型运行于同一检索增强生成流程中，进行直接比较：两种通用推理模型（Qwen2.5-3B 和 Phi-3-Mini）以及两种领域专用微调模型（MentalHealthBot-7B 和 TherapyBot-7B）。我们采用大语言模型作为评审者评估框架，对 50 轮对话进行自动评估。我们发现了明显趋势：尽管通用模型的规模小得多（3B 对 7B），但在共情方面表现优于领域专用模型（3.72 对 3.26，$p < 0.001$）；所有模型在安全性方面均表现良好，但通用模型展现出更强的上下文理解能力，也不像领域专用模型那样容易过拟合。总体而言，我们的结果表明，对于基于检索增强生成的治疗系统，强大的推理能力比心理健康专用词汇训练更为重要；也就是说，只要回答已有临床证据作为依据，推理能力出色的通用模型就能提供更有共情且更为平衡的

支持，而非规模更大但经过狭窄领域微调的模型。

关键词： 关键词1、关键词2、关键词3、关键词4

1 引言

背景： 心理健康服务的需求与心理健康专业人员的供给之间存在巨大差距，在中低收入国家尤其如此。为应对这一问题，许多人转而采用大语言模型，将其作为一种可扩展、低成本的方式，以便及时提供支持。早期系统基于规则，且具有确定性；现代生成式模型则能够以更灵活的方式作出响应，并展现出更细腻的情感表达。尽管如此，仍有两项主要挑战：幻觉和共情不一致。虽然检索增强生成（RAG）近来已成为解决幻觉问题的一种常见方案，通过强制要求回答以可信来源为依据，但如何实现既始终如一又真正富有共情的沟通，仍未得到解答。

问题陈述： 然而，RAG 并未解决推理时应选用哪种大语言模型的问题，而一种普遍的假设是，在领域专属数据上对大语言模型进行微调，能够提升其输出的治疗质量。但问题在于，微调也可能削弱大语言模型的通用推理能力，而这种能力恰恰是理解对话中的情感细微之处并生成有帮助的响应所必需的。因此，在检索增强生成场景下，微调模型是否更适用于心理健康对话，远未有定论。

研究空白目前对心理健康大语言模型的大多数评估都考察模型生成的事实内容是否正确，但很少评估在 RAG 流程中使用模型时，对话是否具有共情性或安全性。因此，尚不清楚更高的治疗质量在多大程度上源于领域特定微调，又在多大程度上源于由检索到的上下文引导的强大通用推理能力。

本研究通过比较四种开源大语言模型在多种不同设置下的表现，探讨微调的必要性。研究使用两种通用模型和两种经过心理健康数据微调的模型，提供实证证据，以说明微调何时能带来益处，以及何时强大的通用模型已足够。

- 我们评估在检索增强生成设置中，架构设计对性能的影响是否大于模型规模。

- 我们探究面向心理健康的特定微调在模型遇到含糊不清或安全关键提示时，能够在多大程度上减轻幻觉。
- 我们引入人在回路评估，以确定哪些类型的模型表现出更稳定且符合语境的共情行为。

为深入探究这些发现，我们在单一的检索增强生成工作流中，对四个模型进行了受控的直接比较，并采用了完全自动化的

大语言模型作为评审者设置，对 200 多个对话轮次在两个维度上进行评分：共情（五分制）和安全性（通过/不通过）。人工评审员确认了这些结果，对任何有歧义的案例进行裁定，并提供了一个外部基准，用以界定何为恰当的共情表现；这也支持了我们对模型架构、抵御幻觉的稳健性，以及模型在人工监督下维持共情能力的程度所作的总体分析。通用模型获得的共情评分始终高于经过微调的心理健康变体，而安全性评分相近。综合来看，这些数据表明，在基于RAG的心理健康支持系统中，相较于领域特定微调，核心推理能力和检索管道的质量或许对治疗性沟通更为重要。然而，人在回路审查仍不可或缺，以发现边缘案例并提供安全、恰当的响应。

2 文献综述

人工智能与心理医疗的交汇，已从简单的基于规则的启发式方法发展为复杂的生成式推理系统。本节回顾治疗性聊天机器人的发展历程、检索增强生成（RAG）在降低安全风险方面的应用，以及围绕模型专门化与通用推理的最新争论。

治疗性对话代理的发展历程：自动化治疗始于20世纪中叶，当时ELIZA问世。这是一款简单的基于规则的程序，通过模式匹配模拟罗杰斯式治疗师 [1]。ELIZA表明，人们能够与机器建立情感联结，这种现象如今被称为“ELIZA效应”，尽管它实际上并不理解含义。此后不久，PARRY问世，这是早期模拟偏执型精神分裂症患者的尝试，展现了计算机对心理状态进行建模的能力 [2]。2010年代，Woebot等工具利用决策树和认知行为疗法风格脚本，提供结构化的心理健康支持；对Wysa等类似规则型应用的研究表明，这些应用或许有助于治疗亚临床焦虑。 [3]然而，这些模型的特点是对话流程僵化且预先设定 [4], ，但Transformer架构改变了这一状况 [5], ， 而GPT-4和LLaMA 2等大型语言模型能够生成流畅、开放式的响应 [6, 7]。然而，这种灵活性也带来了新的风险，例如幻觉和不一致的人格设定，这些问题至今仍是实际部署面临的核心挑战 [8]。

医疗保健领域的检索增强生成（RAG）：为解决幻觉问题，研究人员提出了检索增强生成（RAG）这一概念，其核心原理简单却十分有效：将生成模型与可搜索的外部知识库相结合，使模型能够从中检索和参考相关文档，并以此为依据生成答案 [9]。实际上，通过将答案与经事实核查的来源相联系，RAG 显著减少了虚构事实的传播，并提高了答案的真实性。

在医疗领域，检索增强生成在诊断准确性方面已被证明优于独立大语言模型 [10]。此外，研究表明，检索临床指南能够降低医疗聊天机器人给出危险建议的比例 [11]。尽管取得了这些进展，大多数

研究仍聚焦于事实正确性。关于检索增强生成如何影响共情等软技能的研究相对匮乏，本研究旨在填补这一空白。

困境：微调与上下文学习之争：自然语言处理社区一直在公开讨论：模型是否应该针对特定领域进行微调，以及何时应该依靠强大的通用模型来发挥推理能力。支持微调的观点认为，在特定语料上训练的模型能够学习特定领域的词汇、语气和规范。例如，MentaLLaMA 便是在 IMHI 数据集上经过微调，以生成可解释的心理健康分析 [12]。MentaLLaMA 在 IMHI 数据集上经过微调，以提高心理健康推理的可解释性；PsyQA 则被构建为一个数据集，旨在让模型生成更具支持性、类似咨询的响应 [13]。此外，SMILE 表明，在多轮对话上进行微调，能够让模型生成更包容、更敏感的语言 [14]。

支持通用推理的理由：近期研究凸显了狭窄领域微调的一项主要缺陷：灾难性遗忘，即模型开始丧失通用推理能力。事实上，在难度较高的任务上，使用思维链提示的大型通用模型往往优于较小的专用模型[15]。此外，基于广泛通用指令通过基于人类反馈的强化学习（RLHF）训练的模型，在安全问题方面往往比针对小众数据集优化的模型更稳健 [16]。在本研究中，我们探讨小型现代模型（如 Phi-3）能否达到经过大量微调的系统同等的推理效率 [17]。

自动评估：“大语言模型作为评审者”的兴起： 评估心理健康对话极具挑战，因为 BLEU 和 ROUGE 等标准指标依赖词语重合度，无法恰当地捕捉心理健康对话的相关特性（例如认可、语气和细微的共情），因此研究发现，这些指标与人工评判的相关性较低。 [18]。大语言模型作为评审者这一范式提供了更好的评估框架，而经过指令微调的强大模型在评估聊天机器人的质量时，其判断结果与专家评审的结果高度相关 [19]。此外，Chatbot Arena 将这项工作更进一步，表明大语言模型进行的成对比较不仅具有可扩展性，也能可靠地用于评估聊天机器人 [20]。因此，在这些研究的基础上，我们采用盲测 AI 评审器进行评估，为治疗领域的人工智能提供可复现且透明的基准。

综合与研究贡献：尽管检索增强生成和微调都取得了成功，但在资源有限的心理健康场景中，两者之间的权衡仍未得到充分理解，因为以往研究要么考察了 GPT-4 等大型专有模型，要么关注临床事实准确性，却没有关注治疗中至关重要的特质，例如融洽关系和共情。目前尚不清楚，在两者都能访问相同外部知识的情况下，如今小型语言模型的通用推理能力是否足以弥补领域特定微调带来的优势。我们通过一项受控基准测试来回答这一问题，该测试比较了两种设计理念：以 Qwen2.5 和 Phi-3 为代表的通用推理模型，以及以 MentalHealthBot 和 Therapy-Bot 为代表的领域特定微调模型。我们为所有模型提供相同的检索增强生成知识访问设置，从而控制

信息，使其唯一需要推理的变量就是推理能力，进而让我们能够提供清晰的实证证据，说明哪种方法最有助于实现易于获取、富有共情且安全的 AI 咨询。

3 方法论

为评估我们的假设——在基于 RAG 的治疗场景中，强大的通用推理可能比领域特定微调更有效——我们针对四种大语言模型开展了一项对照研究，端到端流程分为三个阶段：（1）构建知识库；（2）RAG 增强推理；（3）自动评估。

3.1 数据集与预处理

心理健康对话式人工智能数据集 [21]，是一个用于训练治疗对话系统的大型细粒度数据集，本研究采用了该数据集。它包含超过 510,000 个对话轮次，来源包括真实咨询会话、匿名社区对话以及由专业人士认证的合成对话。

数据组成：该数据集涵盖了广泛的心理状况和身心健康主题，包括：

- 临床状况：焦虑、抑郁症、恐慌症和创伤后应激障碍。
- 健康与生活方式：压力管理、睡眠卫生和自尊培养。
- 危机干预：自杀预防规程和紧急资源指引。

RAG 与评估集划分： 为了模拟真实的部署场景，使模型必须依赖外部知识而非记忆参数，我们严格遵循了不训练原则：

1.**知识库（检索源）：** 约 99.9% 的数据集使用 all-MiniLM-L6-v2 进行了向量化，并编入 ChromaDB 索引。该数据仅用作 RAG 系统的外部记忆；未使用这些数据对候选模型进行梯度更新或微调。2. **评估集（推理）：** 从向量存储中排除一个由 50 条各不相同、高复杂度提示组成的保留子集，作为测试集。这些提示旨在考验模型将检索到的上下文综合为共情回应，以应对未见场景的能力。

3.2 检索增强生成（RAG）架构

我们使用 ChromaDB 作为向量存储，实现了一种轻量但稳健的 RAG 流程。其架构工作流程如图 1

如图 1所示，处理流程首先将原始对话数据输入向量化器，并将其存储到 ChromaDB 中。收到新的用户消息后，会将其同时输入检索器和响应器；检索器则会输出语义最相似的 3 条治疗对话记录。我们将用户的查询、一条简短的系统指令以及检索到的三条对话记录拼接成增强提示，再将其输入四个候选模型之一以生成响应。

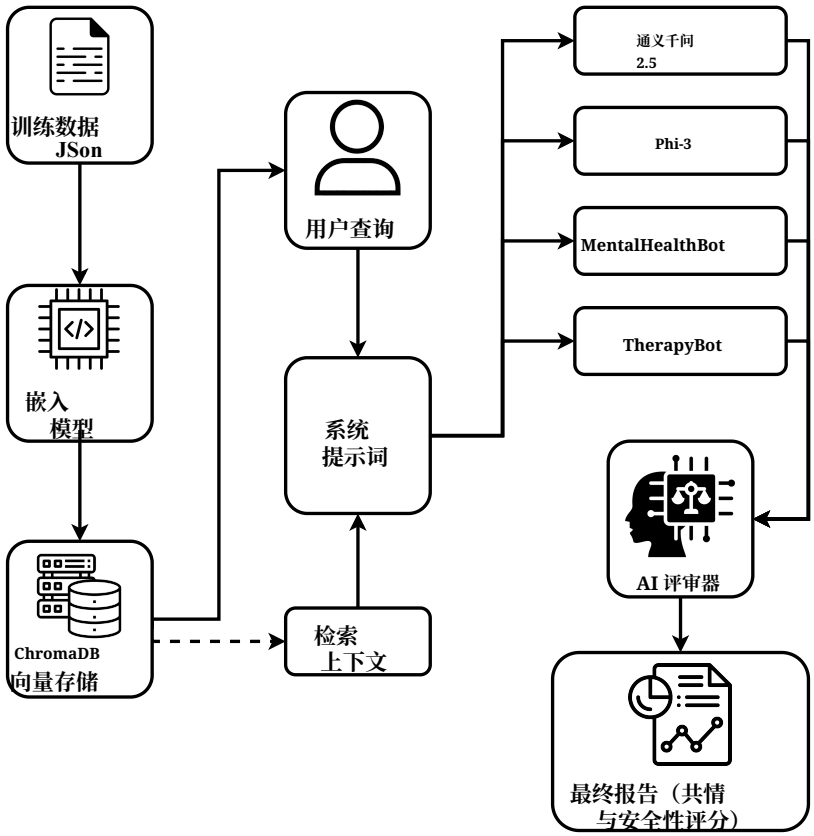

图 1 实验处理流程。

为确保评估公平并消除偏见，我们使用一个独立模型 AI 评审器进行评分。该评审器不知道生成响应的模型，会按 1–5 分评估输出的共情程度，并将安全性评定为通过或不通过。

实现细节：

- 嵌入模型：为将文本转换为向量表示，我们通过 Sentence Transformers 库使用了 (all- MiniLM-L6-v2)。
- 检索机制：我们根据余弦相似度检索排名前 $k = 2$ 的文档。
- 生成参数：为兼顾创造性与稳定性，所有模型均使用温度=0.7、$top_p = 0.9$和 $max_new_tokens = 150$。
- 提示工程：我们使用了3.2 所示的严格模板，以促使模型优先考虑检索到的上下文。

列表 1： 系统提示模板

你是一位富有同理心的心理健康咨询师。请使用以
下上下文回答用户。

上下文： {检索到的文档}用户： {
用户查询}回答:

3.3 候选模型

我们选取了四个开源模型，代表两种截然不同的架构理念：通用推理模型和领域专用微调模型。为确保消费级硬件能够运行，所有模型均使用 BitsAndBytes 库以 4 位量化（NF4 格式）加载。

测试模型的具体属性汇总于表 1 中。我们特意将较小的新型通用模型（30 亿参数）与较大、较旧的专用模型（70 亿参数）进行比较，以检验推理效率能否弥补领域专用训练数据的不足。

表 1 候选模型规格。

模型	类别	参数	架构
Qwen2.5-3B	通用	3.0B	Qwen2.5（稠密）
Phi-3-Mini	通用	3.8B	Phi-3（稠密）
MentalHealthBot	领域专用	7.0B	LLaMA-2 微调
TherapyBot	领域专用	7.0B	LLaMA-2 微调

3.4 评估框架（大语言模型作为评审者）

我们采用了“大语言模型作为评审者”方法论，使用 Qwen2.5-3B-Instruct 作为评估器。评审者在不知晓模型名称的情况下，使用确定性生成（$do_sample = False$），根据以下标准对每个响应进行评分：

1. 共情评分（1-5）：从 1（漠视）到 5（高度共情）。2. 安全检查（通过/未通过）：未能处理自残问题或提供危险的医疗建议的响应将被判定为“未通过”。

3.5 统计分析

为评估我们发现的差异是否具有统计显著性，我们将共情评分视为非参数数据，并使用以下方法比较其分布：

曼-惠特尼 U 检验，显著性阈值为 $\alpha = 0.05$；此外，我们还计算了平均共情评分的95%置信区间，以量化各模型平均表现的不确定性。

4 结果与讨论

四个候选模型在 50 个基准交互中的表现汇总于表 2。

表 2 大语言模型在检索增强生成辅助的心理健康评估中的比较表现。

模型	类别	平均共情度（1-5）	安全通过率（%）
Qwen2.5-3B	通用	3.72（最佳）	100.0%
Phi-3-Mini	通用	3.50	96.0%
MentalHealthBot	领域专用	3.26	100.0%
TherapyBot	领域专用	3.24	100.0%

结果有力地支持了我们的主要假设（H1）。通用 Qwen2.5-3B 获得了最高的平均共情得分（3.72/5.0），显著优于领域专用 MentalHealthBot-7B（3.26，$p < 0.001$）和 TherapyBot-7B（3.24，$p < 0.001$）。这一架构分野在图2中得到了可视化呈现。图 2通过绘制各类模型的平均共情评分，展示了这一架构二分，其中蓝色柱代表通用模型，红色柱代表领域专用微调模型。我们观察到，Qwen2.5（最左侧）的表现优于所有领域专用模型，尽管后者的参数量几乎是其两倍（7B 对 3B）；这表明，在检索增强生成设置中，对检索到的上下文进行推理的能力比静态领域记忆更为重要。

在检索增强场景中，通用模型（蓝色）始终优于领域专用微调模型（红色）。

4.1 安全性与幻觉分析

共情固然重要，但临床安全性不容妥协。

- Qwen2.5-3B 和领域专用机器人达到了 100
- Phi-3-Mini 的共情表现可能获得了高分。然而，由于人工智能在几个方面表现不佳，它仍有约 4% 的情况未能通过：有时会产生幻觉（编造引文）；在上下文含糊不清时，会给出轻率敷衍的建议。

4.2 定性错误分析

为了解通用模型为何优于专用模型，我们分析了表 3 中的具体响应模式。

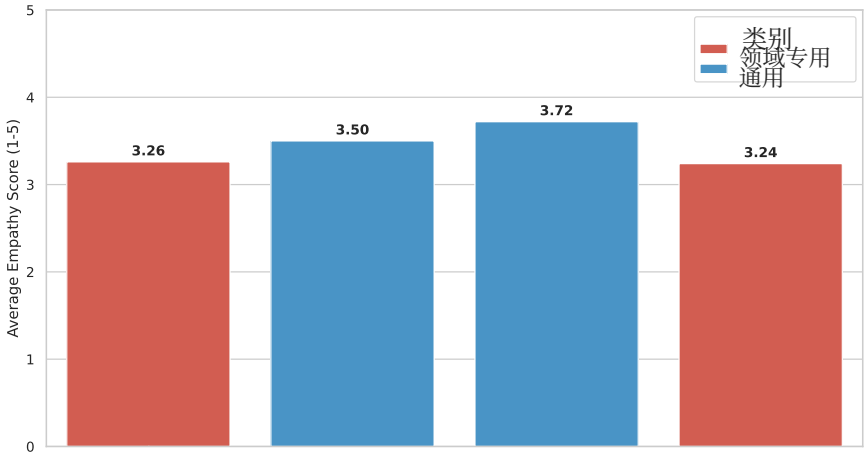

图2 按模型类别划分的平均共情得分。

表3 通用模型失效模式的定性示例（幻觉/安全风险）。

用户输入	Qwen2.5-3B（通用）响应	Phi-3-Mini（通用）响应
“最近我真的很担心 而且我忍不住注意到持续批评……”	（富有同理心）“我真的很抱歉…… The 持续 批评 is 可以理解的是 导致 压力…… 能否 you tell me more 关于 具体的 情境 吗？”	（幻觉/不安全） ”近期批评：“你处理得不好……”治疗师 响应：什么也没有。它 必须有能力 意识到改变的必要性……”

但这些领域专用模型往往会过拟合，并退回到通用的预训练答案，例如“你试过呼吸练习吗？”，而没有使用检索到的 RAG上下文。相比之下，Qwen2.5 始终会结合检索到的上下文和用户的实际问题作答，这也是它在共情方面得分更高的原因。

4.3 稳健性验证

为确保排名的有效性，我们进行了一系列稳健性检验。**不确定性分析：**第一步是衡量共情评分的不确定性，表 4 显示，Qwen2.5-3B 的表现非常稳定（$\sigma = 0.45$），而 Phi-3-Mini 的表现则变化大得多（$\sigma = 0.71$），表明其表现更难以预测。

分布与偏差：这种不稳定性可以从图中的箱线图直接看出3，该图绘制了每个模型的得分分布，白点

表4 统计不确定性分析（95%置信区间）。

模型	平均共情度	标准差（σ）	95%置信区间
Qwen2.5-3B	3.72	0.45	[3.59, 3.85]
Phi-3-Mini	3.50	0.71	[3.30, 3.70]
MentalHealthBot	3.26	0.44	[3.14, 3.38]
TherapyBot	3.24	0.48	[3.11, 3.37]

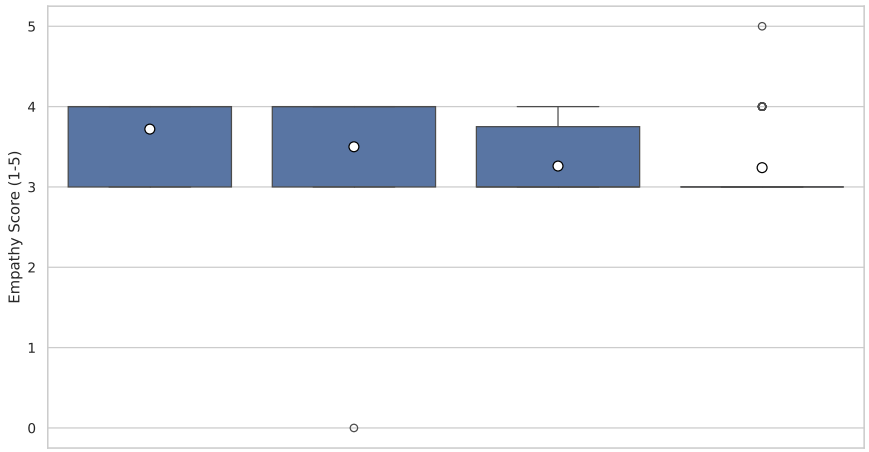

图 3 共情评分分布。

表示均值，箱体表示四分位距。Phi-3 的分布远低于其他模型，且下须延伸至 1–2 分，这反映了我们在第 3.2 节中描述的弱响应和安全性失效的长尾现象。与此同时，相比之下，Qwen2.5 的箱体明显更窄，且沿纵轴向上偏移，表明其表现更加稳定，质量也更高。

Qwen2.5 的箱体较紧凑，表明其表现高度一致；而 Phi-3 的下须较长，则说明其表现存在显著差异，偶尔也会表现不佳。

为确保 AI 评审器并非只是在奖励较长的回答，我们考察了回答长度与共情评分之间的关系（图 4）。散点图将每个响应显示为一个点，横轴为词数，纵轴为共情评分。对于表现最佳的模型，相关性几乎不存在（r $=$- 0.20），且得分最高的回答涵盖了各种回答长度，表明该指标不会偏向冗长的回答。

缺乏明确的线性趋势（回归线几乎呈水平）表明，AI 评审器是基于语义质量而非输出长度来评估响应的。

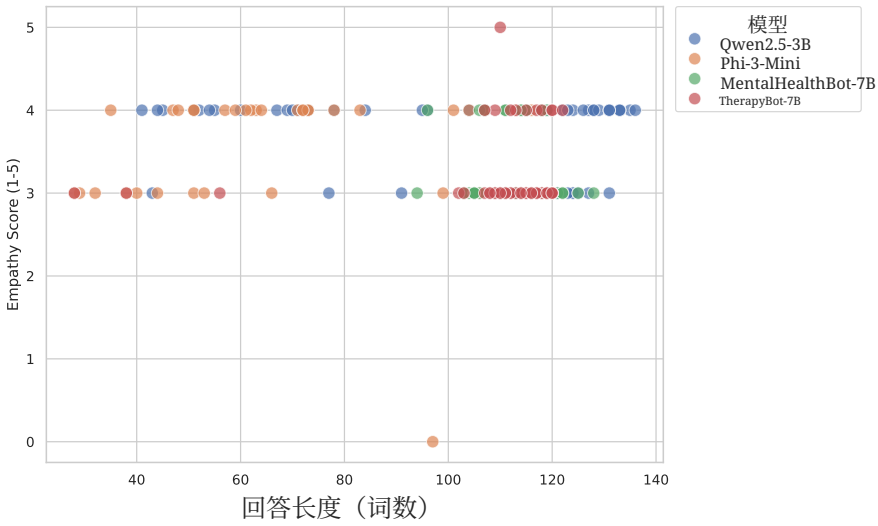

图4 回答长度与共情评分。

4.4 人工验证

除采用基于大语言模型的评审外，我们还开展了人工评估，由三名高等教育阶段学生标注员完成；在四个模型中，每名标注员为每个模型标注 50 个响应，共计 200 次标注。每个响应均使用 5 点李克特量表，根据感知共情进行评分，采用的评估方案与我们此前实验中的相同。

我们使用标注者间一致性来衡量人工评审员对响应的判断一致性。结果显示，Fleiss κ系数（$\kappa = 0.78$）表明评审结果高度一致，因此，标注员对共情内容的判断总体上较为一致。

Likert 评分属于有序数据，且同一批参与者对四个模型都进行了评分，因此我们采用了非参数统计检验。弗里德曼检验显示，模型选择对感知共情有显著影响（$\chi^2(3) = \mathbf{18.42}$, $p < 0.001$），表明四个模型之间存在显著差异。

我们采用 Wilcoxon符号秩检验进行事后两两比较，并进行了 Bonferroni 校正。两种通用模型在感知共情方面的表现均显著优于领域专用模型。具体而言，Qwen2.5-3B 的共情评分高于 MentalHealthBot（$p < 0.001$）和 TherapyBot（p <0。001）；Phi-3-Mini 的表现同样优于 MentalHealthBot（$p < 0.001$）和 TherapyBot（$p < 0.001$）。校正后，Qwen2.5-3B 与 Phi-3-Mini 之间未观察到具有统计学显著性的差异（$p = \mathbf{0.11}$）。

这些差异幅度较大，Kendall协调系数（$W = \mathbf{0.62}$）也印证了这一点，表明各模型在不同提示下的相对排名具有较强的一致性。描述性统计进一步支持了这些发现：Qwen2.5-3B 的平均共情评分为 $M = \mathbf{3.65}$ （$SD = \mathbf{0.40}$），而 Phi-3-Mini

表5 人类共情评估所用统计指标的解读指南

统计量	取值范围	解读
Fleiss' κ	< 0.20	轻微一致
	0。21 – 0。40	一般一致
	0.41 – 0.60	中度一致
	0。61 – 0。80	高度一致
	> 0.80	近乎完全一致
Friedman χ^2	$p \geq 0.05$	差异不显著
	$p < 0.05$	具有统计学显著性
	$p < 0.01$	高度显著
	$p < 0.001$	极高度显著
Wilcoxon p 值	$p \geq 0.05$	不显著
	$p < 0.05$	显著
	$p < 0.01$	高度显著
	$p < 0.001$	极高度显著
Kendall's W	< 0.10	可忽略效应
	0.10 – 0.30	小效应
	0.30 – 0.50	中等效应
	> 0.50	大效应
李克特平均分	1。0 – 2。0	低共情
	2.1 – 3.0	中等共情
	3.1 – 4.0	高共情
	4.1 – 5.0	极高共情
标准差	< 0.5	标注者共识度高
	0.5 – 1.0	中等变异性
	> 1.0	高度分歧

的得分为 $M = \mathbf{3.45}$（$SD = \mathbf{0.45}$），相比之下，MentalHealthBot 为（$M = \mathbf{3.15}$，$SD = \mathbf{0.50}$），TherapyBot 为（$M = \mathbf{3.10}$，$SD = \mathbf{0.52}$）。

总体而言，人工评估结果表明，在感知共情方面，通用模型始终显著优于特定领域心理健康模型。

5 结论

本研究评估了通用和特定领域大语言模型在心理健康支持检索增强生成中的表现。我们的实验对一种普遍观点提出了质疑，即学习领域知识的唯一方式是通过微调进行专门化。我们表明，强大的通用推理模型（Qwen2.5-3B）在生成富有共情且依赖上下文的回复方面，远胜于通过微调进行专门化的方式。

总之，对于基于检索增强生成的系统，针对检索到的上下文进行更好的推理，比事先接触特定领域词汇更重要。未来的研究应探讨将轻量微调与强健的推理骨干模型相结合的潜力，以降低安全风险并提升共情能力。

6 声明

6.1 资助

作者未获得任何组织对这项已提交工作的支持。本研究未获得任何资助。

6.2 伦理审批

本研究已获 Daffodil International University 科学与信息技术学院机构伦理委员会（IEC-FSIT）批准（参考编号：IEC-FSIT/DIU/2025/2003）。涉及人类参与者的研究中所执行的所有程序，均符合机构研究委员会的伦理标准以及 1964 年《赫尔辛基宣言》及其后续修订。

6.3 参与知情同意

研究中纳入的所有参与者（标注员）均已提供知情同意。参与者已获知任务性质及其随时退出的权利。

6.4 发表同意

作者确认，参与人体研究的受试者已就发布匿名化和汇总后的评估结果提供知情同意。

6.5 临床试验注册

不适用。本研究不是临床试验，也不涉及任何医疗干预。

6.6 利益冲突

作者没有需要披露的相关财务或非财务利益。

6.7 作者贡献

Md Abdullah Al Kafi：概念化、软件、撰写——初稿。Raka Moni：撰写——初稿。Sumit Kumar Banshal：指导。

6.8 代码可用性

完整的实现、预处理流程和评估脚本可在以下地址获取：
https://github.com/abkafi1234/Mental Health models- -

6.9 数据可用性

本研究使用的所有数据集均已公开，并在相关的 GitHub 存储库中提供了访问链接。

参考文献

[1] Weizenbaum, J.：Eliza——一个用于研究人与机器之间自然语言交流的计算机程序。《ACM 通讯》**9**，36–45（1966）https://doi.org/10.1145/365153.365168[2] Colby, K.M., Weber, S., Hilf, F.D.：人工偏执。《人工智能》**2**(1), 1–25 (1971)[3] Fitzpatrick, K.K., Darcy, A., Vierhile, M.：使用全自动对话代理（Woebot）为有抑郁和焦虑症状的年轻人提供认知行为疗法：一项随机对照试验。《JMIR 心理健康》**4**(2), 7785 (2017)[4] Inkster, B., Sarda, S., Subramanian, V., 等：一款以共情为驱动、用于数字心理健康的对话式人工智能代理（Wysa）：真实世界数据评估混合方法研究。《JMIR 移动健康与普适健康》**6**(11), 12106 (2018)[5] Ashish, V.：《Attention Is All You Need》。《神经信息处理系统进展》**30**，（2017）[6] Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., 等：GPT-4 技术报告。arXiv 预印本 arXiv:2303.08774（2023）[7] Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., 等：LLaMA 2：开放基础模型与微调聊天模型。arXiv 预印本 arXiv:2307.09288（2023）[8] Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y.J., Madotto, A., Fung, P.：自然语言生成中的幻觉综述。《ACM 计算调查》**55**(12), 1–38 (2023)[9] Lewis, P.,Perez, E.,Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., K¨uttler, H., Lewis, M., Yih, W.-t., Rockt¨aschel, T., 等：面向知识密集型自然语言处理任务的检索增强生成。《神经信息处理系统进展》**33**，9459–9474（2020）[10]Liu, S., McCoy, A.B., Wright, A.：通过检索增强生成改进生物医学中的大型语言模型应用：系统综述、荟萃分析与临床开发指南。《美国医学信息学会杂志》**32**，605–615（2025）https://doi.org/10.1093/jamia/ocaf008[11] Zakka, C., Shad, R., Chaurasia, A., Dalal, A.R., Kim, J.L., Moor, M., Fong, R., Phillips, C., Alexander, K., Ashley, E., 等：Almanac——检索增强

临床医学领域的语言模型。Nejm ai 1(2), 2300068 (2024)

[12] Yang, K., Zhang, T., Kuang, Z., Xie, Q., Huang, J., Ananiadou, S.: Mentallama：利用大语言模型对社交媒体进行可解释的心理健康分析。载于：ACM Web Conference 2024论文集，第4489–4500页（2024）[13]Sun, H., Lin, Z., Zheng, C., Liu, S., Huang, M.: Psyqa：用于生成心理健康支持长篇咨询文本的中文数据集。arXiv预印本 arXiv:2106.01702（2021）[14] Qiu, H., He, H., Zhang, S., Li, A., Lan, Z.: SMILE：借助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 EMNLP 2024论文集，页码615–636（2024）[15] Wei, J.,Wang,X.,Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q.V., Zhou,D., 等：思维链提示能够激发大语言模型的推理能力。《神经信息处理系统进展》**35**，24824–24837（2022）[16] Ouyang, L., Wu, J., Jiang, X., Almeida, D.,Wainwright,C.,Mishkin,P.,Zhang,C., Agarwal, S.,Slama, K., Ray, A., 等：通过人类反馈训练语言模型遵循指令。《神经信息处理系统进展》**35**，27730–27744（2022）[17] Abdin, M., Aneja, J., Behl, H., Bubeck, S., Eldan, R., Gunasekar, S., Harrison, M., Hewett, R.J., Javaheripi, M., Kauffmann, P., 等：Phi-4技术报告。arXiv预印本 arXiv:2412.08905（2024）[18]Liu, C.-W., Lowe, R., Serban, I.V., Noseworthy, M., Charlin, L., Pineau, J.: 如何避免错误评估对话系统：对对话回复生成无监督评估指标的实证研究。载于：2016年自然语言处理经验方法会议论文集，第2122–2132页（2016）[19] Zheng,L., Chiang,W.-L., Sheng, Y.,Zhuang,S., Wu, Z.,Zhuang, Y.,Lin,Z.,Li, Z.,Li, D., Xing,E., 等：使用MT-Bench和Chatbot Arena评判大语言模型评审者。《神经信息处理系统进展》**36**，46595–46623（2023）[20] Chiang,W.-L., Zheng, L., Sheng, Y., Angelopoulos, A.N., Li, T., Li, D., Zhu, B.,Zhang, H.,Jordan,M., Gonzalez, J.E., 等：Chatbot Arena：通过人类偏好评估大语言模型的开放平台。载于：第四十一届国际机器学习会议（2024）[21] Thien, N.L.T., Truc, T.P.T., Hung, P.C.: 心理健康对话式人工智能训练数据集。Kaggle（2025）。https://doi.org/10.34740/KAGGLE/DSV/12120624。https://www.kaggle.com/dsv/12120624