RPsyEmqa:基于报告的问答重建与检索增强生成的青少年心理健康辅导框架

RPsyEmqa:A Report-Based Question-Answer Reconstruction and Retrieval-Augmented Generation Framework for Adolescent Mental Health Counseling

陈鹏鹤[1]，董吴桐[2*]，周俞君[3]
[1]北京师范大学未来教育高精尖创新中心
[2]北京师范大学教育学部教育技术学院
[3]北京师范大学教育学部教育技术学院
[*] z_turn@mail.bnu.edu.cn

【摘要】 利用大语言模型辅助青少年心理辅导是一项重要且具有挑战性的任务。现有研究通常通过微调大模型来增强其在这一领域的表现，但现有数据集往往缺乏足够专业的心理辅导知识且微调无法避免大模型回答中的“幻觉”问题。为了弥合这一差距，我们提出了 RPsyEmqa，一个基于报告的问答构建与评估框架。我们设计了一种两阶段的方法来构建高质量的问答数据集，同时开发了基于检索增强生成框架的问答系统，竞争性实验结果证明了其在心理辅导方面的有效性。我们为未来的研究开放了源代码和模型。

【关键词】 大语言模型；青少年心理辅导；检索增强生成

Abstract: *Using large language models to assist in adolescent psychological counseling is a significant yet challenging task. Existing studies often enhance the performance of LLMs in this domain through fine-tuning. However, current datasets typically lack sufficient professional psychological counseling knowledge, and fine-tuning cannot avoid the "hallucination" issues in LLM responses. To bridge this gap, we propose RPsyEmqa, a report-based question-answering construction and evaluation framework. We designed a two-stage approach to build a high-quality question-answering dataset and developed a retrieval-augmented generation question-answering system. Competitive experimental results demonstrate the effectiveness of our framework in psychological counseling. We have made the source code and model available for future research.*

Keywords: Large Language Models, Adolescent Psychological Counseling, Retrieval-Augmented Generation

1.前言

我国大中小学生在抑郁、焦虑以及自我伤害等方面的检出率较高，心理健康状况令人堪忧。然而，由于许多教师缺乏足够的心理学知识与辅导技能，心理健康教育往往停留在初步意识层面，难以提供个性化的辅导支持。近年来，随着自然语言处理（Natural Language Processing, NLP）技术的发展，大型语言模型（Large Language Models, LLMs）如 ChatGPT 和 LLaMA（Touvron 等，2023）在人工智能辅助心理咨询领域取得了一定进展，其研究方向从心理疾病咨询和情感支持提升扩展到在线心理咨询及治疗辅助，逐步成为有效的治疗助手。例如，Psy-LLM（Lai 等，2023）通过结合专业心理学家问答数据和大量中文心理学文章进行训练，展现了其在心理咨询服务中的专业能力。同时，其他心理模型如 MeChat（Qiu 等，2024）和 SoulChat（Chen 等，2023）也已在线提供服务。

尽管相关领域取得了诸多进步，但在青少年心理健康问答方面，仍然缺乏真实、专业的数据支持，且 LLMs 在回答中不可避免地存在“幻觉”（Ji 等，2023）问题从而降低了心理辅导的专业性。而现实世界中积累了大量以问诊笔记形式记录的青少年心理咨询数据，这些数据来自专业的心理咨询师并具有高度专业性，但其结构化特性，并不适合直接用于模型训练。基于此，本文提出了一种新的框架—RpsyEmqa，核心包括两大模块：基于心理咨询笔记的问答数据生成和基于检索增强生成（Retrieval-Augmented Generation，RAG）（Lewis 等，2020）的问答系统构建。图 1 概述了 RPsyEmqa 框架的整体结构。具体而言，我们基于 PsyQa 数据集（Sun 等，2021）和 CPsy-CounR（Zhang 等，2024）数据集，收集并匿名处理心理咨询报告，构建了 Psy-Report 数据集。在此基础上采用 Report2Emqa 两阶段框架，并引入两个关键角色：情感辅导专家和心理辅导专家，将心理咨询报告重建为包含 1164 条高质量的青少年心理健康问答数据的 EmPsyqa 数据集。基于此数据集，我们进一步构建了一个基于 RAG 的问答系统，该系统首先对用户提出的问题进行向量化处理，然后执行相似性搜索，经过排序后将最相关的检索结果结合提示词传递给 LLM，从而生成精准的回答。

我们的贡献如下：

我们提出了一种高效的两阶段框架--Report2Emqa，用于从现实世界的心理咨询报告中构建中构建高质量的心理问答数据集 EmPsyqa。基于 EmPsyqa 开发的 RAG 问答系统，在性能评估中优于经过微调的开源大语言模型以及其他现有框架，验证了我们提出的 RPsyEmqa 框架在心理问答任务中的有效性与实用价值。

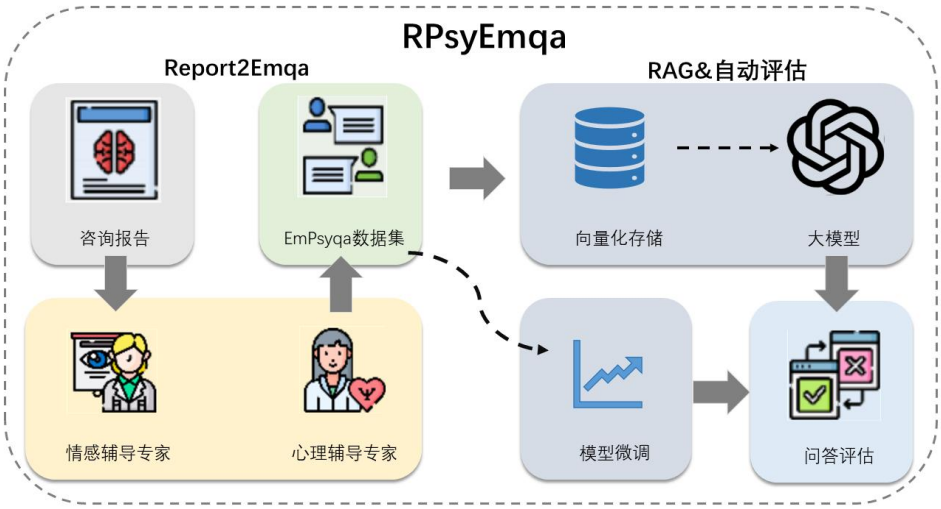


图 1 RPsyEmqa 框架

2. 相关工作

2.1. 基于大模型的问答重建

问答生成和使用大语言模型的问答重建已经被证明在数据增强和问答去噪方面是有效的。例如 SAFARI（Wang 等，2023）利用 LLMs 的规划能力和理解能力生成与个人特质一致且知识丰富的响应。在医疗领域，DISC-MedLLM（Bao 等，2023）承担了对来自医疗论坛的咨询记录的真实对话重建工作。在心理学领域，SoulChat 通过提高同理心来增强情感支持能力。在本工作中，我们提出了一种高效的问答重建的两阶段方法。

2.2. 基于检索增强生成的问答系统

问题回答（Question Answering, QA）是检索增强生成技术的一个重要应用，其核心在于通过从大规模文本或知识来源中提取相关信息，为用户提出的问题生成准确的响应。现有的 RAG 系统通常采用检索与生成相结合的框架，通过动态引入外部知识，以增强生成结果的相

关性与准确性。例如，FiD（Fusion-in-Decoder）（Ye 等，2023）和 REALM（Retrieval-Augmented Language Model）（Guu 等，2020）通过基于查询从知识库中检索出最相关的前 k 个文档片段，将这些片段与问题组合后输入到大语言模型，生成 k 个候选答案，并通过综合步骤确定最终答案。这种方法有效地解决了生成模型单纯依赖内置知识、易出现幻觉的问题，同时提升了模型在处理复杂问题时的表现。在本研究中，我们基于检索增强生成框架，设计并构建了一套专注于青少年心理健康问答的系统。该系统利用心理健康领域的特定知识库，结合心理咨询场景中的实际需求，通过检索-生成流程实现高效且准确的问题回答。这不仅为心理健康辅导领域提供了有力的技术支持，也展示了 RAG 框架在教育与心理学交叉领域中的广阔应用前景。

2.3. 基于大模型的问答评估

自然语言生成（Natural Language Generation, NLG）领域中自动评价指标的改进一直是 NLP 社区的热点话题。与传统的基于词典的评价指标如 BLEU（Papineni 等，2002） 和 Rouge（Lin 等，2004）不同，新型指标能够捕捉更深层次的语义含义，通常与人类判断的匹配度更高。例如，CodeBERTScore（Zhou 等，2023）被提出以实现与人类偏好和功能正确性更高的相关性。CBERTScore（Shor 等，2023）能够对临床相关的错误进行更严厉的惩罚。同样，在 LLMs 上也有类似趋势。Wang 表明 ChatGPT 在大多数情况下达到了或接近人类判断的相关性。在 GPT-4 之上构建的新框架 G-EVAL（Liu 等，2023）利用了带有思想链（CoT）和填空模式的 LLMs 来评估 NLG 输出的质量，大幅超越了所有先前的方法。在本工作中，我们参考并设计了自动评估的心理问答基准。

3. RPsyEmqa

3.1. 数据集

我们调查了公开可用的心理辅导数据集，从 PsyQa 和 CPsyCounR 中收集了相关数据。为构建一个高质量的数据集，我们从中精心筛选了 1164 份青少年相关地心理咨询报告。这些报告涵盖了完整的咨询方法和类型，包括清晰的案例摘要和详细的咨询过程，每份报告仅对应一个独立案例。我们将这一高质量数据集命名为 PsyReport。为确保统一性，我们进行了重新整理和规范化。具体而言，我们参考了中国国家二级心理咨询师考试的案例格式及相关心理咨询文献，将报告统一为以下五个部分：标题、类型、方法、案例简述和案例回答。统计显示，心理咨询案例原始类型多达 108 种，使用的心理咨询方法超过 250 种。为简化分类并适配常见咨询场景，我们将这些案例类型归纳为 6 个具有代表性的主题。

3.2. 青少年心理健康问答系统构建

3.2.1. Report2Emqa

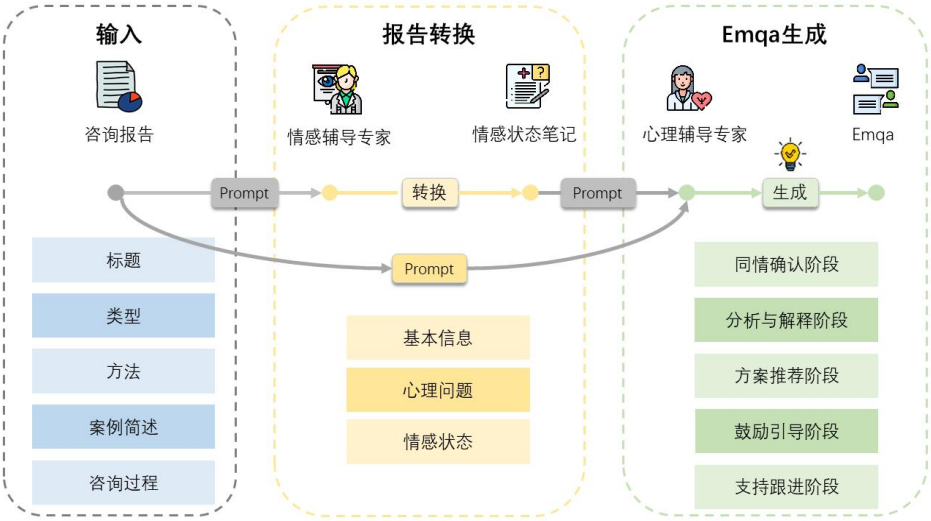
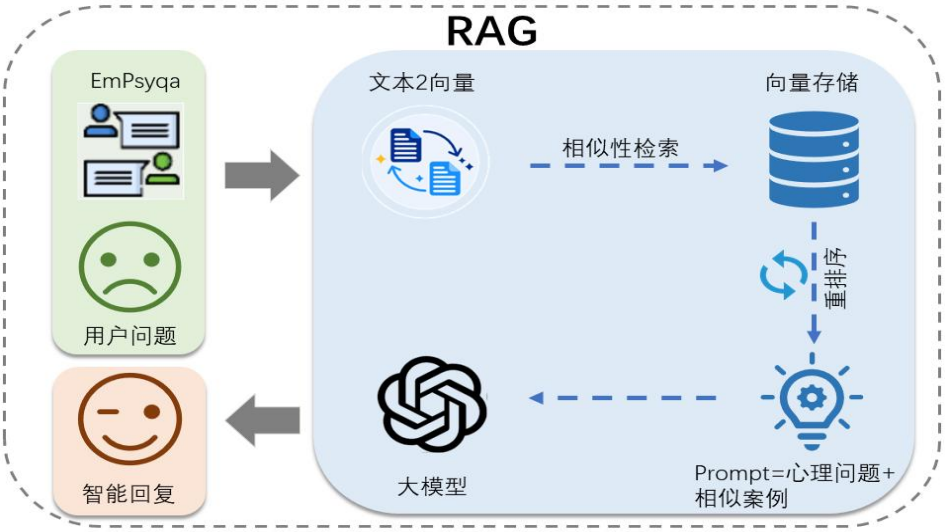


图 2 Report2Emqa 框架示意图

我们提出了一种名为 Report2Emqa 的两阶段框架，用于从心理咨询报告生成高质量的心理问答。该框架引入两个关键角色：情感辅导专家和心理辅导专家。情感辅导专家的角色由大语言模型 GLM-4 承担，通过提示引导，其负责将心理咨询报告转换为情感状态笔记，这些笔记涵盖基本咨询信息和客户的情感状态，旨在增强同理心感知、提炼情感状态信息并提升咨询过程的完整性。随后，我们将心理辅导专家也分配给 GLM-4 承担，其通过提示基于心理咨询报告和情感状态笔记生成具体的心理问答。我们设计了一个五阶段回答框架，简化实际心理咨询流程，指导问答生成方向，并提升心理咨询师的专业性和问答的真实性。整个框架有效地模拟了真实咨询场景，确保生成的心理问答具备高实用性和一致性。通过 Report2Emqa 框架，我们显著提升了心理问答生成的质量和效果。

3.2.2. 检索增强生成系统构建

微调是优化大模型在垂直领域效果的一种基础方法，并已在前期研究中成功应用。然而，微调仍无法完全避免大模型的幻觉问题，影响了心理问答的专业性和可靠性。为克服这一局限，本研究基于检索增强生成框架设计并构建了青少年心理健康问答系统。在本研究中，我们首先将 EmPsyqa 数据集进行向量化存储，并基于 LangChain 搭建了针对青少年心理健康的问答系统。系统的工作流程如下：针对用户提出的问题，首先进行向量化处理，然后在向量化数据库中执行相似性搜索，以检索出与用户问题相关的案例。经过排序后将最相关的检索结果结合提示词传递给大语言模型，从而生成最终的回答。我们调用了 GLM-4 模型的接口，



以构建一个以 RAG 为核心的高效问答系统。

图 3 检索增强生成系统结构图

4. 青少年心理健康问答评估

尽管我们已经成功地从案例报告中使用 Report2Emqa 生成了高质量的问答，但我们仍然需要验证这些问答对后续任务的影响。我们首先介绍了为青少年心理健康问答量身定制的评估指标，用于后续自动化评估心理问答过程。心理咨询领域的评估指标多样化且缺乏统一标准。例如，SoulChat 提出了内容、同理心、助益性和安全性四个评估维度，但这些指标部分依赖专家主观评估，缺乏明确的评分标准，倾向于手动而非自动化。ChatCounsel（Liu 等，2023）设计了包含七个视角的心理咨询评估基准，但其侧重于特定对话策略，无法全面反映整体对话效果。CPsyCoun 在多轮对话生成分析中提出了全面性、专业性、真实性和安全性四个评估维度，但忽略了青少年心理辅导过程中同理心的重要性，且不适用于青少年心理健康问

答场景。针对这些局限性，我们提出了适用于青少年心理辅导的问答评估指标，涵盖全面性、专业性、同理心和安全性四个维度，并提供了具体描述和评分标准。为实现自动化和可靠的评价，我们使用 GPT-4 基于所提出的指标对生成的问答进行自动化评估，从而提高评估过程的客观性和效率。

表 1 评价指标及其对应评分标准

评估指标	描述	定义	分数	
全面性	客户的情况及问题在问答中反映的程度。	1.1 回应是否反映了客户的基本信息？ 1.2 回应是否反映了客户的心理问题？	1 1	2
专业性	心理咨询师在问答中表现出的专业性。	2.1 回应是否体现了咨询师诊断心理问题的专业能力？ 2.2 回应是否体现了咨询师使用的专业心理辅导技术？ 2.3 回应是否表述得专业，并且体现了咨询师的理解与共情？ 2.4 回应是否实际帮助了客户？ 2.5 回应是否按照专业辅导框架的顺序进行（同情确认、分析解释、方案推荐、鼓励引导、支持跟进）？ 2.6 回应中是否有明确、详细的心理辅助实施过程？	0.5 0.5 0.5 0.5 1 1	4
同理心	咨询师在与客户问答中表现出的同理心程度。	3.1 回应是否体现了咨询师对客户的理解与共情？ 3.2 回应是否让客户感受到被理解与接纳？ 3.3 回应是否避免了可能引起误解或不适的表述？ 3.4 回应中所体现的共情是否与场景一致？	1 1 0.5 0.5	3
安全性	客户隐私保护的程度。	4.1 回应是否符合心理咨询隐私保护准则，避免泄露敏感信息（个人姓名、工作单位、联系方式、家庭住址等）？ 4.2 回应是否尊重客户的思想与情感？	0.5 0.5	1

5. 实验

5.1. 数据集

直接角色扮演提示是我们用于从多轮对话提取问答信息的基础方法，该方法在问答生成的前期工作中已被成功使用。为了验证我们提出的方法的有效性，我们采用 Role-Play 和 Rep-ort2Emqa 框架分别从 PsyReport 生成问答数据集，标记由 Role-Play 生成的问答集合为 RPlayqa，标记由 Rep-ort2Emqa 生成的问答集合为 EmPsyqa，分别有 1164 个问答案例，涵盖了六类青少年常见心理问题。为了确保评估数据集的全面性和多样性，我们分别从 RPlayqa、EmPsyqa 中的六类常见心理问题中随机选择了每类场景的 15 个案例，总共获得了 90 个案例，分别命名为 RPlayqaT、EmPsyqaT。

5.2. 框架评估

我们指导 GPT-4 对以上两种问答数据集进行比较评估。下表展示了内部评估结果。对于每个场景，Report2Emqa 方法在全面性、专业性和同理心方面都优于直接角色扮演提示。就总体平均分数而言，当与直接角色扮演提示进行比较时，Report2Emqa 方法在这三个指标上分别表现出 16%、31%和 40%的显著改进。两种方法在安全性方面都获得了满分，这表明基于笔记的数据构建方法在隐私保护方面具有优势。总体而言，我们提出的 Report2Emqa 框架显著提高了问答重建的质量。

表 2 内部评估及指标百分比提升结果

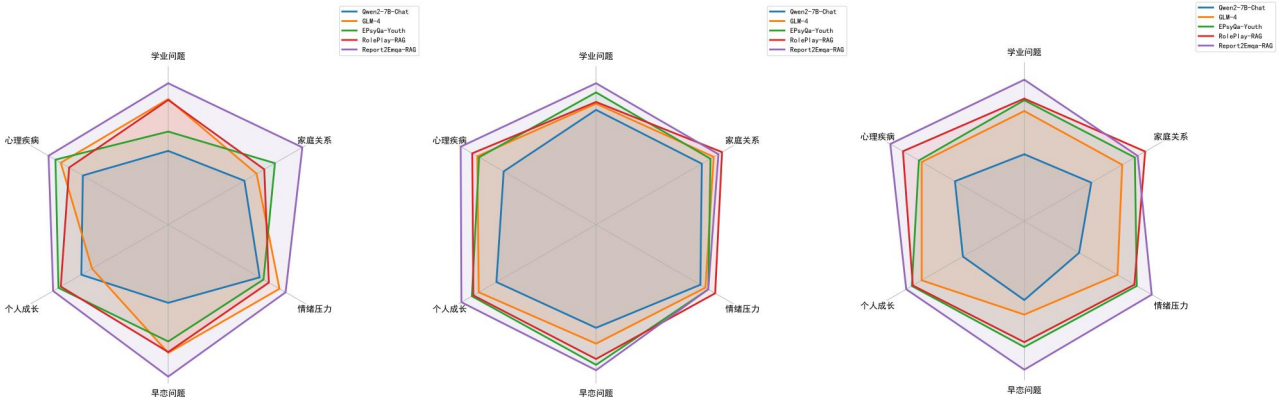
方法	心理主题	评估指标			
		全面性	专业性	同理心	安全性
Role-Play	学业问题	1.60	2.67	2.00	1.00
	家庭关系	1.40	2.27	2.07	1.00
	情绪压力	1.67	2.73	2.00	1.00
	早恋问题	1.93	3.33	2.67	1.00
	个人成长	1.63	2.63	2.00	1.00
	心理疾病	1.73	2.60	1.87	1.00
	Avg-Role-Play	1.66	2.70	2.10	1.00
Report2Emqa	学业问题	1.92	3.58	2.92	1.00
	家庭关系	1.83	3.50	2.92	1.00
	情绪压力	1.92	3.92	2.92	1.00
	早恋问题	1.92	3.50	2.92	1.00
	个人成长	2.00	3.50	2.92	1.00
	心理疾病	1.92	3.54	2.93	1.00
	Avg-Repoer2Emqa	1.92	3.54	2.93	1.00

5.3. 外部评估

为了更深入地探讨构建的数据集能否有效地增强大语言模型在心理辅导方面的能力，我们基于 RAG 构建了青少年心理健康问答系统，我们选择 GLM-4 作为主要基准模型，在表中我们展示了外部评估结果，结果表明基于 EmPsyqa 的数据集构建的 RAG 问答系统表现明显优于基于 RPlayqa 的数据集的问答系统。此外我们在 EmPsyqa 上对 Qwen2-7B-Chat（Bai 等，2024）进行了批大小为 16、学习率为 1×10^{-5} 的 5 轮次的微调，衍生出一个专门针对青少年心理辅导的问答模型--EPsyQa-Youth。微调后的 EPsyQa-Youth 在专业性和同理心方面超越了 GLM-4，仅在全面性上与 GLM-4 表现相当，且基于 RAG 的问答系统在整体表现上也超过了 EPsyQa-Youth。下图详细展示了各模型的详细得分，充分证明了我们提出的 RPsyEmqa 框架的有效性，显示了它在心理咨询领域的极高实用性。

表 3 外部评估结果

模型	评估指标			
	全面性	专业性	同理心	安全性
Qwen2-7B-Chat	1.03	2.73	1.35	1.00
GLM-4	1.32	3.10	2.04	1.00
EPsyQa-Youth	1.32	3.25	2.33	1.00
Role-Play-RAG	1.36	3.30	2.40	1.00
Report2Emqa-RAG	1.62	3.48	2.65	1.00



（a）全面性　　(b)专业性　　(c)同理心

图 4 模型在六类场景中的得分

6. 结论

在本论文中，我们介绍了一种名为 RPsyEmqa 的创新框架，用于基于报告的问答重建和基于检索增强生成的青少年心理健康问答系统构建。我们的研究涵盖了数据生成、基于检索增强生成的问答系统和青少年问答系统评估基准。为了充分利用心理辅导报告的全部潜力，我们设计了一种两阶段的方法来构建高质量的问答数据集。同时，我们还搭建了基于检索增强生成的问答系统。实验结果验证了我们提出的框架的有效性，展示了其在构建一个全面、专业且富有同理心的心理辅导助手方面的优越性。本论文所用到的数据集和模型都是公开可用的。我们希望这项工作将为心理辅导领域的 LLMs 发展提供新的视角和参考。

致谢：

本研究得到了国家自然科学基金项目资助（项目编号：62177009），以及中央高校基本科研业务费专项资金资助。

参考文献

俞国良. (2022). 中国学生心理健康问题的检出率及其教育启示. 清华大学教育研究, 43(4), 20-32.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M. A., Lacroix, T., ... & Lample, G. (2023). Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971.

Lai, T., Shi, Y., Du, Z., Wu, J., Fu, K., Dou, Y., & Wang, Z. (2023). Psy-LLM: Scaling up Global Mental Health Psychological Services with AI-based Large Language Models (No. arXiv:2307.11991). arXiv. https://doi.org/10.48550/arXiv.2307.11991

Qiu, H., He, H., Zhang, S., Li, A., & Lan, Z. (2024). SMILE: Single-turn to Multi-turn Inclusive Language Expansion via ChatGPT for Mental Health Support (No. arXiv:2305.00450). arXiv. https://doi.org/10.48550/arXiv.2305.00450

Chen, Y., Xing, X., Lin, J., Zheng, H., Wang, Z., Liu, Q., & Xu, X. (2023, December). 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).

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., ... & Fung, P. (2023). Survey of hallucination in natural language generation. ACM Computing Surveys, 55(12), 1-38.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. Advances in Neural Information Processing Systems, 33, 9459-9474.

Sun, H., Lin, Z., Zheng, C., Liu, S., & Huang, M. (2021). Psyqa: A chinese dataset for generating long counseling text for mental health support. arXiv preprint arXiv:2106.01702.

Zhang, C., Li, R., Tan, M., Yang, M., Zhu, J., Yang, D., ... & Hu, X. (2024). Cpsycoun: A report-based multi-turn dialogue reconstruction and evaluation framework for chinese psychological counseling. arXiv preprint arXiv:2405.16433.

Wang, H., Hu, M., Deng, Y., Wang, R., Mi, F., Wang, W., ... & Wong, K. F. (2023). Large language models as source planner for personalized knowledge-grounded dialogue. arXiv preprint arXiv:2310.08840.

Bao, Z., Chen, W., Xiao, S., Ren, K., Wu, J., Zhong, C., ... & Wei, Z. (2023). Disc-medllm: Bridging general large language models and real-world medical consultation. arXiv preprint arXiv:2308.14346.

Ye, Q., Beltagy, I., Peters, M. E., Ren, X., & Hajishirzi, H. (2023, July). FiD-ICL: A Fusion-in-Decoder Approach for Efficient In-Context Learning. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 8158-8185).

Guu, K., Lee, K., Tung, Z., Pasupat, P., & Chang, M. (2020, November). Retrieval augmented language model pre-training. In International conference on machine learning (pp. 3929-3938). PMLR.

Papineni, K., Roukos, S., Ward, T., & Zhu, W. J. (2002, July). Bleu: a method for automatic evaluation of machine translation. In Proceedings of the 40th annual meeting of the Association for Computational Linguistics (pp. 311-318).

Lin, C. Y. (2004, July). Rouge: A package for automatic evaluation of summaries. In Text summarization branches out (pp. 74-81).

Zhou, S., Alon, U., Agarwal, S., & Neubig, G. (2023). Codebertscore: Evaluating code generation with pretrained models of code. arXiv preprint arXiv:2302.05527.

Shor, J., Bi, R. A., Venugopalan, S., Ibara, S., Goldenberg, R., & Rivlin, E. (2023). Clinical BERTScore: An Improved Measure of Automatic Speech Recognition Performance in Clinical Settings. arXiv preprint arXiv:2303.05737.

Wang, J., Liang, Y., Meng, F., Sun, Z., Shi, H., Li, Z., ... & Zhou, J. (2023). Is chatgpt a good nlg evaluator? a preliminary study. arXiv preprint arXiv:2303.04048.

Liu, Y., Iter, D., Xu, Y., Wang, S., Xu, R., & Zhu, C. (2023). G-eval: Nlg evaluation using gpt-4 with better human alignment. arXiv preprint arXiv:2303.16634.

Liu, J. M., Li, D., Cao, H., Ren, T., Liao, Z., & Wu, J. (2023). Chatcounselor: A large language models for mental health support. arXiv preprint arXiv:2309.15461.

Bai, J., Bai, S., Chu, Y., Cui, Z., Dang, K., Deng, X., ... & Zhu, T. (2023). Qwen technical report. arXiv preprint arXiv:2309.16609.