心理科学进展 2025, Vol. 33, No. 10, 1783–1793 https://doi.org/10.3724/SP.J.1042.2025.1783
Advances in Psychological Science https://cstr.cn/32111.14.2025.1783

·研究前沿(Regular Articles)·

大语言模型的共情模拟：评估、提升与挑战*

周倩伊[1#] 蔡亚琦[1#] 张 亚[1,2]

([1]华东师范大学心理与认知科学学院, 上海市心理健康与危机干预重点实验室, 上海 200062)
([2]青少年心理健康与危机智能干预安徽省哲学社会科学重点实验室, 合肥师范学院, 合肥 230601)

摘 要 随着大语言模型(Large Language Models, LLMs)在自然语言生成和情感计算方面的技术进步, 其在心理咨询、医患沟通和客户服务等领域的共情模拟能力受到广泛关注。LLMs 的共情模拟主要表现为认知共情模拟, 而非情感共情模拟, 主要包括情绪识别、共情回应和语境适应。当前评估 LLMs 共情模拟的方法包括人工评估、自动化评估、任务驱动评估, 三种方法各有优劣, 适用场景有所差异。LLMs 共情模拟能力与人类共情能力相比, 在共情生成任务方面表现出色, 但仍面临情感理解的局限性; 为了进一步提升LLMs的共情模拟能力, 可以采用数据增强、模型框架与架构优化、强化学习、引导词优化等方法加以改进。同时, 模型使用过程中的伦理规范与潜在风险仍需引起关注。

关键词 大语言模型, 共情模拟, 共情评估, 伦理问题
分类号 B842; R395

1 引言

近年来, 大语言模型(Large Language Models, LLMs)不断发展, 展现出与人类相当的智能(Wang, Ma, et al., 2024)。目前, 大语言模型已被广泛应用于语言生成相关任务, 例如问题解答、文章概括、逻辑推理等(Laskar et al., 2023; Zhuang et al., 2023), 其角色扮演的能力还让其能够模拟人类被试进行实验(Dillion et al., 2023)。为了使 LLMs 在实际应用中更好地适应人类需求, 关于其模拟人类共情(Empathy)的相关研究已然成为热点。特别是在医患沟通、心理咨询和客户服务等领域, LLMs 的共情模拟能力展现出重要的应用价值。

具体来说, 共情是指识别他人情绪, 理解其情感状态, 并做出适应性反应的心理能力。在人机交互(Liu-Thompkins et al., 2022; Svikhnushina & Pu, 2022)中, 准确理解对方讲述中所包含的想法、情感, 并给出恰当的回应极其重要。已有的研究发现共情通常包括认知共情和情感共情, 前者指理解他人情感的能力, 与心理理论密切相关; 后者指对他人情绪产生共鸣并作出情绪反应的能力, 两者可能对应不同的脑区活动(Ma et al., 2024)。近期对 LLMs 共情模拟特点的综述研究表明, LLMs 的共情模拟能力主要表现为认知共情模拟, 而非情感共情模拟。这是因为其共情性回应依赖于对训练数据中语言模式的统计学习, 而非内在的真实情感体验或动机, 与人类基于主观体验、情感共鸣和道德动机的共情, 在机制上有本质区别。从根本上看, LLMs 无法真正体验人类情绪。然而, 在识别情绪并生成共情性回应的任务中, LLMs 在某些情境下的表现甚至优于人类(Sorin et al., 2024)。尽管如此, LLMs 要有效进行共情模拟, 仍需根据具体情境准确展现恰当的情感反应, 以实现与人类的情感对齐。相较于逻辑推理、文章概括或信息提供, 共情模拟对 LLMs 而言更具挑战性, 因其作为一种复杂且多层次的反应, 不仅要求熟练掌握语言, 还需深入理解人类心理、情感及社会背景(Huang et al., 2025)。

收稿日期：2025-03-09
* 青少年心理健康与危机智能干预安徽省哲学社科重点实验室开放基金重大项目(SYS2024XXX)资助。
# 周倩伊和蔡亚琦为本文的共同第一作者
通信作者：张亚, E-mail: yzhang@psy.ecnu.edu.cn

LLMs 共情模拟能力的研究对于增强人机交互体验, 提高用户信任度和满意度具有重要意义。例如, 在医疗健康领域, LLMs 能够生成富有共情的回应, 改善患者在医疗互动中的体验, 增强患者满意度(Luo et al., 2024); 在心理咨询领域, LLMs 可作为心理健康支持工具, 提供即时的情感支持和建议, 帮助用户应对情绪困扰; 结合多模态数据, LLMs 可以更准确地识别用户情绪, 模拟心理咨询师的共情和主动引导能力, 辅助用户进行自我反思和情绪管理(Ren et al., 2024)。

2 LLMs 共情模拟能力的评估

2.1 人工评估

人工评估有两种常用的方式。第一种依靠人类标注员或目标用户对 LLMs 的共情模拟能力进行 Likert 主观评分, 例如, “请根据您在该情境下会给出的回应, 对该回答的共情水平进行评分” (Welivita & Pu, 2024)。Likert 评分范围通常为 1 (完全没有共情性) ~ 5 (极具共情性) (例如 Ayers et al., 2023; Lee, Suh, et al., 2024)或 1 (完全没有共情性) ~ 3 (极具共情性) (例如 Welivita & Pu, 2024)。评估维度通常可以包括：1)情感理解, 即评估模型是否正确识别用户的情感, 例如“我今天很悲伤”, 模型是否能表现出对悲伤情绪的正确理解。2)情感反应, 即评估生成的文本是否能展现出适当的共情性, 例如安慰、支持或提供建议。3)语境适应性, 即检查模型的回应是否符合对话上下文, 避免出现不相关或不适当的回答。4)语言自然度, 即评估模型生成文本的流畅性、用词适当性以及是否符合人类沟通习惯。此外, 还可以邀请心理咨询、医疗服务等领域的专业人士, 对模型表现进行专家评估(例如 Ayers et al., 2023)。

第二种则是由人类评估者判断在同一情境下的两条回答(其中一条由人类给出, 一条由 LLMs 生成)何者更具有共情性, 例如, 在比较 LLMs 和人类医生回复的共情性时, 请评估者阅读患者的问题和对应的两个回答, 判断何者更具共情性(Luo et al., 2024)。该方法使用的广泛性不及第一种 Likert 主观评分法, 可能是由于 Likert 评分法可以对不同模型的共情水平进行细致的量化评定(例如 1~5 分), 但是, 仅将人类回答和模型回答进行优劣比较, 提供的信息较为简单, 且在需要横向比较多个模型时难以适用。

2.2 自动化评估

自动化评估方法利用算法或现有的机器学习技术, 对 LLMs 的共情模拟能力进行快速、客观的量化分析。例如, 在进行情感相似性评估时, 可采用预训练大语言模型(如 BERT[1]、RoBERTa[2])对文本情绪进行分类, 并比较二者的一致性; 使用余弦相似度(Cosine Similarity[3])来计算用户输入与生成文本在情感向量空间的相似度(例如 Reimers & Gurevych, 2019)。自动化评估方法(如情感分类和余弦相似度计算)为 LLMs 共情模拟能力的评估提供了高效且可量化的手段。该类方法适用于大规模数据的分析, 能够帮助研究人员和开发者快速优化模型, 从而提升其在共情对话中的表现。

2.3 任务驱动评估

任务驱动评估通过让模型执行特定的情感识别和共情任务来评估其共情模拟能力。例如, 对话中的情感识别任务(Emotion Recognition in Conversations, ERC), 通过衡量 LLMs 对于对话中情感分类的准确性, 评估其共情模拟能力(Zhao et al., 2023)。在此基础上, 对话中的情感原因识别任务(Recognizing Emotion Cause in Conversations, RECCON, Poria et al., 2021)进一步拓展了情感识别的深度, 该任务关注对话中情感的来源, 包括两个子任务：因果跨度提取(Causal Span Extraction, CSE, 该任务旨在识别讲述者非中性话语的因果跨度, 即情感原因)和因果情感蕴含(Causal Emotion Entailment, CEE, 该任务给定一个讲述者的非中性话语, 目标是预测对话历史中哪些特定话语导致了讲述者话语中的非中性情绪)。RECCON 任务要求模型分析对话中的因果关系, 分析是哪些发

[1] BERT 是由 Google AI 在 2018 年提出的一种基于 Transformer 架构的自然语言处理(NLP)模型。它是一种预训练语言模型, 可以通过无监督学习在大规模文本数据上进行训练, 然后在特定任务(如问答、情感分析等)上进行微调。

[2] RoBERTa 是由 Facebook AI 在 2019 年提出的一种基于 BERT 模型的自然语言处理(NLP)模型。它采用了与 BERT 相似的 Transformer 架构, 但在训练过程中进行了优化, 使用了更多的训练数据、更长的训练时间, 并且去除了 BERT 中的一些限制(如 Next Sentence Prediction 任务), 从而提升了模型的性能。

[3] 余弦相似度(Cosine Similarity)是一种常用于自然语言处理(NLP)和信息检索领域的度量方法, 用来衡量两个向量在多维空间中的方向相似性。在情感相似性评估中, 余弦相似度被用来计算用户输入文本与生成文本在情感向量空间中的相似程度, 反映它们在情感表达上的接近程度。

表 1　评估大语言模型共情模拟能力的三种方法比较

评估方式	定义	方法	优点	局限性	适用场景
人工评估	依靠人类(标注员、用户或专家)对 LLMs 的共情性回应进行主观或比较性评估。	1. Likert 主观评分：人类对模型回应按量表评分(如 1~5 分), 评估情感理解、情感反应、语境适应性、语言自然度等维度。 2. 比较评估：人类判断人类回应与 LLM 回应在同一情境下的共情性优劣。	1. 能够捕捉人类主观感受，提供细致量化评分(如 Likert 量表)。 2. 专家评估可提供专业视角(如心理咨询领域)。 3. 维度明确，评估全面。	1. 主观性强，受评估者背景影响。 2. 耗时耗力，成本高。 3. 比较评估信息量有限，难以横向比较多个模型。	1. 小规模、高精度研究。 2. 需要专业视角的场景(如心理健康咨询)。 3. 评估模型与人类共情回应的细微差异。
自动化评估	利用算法或机器学习技术对 LLMs 共情能力进行客观量化分析。	1. 情感分类：使用预训练模型进行情感分类, 比较用户输入与生成文本的情感一致性。 2. 余弦相似度：计算用户输入与生成文本在情感向量空间的相似度。	1. 高效、快速，适合大规模数据分析。 2. 客观性高，减少人为偏见。 3. 可量化情感相似度，便于模型优化。	1. 可能忽略语境或细微情感差异。 2. 依赖预训练模型质量和词典覆盖率。 3. 难以评估语言自然度或主观共情感受。	1. 大规模模型性能测试。 2. 快速迭代开发的初期评估。 3. 情感分类或向量分析相关研究。
任务驱动评估	通过让 LLMs 完成特定情感任务，评估其情感识别和共情能力。	1. 情感识别任务：评估模型对对话中情感分类的准确性。 2. 情感原因识别：分析对话中情感来源及因果关系。 3.心理学实验任务：分析对话或视频描述, 推测情感和想法。 4. 心理学量表：使用共情量表评估 LLMs 的共情表现。	1. 任务明确，评估结果可量化。 2. 能深入分析情感因果, 提升模型可解释性。 3. 适合测试模型在复杂对话中的表现。	1. 任务设计复杂，需大量标注数据。 2. 可能无法全面反映主观共情能力。 3. 评估范围受任务定义限制。 4.心理学量表可能不完全适用于非人类主体。	1. 情感分析和对话系统研究。 2. 需要评估模型对复杂情感动态的理解。 3. 学术研究中的特定任务测试。

言触发了怎样的情感变化。识别对话中的情感原因有助于全面了解用户的情感状况, 并有助于提升基于情感的模型的可解释性和性能(Zhao et al., 2023)。

除了机器学习领域常用的任务, 心理学领域的众多任务也可用于测量 LLMs 的共情模拟能力。Hagendorff 等人(2023)指出, 将 LLMs 视作心理学实验中的参与者有助于使用心理学的方法测试 LLMs 复杂而新颖的行为模式, 发现更多传统自然语言处理方法无法检测到的能力。例如, 共情准确性的测量已发展出较为完整的范式, 可以通过让 LLMs 完成该任务来评估其共情模拟能力。Ickes 等人(1990)以视频为共情材料设计了自然主义任务, 该范式要求被试观看双人交谈视频并推断其中一方的情感和想法, 衡量的是共情准确性, 即被试在推断他人情感和想法时的准确程度。将这一范式应用于 LLMs 时, 可以通过让模型分析对话或视频描述, 推测主人公的情感和思想, 再与实际情况进行对比, 从而评估模型的认知共情模拟能力。此外, 还有研究者使用人际反应指针量表(International Reactivity Index, IRI), 基本共情量表(Basic Empathy Scale, BES)等心理学量表测量LLMs的共情模拟能力(Pan et al., 2024)。

综上所述, 对 LLMs 共情模拟能力的评估集中于认知共情领域, 评估方式既包括人工主观评估, 又包括利用算法和任务的客观指标评估。人工评估能够考虑情境因素的影响, 通过人工反馈优化模型; 自动化评估能够快速进行量化分析; 任务驱动评估则在实际任务情景中测量 LLMs 的共情能力, 更接近应用场景。不同的评估方法各有特色, 但彼此之间可能存在矛盾和互补之处(见表 1), 未来可以考虑发展统一的标准化测量框架, 从而更加系统化衡量 LLMs 的共情模拟能力。例如, Huang 等人(2025)尝试提出了统一的情感智能测试框架 EmotionBench[4]。但在 LLMs 共情模拟测量方面, 这样的统一框架仍待发展。此外, 近年来

[4] EmotionBench 是一个用于评估和基准测试模型(特别是 LLMs)情感理解与生成能力的框架或数据集。它旨在系统性地测量模型在处理、识别或生成情感相关内容时的表现, 通常包括共情性、情感准确性、语境适应性等维度。EmotionBench 可能涉及多种任务, 如情感分类、情感生成、对话中的共情回应等, 用于比较不同模型或优化模型的情感智能。

使用心理学标准化量表和共情测量范式对 LLMs 的共情模拟能力进行评估成为了新的发展方向，例如，使用心理学量表能够进一步提升评估的准确性；使用心理学范式(如共情准确性实验)则能够探索 LLMs 复杂而新颖的反应模式。

3 LLMs 共情模拟能力发展现状

近两年来，越来越多的研究支持 LLMs 生成共情性回复的能力优于人类或与人类相当的观点。例如, Ayers 等人(2023)研究指出，相较于在线医疗平台的医生给出的回复，人类被试认为 GPT-3 在医疗领域生成的回复共情性更高; Luo 等人(2024)指出由 LLMs 提供支持的聊天机器人有可能在提供共情性沟通方面超越人类医生; Welivita 和 Pu (2024)基于 EmpatheticDialogues (ED)数据集，比较同一情境下 LLMs 生成的共情回复与原数据集中回复的共情水平，发现人类评分者认为 LLMs 的回复更加具有共情性; Lee, Suh 等人(2024)让多个 LLMs 针对描述常见生活经历的帖子(如职场情境、育儿、亲密关系以及其他引发焦虑和愤怒的情境)生成共情性回复，发现 LLMs 生成的回复在共情性评分上始终高于人类撰写的回复。

此外，语言学分析表明，这些模型在标点符号、表情符号以及特定词汇的使用方面，表现出独特且可预测的“风格” (Lee, Suh, et al., 2024)。Loh 和 Raamkumar (2023)在心理学领域设计了一项实验，旨在评估 LLMs 在模拟心理健康咨询对话情境中生成共情性回复的能力，他们发现在大多数情况下, LLMs 的回答相较于人类明显更具共情性。这些研究结果突显了 LLMs 在需要共情能力的场景中具有给予共情性回复、为人类提供支持的潜能。

也有研究发现了 LLMs 在共情模拟的一些子任务上的表现不尽如人意。例如, Zhao 等人(2023)将 LLMs 的情感对话能力区分为理解能力(包括情感识别、情绪原因识别以及行动分类)和生成能力(包括共情性反应和共情支持)，该研究通过一系列下游实验任务来评估 ChatGPT 在情感对话理解和生成方面的表现。结果发现，ChatGPT 在识别人类情绪方面依然有待提高，例如，在医生和患者的对话中，当患者描述头痛症状时，数据集注释将情绪归类为中性，而 ChatGPT 则将其视为悲伤，这种标准的不一致可能并非源于 ChatGPT 的能力，而是缺少样本提示导致的。

在共情反应方面，ChatGPT 经常先复述用户情绪，然后再扩展信息，这种模式可能会让用户感到厌烦；一旦确认用户面临的困境，ChatGPT 就过于着急地给出相应的建议和应对策略，而忽略了对用户情绪的抚慰和关怀。此外，Pan 等人(2024)使用心理学量表(基本共情量表和人际反应指针量表)比较了 1200 名由 GPT-4 模拟生成的参与者与 1200 名真人参与者在量表上的得分，结果发现，GPT-4 模拟参与者在两个量表上的得分水平显著低于人类。虽然以往一些研究通过特定情境评估 LLMs 的共情模拟能力时发现它们表现出一定水平的共情识别和反应能力，但通过标准化问卷的评估结果发现, LLMs 目前尚未达到人类的共情水平。

综上所述, LLMs 在特定情境下生成共情性回复的能力已经与人类相当，甚至超越人类，在需要共情的情感场景中能够给予人类情感支持；而在共情模拟的一些子任务以及标准化共情量表测试的结果上依然存在上升空间，模拟共情的能力还有待于进一步提升。随着 LLMs 快速发展，目前的研究结果也许不能完全代表 LLMs 共情模拟能力的发展状况。例如，目前仍没有研究对 DeepSeek 模型[5]的共情能力进行测量。

更重要的是，以上有关 LLMs 共情模拟的研究正在动摇人类关于共情本质的理解。长期以来，共情被认为是人类独有的特质，其定义通常基于人与人之间的互动。共情的复杂性受个性、文化和情境的影响，导致该术语的定义存在模糊性。LLMs 表现出的共情模拟本质上与人类共情不同，因为算法并不会真正经历情感体验。因此，人类中心主义的共情定义可能并不适用于 LLMs。但是，如果 LLMs 生成的回应符合人类的期望或偏好，那么这些可观察到的响应是否可以被视为共情？当人们无法区分人类和 LLMs 生成的回应，甚至人们更偏

[5] DeepSeek 模型是由中国人工智能公司 DeepSeek (杭州深度求索人工智能基础技术研究有限公司)开发的一系列开源大型语言模型，旨在以低成本、高效率的方式提供高性能的 AI 解决方案，适用于通用对话、编程、推理和多模态任务。这些模型以开源形式发布(通常采用 MIT 许可证)，广泛应用于研究、开发和商业场景，挑战了如 OpenAI、GPT-4 等闭源模型的市场主导地位。

好 LLMs 生成的回答时, 这可能表明 LLMs 模拟共情本身已足够(Huang et al., 2025)。

4　LLMs 共情模拟能力的提升策略

LLMs 对于人类情感和想法的推断能力还有相当的提升空间, 进一步考虑文化和情境的影响, 发展与人类情感对齐的 LLMs 还需要进行数据增强、架构优化、强化学习以及合理的引导词设计。

4.1　数据增强

构建包含更多情感信息的共情对话数据集, 以实现对大语言模型的微调。目前最为常用的情感对话数据集是 Rashkin 等人(2018)提出的基于情感情景的对话数据集 EmpatheticDialogues[6](ED), 该数据集包含 25000 轮情感对话。为了进一步提高 LLMs 的共情模拟能力, 许多研究者基于 ED 数据集进行数据增强, 并将增强后的数据用于 LLMs 微调, 从而提高其生成共情对话的能力。例如, Cao 等人(2025)使用大模型对 ED 数据集中包含情感信息的语段进行筛选, 并完成情感信息标注, 构建了 TOOL-ED 数据集; 发现利用该数据集对 LLMs 进行微调, 可以提高其生成共情对话的能力。

增大数据集、提高数据质量、增加对话轮次也能够增强共情对话数据集, 从而提高 LLMs 的共情模拟能力。例如, Synth-Empathy 是一种基于 LLMs 的数据生成与筛选方法, 用于自动生成高质量的共情对话数据(Liang et al., 2024)。它通过优化数据生成流程, 高效、低成本地创建大规模共情数据集, 同时通过质量与多样性选择机制, 自动保留高质量数据并剔除低质量数据, 确保生成的对话数据在共情性、相关性和多样性方面达到高标准。Chen 等人(2023)构建了一个包含超过 200 万个样本的多轮共情对话数据集, 训练模型在面对不同情境时能够使用提问、安慰、认可、倾听、信任等多种情感支持的表达。实验表明, 通过使用更接近心理咨询师表达的多轮对话进行微调, LLMs 的共情模拟能力可以显著增强。

4.2　模型架构与框架的优化

通过改进和调整模型的结构、组件和信息流动方式, 可以提升 LLMs 在生成共情对话任务上的表现。Cai 等人(2024)设计的新模型 Pecer[7], 能够动态捕捉历史对话中的情感信息和人格信息, 并将获取的信息融入到共情反应的生成中, 从而得到更好的共情效果。Liu 等人(2024)结合语法依赖树和情感词汇模块(National Research Council Emotion Lexicon, NRC[8])从历史对话中提取情感词, 构造情感依赖树, 从而更好捕捉情感信息, 对讲述者进行共情。

除了优化 LLMs 的底层架构, 构建 LLMs 的使用框架也能提高 LLMs 的共情模拟能力。LLMs 在生成共情回复方面十分出色, 但缺乏深入理解情绪和认知情感之间细微差别的能力, 而小型共情模型(small-scale empathetic models, SEMs[9])则与之相反。混合共情框架(Hybrid Empathetic Framework, HEF, Yang et al., 2024)将 SEMs 视为灵活的插件, 可以提高 LLMs 细微的情感和认知理解, 缓解了 LLMs 在细粒度情绪检测方面的困难, 并提高了 LLMs 识别情绪原因的能力。而生成前共情(Empathizing Before Generation, EBG)框架(Zhu et al., 2024)能够使 LLMs 在生成回复之前分析思维链(chain of thought, COT[10]), 提高了推断

[6] EmpatheticDialogues 是一个包含约 25, 000 段对话的开源数据集, 专注于共情对话研究, 涵盖 32 种情感类别, 广泛用于训练和评估 LLMs 的共情能力。其对话数据支持情感分析、模型优化和心理健康应用, 但需注意语言和文化局限性。

[7] Pecer 全称为 Prompt-enhanced Empathetic Conversation and Evaluation Model (提示增强的共情对话与评估模型)。该模型旨在通过结合引导词工程(prompt engineering)和强化学习(RL)技术, 生成高质量的共情对话数据, 同时能够评估和优化对话中的共情性。Pecer 特别适用于生成符合人类情感期望的对话, 广泛应用于情感支持、心理健康辅助和客户服务等场景。

[8] NRC Emotion Lexicon (National Research Council Emotion Lexicon, 简称 NRC 情感词典)是由加拿大国家研究委员会(National Research Council Canada)开发的一种情感分析资源, 广泛用于自然语言处理(NLP)和情感计算领域。它是一个包含大量词汇及其情感标注的词典, 用于识别文本中的情感信息(如喜悦、悲伤、愤怒等)。NRC 情感词典通过将词汇与特定情感类别和情感强度关联, 帮助研究者和开发者分析文本的情感倾向、情感强度以及相关情感特征。

[9] 指专门设计用于情感理解和共情任务的小型神经网络模型。相较于 LLMs, SEMs 参数规模小, 但在细粒度情绪检测和认知细微差别(如情感原因推断)方面表现更优。SEMs 通常通过特定情感数据集(如 EmotionBench)训练, 专注于捕捉对话中的情感细节。

[10] 一种提示策略, 引导 LLMs 在生成最终输出前逐步推理, 分解复杂问题为多个逻辑步骤。COT 通过显式表达推理过程(如情感推断的中间步骤), 提高模型在情感分析、逻辑推理和问题解决中的准确性。在 EBG 中, COT 用于分析情绪原因和语境。

情绪的能力和识别情绪标签的准确性。

4.3 强化学习

通过引入人类或 LLMs 对生成回复的反馈, 从而优化模型在情感表达和自然度上的表现。该方法通过奖励机制引导模型生成更贴合人类期望的回应。例如, Wang, Xu 等人(2024)引入了一个名为Muffin的反馈框架, 通过多方面的AI反馈来评估生成回复的有效性。具体来说, 该框架通过对比“有效” (helpful)和“无效” (unhelpful)回复来优化该模型, 使其更能区分高质量回应, 降低生成低质量回复的概率。不过, 引入人类反馈的评估流程还有待规范。例如, 在 Sabour 等人(2022)的研究中, 仅招募 3 人比较了 100 组不同系统生成的共情对话内容, 以此来对比哪个系统更好。这种评价方式很难保证结果的稳定性以及在解决实际问题中的有效性。Sharma 等人(2023)使用随机对照实验对比人类和人类+人工智能两种回复中的共情反应。但该研究中所使用的评价标准仍只是第三方视角下对话内容所展现出的共情性, 而非被共情者本人真实感受到的共情, 应考虑引入规范的被共情者反馈引导模型生成更贴合人类期望的共情性回应。

4.4 引导词设计

通过调整输入提示或上下文信息, 可以引导LLMs 生成符合特定情感需求的回应。Welivita 和 Pu (2024)发现在引导词(prompt)中加入对于共情的详细定义将有助于 LLMs 生成共情性的回复。Lee, Lee 等人(2024)将咨询心理学的理论模型融入 Prompt 中, 提出共情链(Chain-of-Empathy Prompting, CoE[11]), 从而优化 LLMs 的共情性回复。例如, 在结合认知行为疗法优化共情反应时, 引导词提示 LLMs 分析用户表达的情感及其表述中可能存在的认知陷阱, 然后再给出回应。经过提示后的 LLMs 给出的回应更具有共情性。

总之, LLMs 的共情模拟能力日益受到关注, 其在共情对话中的表现优化已成为情感计算领域的核心研究方向。研究者通过多维度策略提升LLMs 的共情模拟能力, 包括数据增强、模型架构优化、强化学习和精细输入提示(见表 2)。这些策略虽均增强了 LLMs 的共情模拟能力, 但侧重点各异, 未来需结合实际应用需求优化策略组合, 以生成更贴合人类期待的共情回应。

当前, 大多数优化策略的比较基线依赖EmpatheticDialogues 数据集, 限制了评估的多样性和现实适用性。未来研究可结合人工评估和心理学实验范式(如控制实验、情感量表测量), 在多样化的实际应用场景(如心理咨询、客户服务)中验证优化后模型的共情能力。此外, 现有提升策略多基于 LLMs 的预训练与微调, 依赖其语言生成和逻辑推理能力。因此, LLMs 基础能力的持续改进对共情优化至关重要, 不应忽视。

5 LLMs 共情模拟能力的挑战与局限

5.1 共情模拟的真实性受限

LLMs 依靠自然语言处理和统计学习来生成文本, 它们的“共情”表现主要来自于模式学习(侯悍超 等, 2024), 即其机制本质上是基于对海量统计数据的分析与模拟, 而非源于人类真实的情感体验和内在理解(Shao et al., 2024; Sorin et al., 2024)。此外, LLMs 能够识别文本中的情感倾向, 并生成相应的共情回复; 通过上下文理解可以选择适当的语言风格, 使其表达更加符合语境。但是, LLMs 所产生的共情性回应, 常表现为情感词汇的过度堆砌以及回复模式的刻板化(Naik et al., 2024; Sorin et al., 2024)。这导致其难以展现对情感真切、细腻的深层理解, LLMs 共情表达的真实性也因此受到质疑。近期一项研究比较了基于 LLMs 的聊天机器人辅助宣泄与传统日记平台宣泄在减少负面情绪和增加感知社会支持方面的有效性。结果发现, 聊天机器人辅助宣泄有效地降低了高唤醒和中等唤醒的负面情绪, 如愤怒、沮丧和恐惧。然而, 参与者在聊天机器人辅助宣泄条件下的感知社会支持并未显著增加, 并且仍然感知到孤独, 这表明参与者可能并未将聊天机器人的有效辅助视为社会支持(Hu et al., 2025)。

5.2 难以应对复杂情感与情境

人类的情感表达和感受具有高度的复杂性, 缺乏情感理解的 LLMs 难以进行恰当回应。例如, 个体可能体验到悲喜交加等复合情感状态, 并可

[11] CoE (Chain-of-Empathy Prompting, 共情链提示) 是一种引导词工程(prompt engineering)方法, 通过将咨询心理学的理论模型融入提示设计, 优化 LLMs 的共情输出。具体来说, CoE 引导 LLMs 在生成回复前, 通过一系列结构化的推理步骤(类似思维链, COT), 分析用户的情感状态、潜在的认知偏差或心理需求, 从而生成更具共情性、语境贴合性和心理支持效果的回应。

表 2　提升大语言模型共情模拟能力的 4 种策略比较

提升策略	定义	方法	优点	局限性	适用场景
数据增强	构建或优化包含更多情感信息的共情对话数据集, 并在数据集上对 LLMs 进行微调, 从而优其共情表现。	1. 基于现有数据集(如 EmpatheticDialogues, ED)进行增强, 筛选并标注情感信息, 构建新数据集。 2. 自动生成高质量共情数据, 丢弃低质量数据。 3. 构建多轮共情对话数据集, 融入提问、安慰等情感支持表达。	1. 提升数据质量和多样性, 直接增强模型共情能力。 2. 可生成大规模数据集, 降低数据收集成本。 3. 多轮对话数据贴近真实场景, 效果显著。	1. 数据标注和筛选需大量计算资源和人力。 2. 增强数据可能引入新偏见或噪声。 3. 依赖原始数据集的质量和覆盖范围。	1. 对话系统开发需要高质量情感数据时。 2. 心理咨询或情感支持场景的模型训练。 3. 数据资源有限但需快速扩展的场景。
模型架构与框架优化	通过改进 LLMs 的结构、组件或使用框架, 增强其在共情对话任务中的表现。	1. 设计新模型(如 Pecer)动态捕捉情感和人格信息, 融入共情回应生成。 2. 结合语法依赖树和 NRC 情感词典构建情感依赖树, 捕捉情感信息。 3.将小型共情模型(SEMs)作为插件提升 LLMs 细粒度情感理解。 4. 使用生成前共情(EBG)框架, 通过思维链分析提升情绪推断。	1. 针对性优化模型结构, 效果精准。 2. 框架灵活, 可结合多种技术提升共情能力。 3. 能处理细粒度情感和复杂对话场景。	1. 模型设计和优化需深厚技术背景和高计算成本。 2. 新架构可能增加推理时间或资源需求。 3. 框架复杂性可能降低模型可解释性。	1. 学术研究探索新型共情模型架构。 2. 需要处理复杂情感动态的对话系统。 3. 结合小型模型优化大型模型的场景。
强化学习	通过引入人类或 AI 反馈, 引导生成更贴合人类期望的回应。	1. 使用反馈框架(如 Muffin)通过 AI 反馈和对比学习优化回应质量。 2. 开展随机对照实验, 比较人类与 LLMs 回应的共情性, 将结果反馈给 LLMs。 3. 小规模人类评估对比生成内容。	1. 直接利用人类偏好, 生成更自然的共情回应。 2. 反馈机制可持续优化模型表现。 3. 能针对特定情感场景(如悲伤)优化。	1. 反馈流程缺乏统一规范, 评估稳定性不足。 2. 小规模反馈可能无法代表广泛用户需求。 3. 收集高质量人类反馈成本高, 耗时长。	1. 需要高用户满意度的商业应用。 2. 迭代优化已有模型的共情表现。 3. 针对特定用户群体的个性化共情优化。
引导词设计	通过优化输入提示或上下文信息, 引导 LLMs 生成符合特定情感需求的共情回应。	1. 在提示中加入共情定义, 引导生成共情性回复。 2. 融入咨询心理学理论(如提出共情链提示), 分析情感和认知陷阱。	1. 实现成本低, 易于实施和调整。 2. 可快速提升现有模型的共情表现。 3. 能结合心理学理论, 生成专业化回应。	1. 效果高度依赖提示设计质量。 2. 可能无法解决模型的情感理解局限。 3. 在复杂对话中可能缺乏稳定性。	1. 快速测试或原型开发阶段。 2. 心理咨询或教育场景需要专业化回应。 3. 资源或时间有限时。

能采用反讽等非字面表达方式来传递其情感。然而, 在面对此类复杂的情感和表达时, LLMs 倾向于从用户输入的情绪词中选取进行回应, 表现出的情绪智力较低, 难以实现有效的共情性回复(Naik et al., 2024)。此外, LLMs 在处理涉及社会互动和情感理解的任务时表现出明显的局限性。例如, 在对社交智力和“心理理论” (Theory of Mind)的评估中, GPT-3 等模型在理解社交互动中参与者的意图和反应方面表现不佳, 准确率仅为 55%至 60%, 远低于人类水平(Sap et al., 2022)。

文化情境的差异性进一步加剧了 LLMs 在提供文化敏感性共情内容时所面临的挑战。不同文化背景下, 共情表达方式各异, LLMs 难以灵活适应这些差异。由于训练数据的偏差, 如今大部分 LLMs 生成的文本表现出“以英语文化为中心”的倾向, 即使在回应其他语言的提示时, 仍然更多地体现西方情感表达规范。这表明, 目前的多语言 LLMs 还未能成功学习不同文化中适当的情感表达细微差别(Havaldar et al., 2023)。

5.3　伦理与使用风险

首先, LLMs 所生成的共情模拟内容存在潜在的内容风险。例如, 鉴于训练数据及算法固有的

局限性, LLMs 不可避免地具备生成包含攻击性、歧视性等负面内容信息的可能性(Cuadra et al., 2024; Patil et al., 2024; Shen et al., 2024)。此类不良信息不仅会显著降低用户体验, 更可能对个体心理健康造成负面冲击。

其次, LLMs 的共情模拟能力可能被滥用, 导致信息操纵或用户对模型产生心理依赖, 或许有悖公序良俗。例如, 聊天机器人通过模拟人类行为, 例如友好的语气、情感化表达以及开放式问题的设计, 能够显著降低用户的戒备心理。这种设计策略使用户倾向于将机器人视为具有某种人类意图的实体。即使用户理性上知道聊天机器人是非人类的技术产物, 他们仍然可能会表现出对机器人的情感依赖, 并在交互过程中分享隐私信息(Holmes et al., 2022), 当用户逐渐习惯于从 LLMs 处寻求情感支持与共情反馈时, 反而会加剧个体的孤独体验(Koranteng et al., 2023)。此外, 用户与 LLMs 进行情感交流, 短期内可能会增加用户体验的愉悦度, 但长期使用或将弱化其与真人进行情感互动的主动意愿与能力, 导致人类对真实情感互动的敏感性下降, 并对现实人际关系及社会适应性造成不利影响。

最后, LLMs 的训练数据通常来自互联网、书籍、社交媒体等广泛来源, 这些数据不可避免地包含社会文化偏见和刻板印象。这些偏见可能在 LLMs 生成共情性回复时被无意中放大和传播, 进而引发伦理风险。例如, LLMs 通过统计概率生成回复, 倾向于复现训练数据中高频出现的语言模式。如果数据中某些群体被刻板化(如“亚洲人擅长数学”), LLMs 可能在共情回应中无意中重复这些刻板印象。例如当用户提到一个亚洲裔学生的学业压力时, LLMs 可能生成类似“你的数学成绩一定很好, 但压力也大吧? ”的回应, 强化了刻板印象。又如, 对女性用户的职业困境, LLMs 可能生成“作为女性, 你可能更擅长协调工作, 也许可以考虑行政岗位”的回应, 暗含性别偏见, 却因“共情”语气显得“合理”。 当偏见被 LLMs 以看似善意的共情方式表达和传播时, 其潜在的危害性可能被掩盖, 从而潜移默化地影响使用者的认知和行为, 固化现有偏见和印象, 甚至加剧社会不平等。

6 总结

随着 LLMs 的不断发展, 其在共情模拟能力方面的应用已成为人工智能研究的热点议题。尤其在医疗健康和心理咨询领域, LLMs 展现出了潜在的应用价值。现有研究主要通过评估模型的情感理解、情感反应、语境适应性和语言自然度等多个维度, 探讨如何增强 LLMs 在互动过程中展现类似于人类的共情反应。

目前, LLMs 共情模拟能力的评估方法日趋完善, 研究者从不同角度对其共情模拟水平进行了测量, 包括人工主观评估、算法自动化评估和基于任务驱动的评估方法。多样化的评估手段多维度评估了 LLMs 的共情反应能力, 但仍有不足。未来应当致力于构建系统性测量 LLMs 共情模拟能力的统一框架, 并考虑纳入心理学的共情测量范式以探索 LLMs 更加复杂的共情反应模式。

根据已有的评估方法, 多数研究支持 LLMs 生成共情性回复的能力与人类相当的观点, 但在准确推断人类情感和想法方面的能力仍待提升。目前已有的提升策略包括数据增强、模型架构优化、基于人类反馈的强化学习和精细化的引导词设计, 在不同层面上提高了 LLMs 的共情模拟能力, 未来应根据实际需要进行选择, 以适应更加复杂的情感场景和应用需求。

然而, 尽管 LLMs 在共情表达方面取得了一定进展, 相关研究也凸显了伦理风险的潜在影响, 引发了对技术滥用、用户隐私保护和社会信任度等问题的关注。为了保证伦理与隐私问题得到妥善解决, 设计者需在增强聊天机器人功能与保护用户数据安全和心理健康之间取得平衡, 以防止技术滥用并增强用户对人工智能的信任。

参考文献

侯悍超, 倪士光, 林书亚, 王蒲生. (2024). 当 AI 学习共情: 心理学视角下共情计算的主题、场景与优化. *心理科学进展, 32*(5), 845−858

Ayers, J. W., Poliak, A., Dredze, M., Leas, E. C., Zhu, Z., Kelley, J. B., ... Smith, D. M. (2023). Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. *JAMA Internal Medicine*, *183*(6), 589−596. https://doi.org/10.1001/jamainternmed.2023.1838

Cai, M., Wang, D., Feng, S., & Zhang, Y. (2024). PECER: Empathetic response generation via dynamic personality extraction and contextual emotional reasoning. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 10631−10635). IEEE. http://10.1109/ICASSP48485.2024.10446914

Cao, H., Zhang, Y., Feng, S., Yang, X., Wang, D., & Zhang, Y. (2025). TOOL-ED: Enhancing empathetic response generation with the tool calling capability of LLM. In *Proceedings of the 31st International Conference on Computational Linguistics* (pp. 5305−5320). Association for Computational Linguistics. https://aclanthology.org/2025.coling-main.355/

Chen, Y., Xing, X., Lin, J., Zheng, H., Wang, Z., Liu, Q., & Xu, X. (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* (pp. 1170−1183). Association for Computational Linguistics. https://aclanthology.org/2023.findings-emnlp.83/

Cuadra, A., Wang, M., Stein, L. A., Jung, M. F., Dell, N., Estrin, D., & Landay, J. A. (2024). The illusion of empathy? Notes on displays of emotion in human-computer interaction. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (pp. 1−18). https://doi.org/10.1145/3613904.3642336

Dillion, D., Tandon, N., Gu, Y., & Gray, K. (2023). Can AI language models replace human participants? *Trends in Cognitive Sciences*, *27*(7), 597−600. https://doi.org/10.1016/j.tics.2023.04.008

Hagendorff, T., Dasgupta, I., Binz, M., Chan, S. C., Lampinen, A., Wang, J. X., ... Schulz, E. (2023). Machine psychology. *arXiv*. https://doi.org/10.48550/arXiv.2303. 13988

Havaldar, S., Singhal, B., Rai, S., Liu, L., Guntuku, S. C., & Ungar, L. (2023). Multilingual language models are not multicultural: A case study in emotion. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis* (pp. 202−214). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.wassa-1.19

Holmes, W., Porayska-Pomsta, K., Holstein, K., Sutherland, E., Baker, T., Shum, S. B., & Koedinger, K. R. (2022). Ethics of AI in education: Towards a community-wide framework. *International Journal of Artificial Intelligence in Education*, *32*, 504−526. https://doi.org/10.1007/s40593-021-00239-1

Hu, M., Chua, X. C. W., Diong, S. F., Kasturiratna, K. S., Majeed, N. M., & Hartanto, A. (2025). AI as your ally: The effects of AI‐assisted venting on negative affect and perceived social support. *Applied Psychology: Health and Well‐Being, 17*(1), e12621. https://doi.org/10.1111/aphw.12621

Huang, J. T., Lam, M. H., Li, E. J., Ren, S., Wang, W., Jiao, W., & Lyu, M. R. (2025). Apathetic or empathetic? Evaluating LLMs' emotional alignments with humans. *Advances in Neural Information Processing Systems*, *37*, 97053−97087. https://doi.org/10.48550/arXiv.2308.03656

Ickes, W., Stinson, L., Bissonnette, V., & Garcia, S. (1990). Naturalistic social cognition: Empathic accuracy in mixed-sex dyads. *Journal of Personality and Social Psychology*, *59*(4), 730−742. https://doi.org/10.1037/0022-3514.59.4.730

Koranteng, E., Rao, A., Flores, E., Lev, M., Landman, A., Dreyer, K., & Succi, M. (2023). Empathy and equity: Key considerations for large language model adoption in health care. *JMIR Medical Education*, *9*, e51199. https://doi.org/10.2196/51199

Laskar, M. T. R., Bari, M. S., Rahman, M., Bhuiyan, M. A. H., Joty, S., & Huang, J. (2023). A systematic study and comprehensive evaluation of ChatGPT on benchmark datasets. In *Findings of the Association for Computational Linguistics: ACL 2023* (pp. 431−469). Association for Computational Linguistics. https://aclanthology.org/2023.findings-acl.29/

Lee, Y. K., Lee, I., Shin, M., Bae, S., & Hahn, S. (2024). Enhancing empathic reasoning of large language models based on psychotherapy models for AI-assisted social support. *Korean Journal of Cognitive Science, 35*(1), 23−48. https://doi.org/10.19066/cogsci.2024.35.1.002

Lee, Y. K., Suh, J., Zhan, H., Li, J. J., & Ong, D. C. (2024). Large language models produce responses perceived to be empathic. *arXiv*. https://doi.org/10.48550/arXiv.2403.18148

Liang, H., Sun, L., Wei, J., Huang, X., Sun, L., Yu, B., ... Zhang, W. (2024). Synth-empathy: Towards high-quality synthetic empathy data. *arXiv*. https://doi.org/10.48550/arXiv.2407.21669

Liu, Y., Han, D., Wu, G., & Qiao, B. (2024). KnowDT: Empathetic dialogue generation with knowledge-enhanced dependency tree. *Applied Intelligence*, *54*(17), 8059−8072. https://doi.org/10.1007/s10489-024-05611-x

Liu-Thompkins, Y., Okazaki, S., & Li, H. (2022). Artificial empathy in marketing interactions: Bridging the human-AI gap in affective and social customer experience. *Journal of the Academy of Marketing Science*, *50*(6), 1198−1218. https://doi.org/10.1007/s11747-022-00892-5

Loh, S. B., & Sesagiri Raamkumar, A. (2023). Harnessing large language models' empathetic response generation capabilities for online mental health counselling support. *arXiv*. https://doi.org/10.48550/arXiv.2310.08017

Luo, M., Warren, C. J., Cheng, L., Abdul-Muhsin, H. M., & Banerjee, I. (2024). Assessing empathy in large language models with real-world physician-patient interactions. In *2024 IEEE International Conference on Big Data (BigData)* (pp. 6510−6519). IEEE. https://doi.org/10.1109/BigData62323.2024.10825307

Ma, J., Chen, B., Wang, K., Hu, Y., Wang, X., Zhan, H., & Wu, W. (2024). Emotional contagion and cognitive empathy regulate the effect of depressive symptoms on empathy-related brain functional connectivity in patients with chronic back pain. *Journal of Affective Disorders*, *362*, 459−467. https://doi.org/10.1016/j.jad.2024.07.026

Naik, N., Jenkins, P., Prajapat, S., & Grace, P. (Eds.). (2024). *Contributions Presented at The International Conference on Computing, Communication, Cybersecurity and AI,*

July 3−4, 2024, London, UK: The C3AI 2024 (1st ed.). Springer Cham. https://doi.org/10.1007/978-3-031-74443-3

Pan, S., Fan, C., Zhao, B., Luo, S., & Jin, Y. (2024). Can large language models exhibit cognitive and affective empathy as humans? *OSF Preprints*. https://doi.org/10.31219/osf.io/w5rsu

Patil, D. D., Dhotre, D. R., Gawande, G. S., Mate, D. S., Shelke, M. V., & Bhoye, T. S. (2024). Transformative trends in generative ai: Harnessing large language models for natural language understanding and generation. *International Journal of Intelligent Systems and Applications in Engineering*, *12*(4s), 309−319. https://ijisae.org/index.php/IJISAE/article/view/3794

Poria, S., Majumder, N., Hazarika, D., Ghosal, D., Bhardwaj, R., Jian, S. Y. B., ... Mihalcea, R. (2021). Recognizing emotion cause in conversations. *Cognitive Computation*, *13*(5), 1317−1332. https://doi.org/10.1007/s12559-021-09925-7

Rashkin, H., Smith, E. M., Li, M., & Boureau, Y. L. (2018). Towards empathetic open-domain conversation models: A new benchmark and dataset. *arXiv*. https://doi.org/10.48550/arXiv.1811.00207

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (pp. 3982−3992). Association for Computational Linguistics. https://doi.org/10.48550/arXiv.1908.10084

Ren, C., Zhang, Y., He, D., & Qin, J. (2024). WundtGPT: Shaping large language models to be an empathetic, proactive psychologist. *arXiv* preprint arXiv:2406.15474. https://aclanthology.org/D19-1410/

Sabour, S., Zheng, C., & Huang, M. (2022). CEM: Commonsense-aware empathetic response generation. *Proceedings of the AAAI Conference on Artificial Intelligence*, *36*(10), 11229−11237. https://doi.org/10.1609/aaai.v36i10.21373

Sap, M., Le Bras, R., Fried, D., & Choi, Y. (2022). Neural theory-of-mind? On the limits of social intelligence in LLMs. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing* (pp. 3762−3780). Association for Computational Linguistics. https://aclanthology.org/2022.emnlp-main.248/

Shao, M., Basit, A., Karri, R., & Shafique, M. (2024). Survey of different large language model architectures: Trends, benchmarks, and challenges. *IEEE Access*, *12*, 188664−188706. https://doi.org/10.1109/ACCESS.2024.3482107

Sharma, A., Lin, I. W., Miner, A. S., Atkins, D. C., & Althoff, T. (2023). Human−AI collaboration enables more empathic conversations in text-based peer-to-peer mental health support. *Nature Machine Intelligence*, *5*(1), 46−57. https://doi.org/10.1038/s42256-022-00593-2

Shen, X., Chen, Z., Backes, M., Shen, Y., & Zhang, Y. (2024). "Do anything now": Characterizing and evaluating in-the-wild jailbreak prompts on large language models. In *Proceedings of the 2024 ACM SIGSAC Conference on Computer and Communications Security* (pp. 1671−1685). Association for Computational Linguistics. https://doi.org/10.1145/3658644.3670388

Sorin, V., Brin, D., Barash, Y., Konen, E., Charney, A., Nadkarni, G., & Klang, E. (2024). Large language models and empathy: Systematic review. *Journal of Medical Internet Research*, *26*, e52597. https://doi.org/10.2196/ 52597

Svikhnushina, E., & Pu, P. (2022). PEACE: A model of key social and emotional qualities of conversational chatbots. *ACM Transactions on Interactive Intelligent Systems*, *12*(4), 1−29. https://doi.org/10.1145/3531064

Wang, J., Xu, C., Leong, C. T., Li, W., & Li, J. (2024). Mitigating unhelpfulness in emotional support conversations with multifaceted AI feedback. *arXiv*. https://doi.org/10.48550/arXiv.2401.05928

Wang, L., Ma, C., Feng, X., Zhang, Z., Yang, H., Zhang, J., ... Wen, J. (2024). A survey on large language model based autonomous agents. *Frontiers of Computer Science, 18*(6), 186345. https://doi.org/10.1007/s11704-024-40231-1

Welivita, A., & Pu, P. (2024). Are large language models more empathetic than humans?. *arXiv*. https://doi.org/10.48550/arXiv.2406.05063

Yang, Z., Ren, Z., Yufeng, W., Peng, S., Sun, H., Zhu, X., & Liao, X. (2024). Enhancing empathetic response generation by augmenting LLMs with small-scale empathetic models. *arXiv*. https://doi.org/10.48550/arXiv.2402.11801

Zhao, W., Zhao, Y., Lu, X., Wang, S., Tong, Y., & Qin, B. (2023). Is ChatGPT equipped with emotional dialogue capabilities? *arXiv*. https://doi.org/10.48550/arXiv.2304.09582

Zhu, J., Jiang, Z., Zhou, B., Su, J., Zhang, J., & Li, Z. (2024). Empathizing before generation: A double-layered framework for emotional support LLM. In *Lecture Notes in Computer Science* (pp. 490−503). Springer Nature Switzerland. https://doi.org/10.1007/978-981-97-8490-5_35

Zhuang, Z., Chen, Q., Ma, L., Li, M., Han, Y., Qian, Y., ... Liu, T. (2023). Through the lens of core competency: Survey on evaluation of large language models. In *Proceedings of the 22nd Chinese National Conference on Computational Linguistics (Volume 2: Frontier Forum)* (pp. 88−109). Chinese Information Processing Society of China. https://aclanthology.org/2023.ccl-2.8/

Empathy in large language models: Evaluation, enhancement, and challenges

ZHOU Qianyi[1], CAI Yaqi[1], ZHANG Ya[1,2]

([1] *Shanghai Key Laboratory of Mental Health and Psychological Crisis Intervention, School of Psychology and Cognitive Science, East China Normal University, Shanghai* 200062, *China*)

([2] *Key Laboratory of Philosophy and Social Science of Anhui Province on Adolescent Mental Health and Crisis Intelligence Intervention, Hefei Normal University, Hefei* 230601, *China*)

Abstract: With advances in large language models (LLMs) in natural language generation and affective computing, their ability to simulate empathy has attracted increasing attention in domains such as psychological counseling, doctor-patient communication, and customer service. Empathy simulation in LLMs primarily manifests as cognitive rather than emotional empathy, encompassing emotion recognition, empathic response generation, and contextual adaptation. Current evaluation methods for empathic simulation in LLMs include human evaluation, automated evaluation, and task-based evaluation, each with its own advantages and limitations depending on the application context. Compared to human empathy, LLMs perform well in empathy generation tasks but still face fundamental limitations in emotional understanding. To further enhance the empathic capabilities of LLMs, techniques such as data augmentation, model architecture optimization, reinforcement learning, and prompt engineering can be employed. Meanwhile, ethical considerations and potential risks associated with the use of these models remain important concerns.

Keywords: large language models, empathy simulation, empathy evaluation, ethical issues