第 15 卷 第 1 期 集 成 技 术 Vol. 15 No. 1
2026 年 1 月 JOURNAL OF INTEGRATION TECHNOLOGY Jan. 2026

**引文格式：**
金约汗, 谭明环, 杨敏. 大语言模型在心理健康领域的应用综述 [J]. 集成技术, 2026, 15(1): 85-107.
Jin YH, Tan MH, Yang M. Review on the application of large language models in the mental health field [J]. Journal of Integration Technology, 2026, 15(1): 85-107.

# 大语言模型在心理健康领域的应用综述

金约汗[1] 谭明环[2*] 杨 敏[2*]

[1](华南理工大学 计算机科学与工程学院 广州 510006)
[2](中国科学院深圳先进技术研究院 深圳 518055)

**摘 要** 大语言模型在心理健康领域的应用已成为人工智能与临床心理学交叉领域的核心研究方向。本综述从模型特性与实证依据、临床应用场景及技术发展路径 3 个维度，对该领域的研究进展展开系统性梳理。在模型特性与实证依据层面，本文剖析了大语言模型的核心特质，总结了其适配心理症状诊断与心理疾病干预的实证支撑；在临床应用层面，系统归纳了大语言模型在心理疾病诊断、心理状态评估、虚拟心理治疗及临床决策辅助等场景中的实践案例与应用成效；在技术发展层面，重点梳理了面向心理健康领域的数据构建、模型能力增强及专用评估方法等方向的关键进展。最后，明确指出当前研究仍面临诊断结果与临床实践脱节、治疗模拟深度不足、高质量标注数据稀缺及技术临床转化验证欠缺等核心挑战，并对未来临床应用落地与技术创新研究的发展方向进行了展望。

**关键词** 大语言模型；心理健康；症状诊断；治疗干预
**中图分类号** TP391；[R34] **文献标志码** A doi: 10.12146/j.issn.2095-3135.20251102001
CSTR: 32239.14.j.issn.2095-3135.20251102001

# Review on the Application of Large Language Models in the Mental Health Field

JIN Yuehan[1] TAN Minghuan[2*] YANG Min[2*]

[1]( *School of Computer Science and Engineering*, *South China University of Technology*, *Guangzhou* 510006, *China* )
[2]( *Shenzhen Institutes of Advanced Technology*, *Chinese Academy of Sciences*, *Shenzhen* 518055, *China* )
[*]*Corresponding Authors*: *mh.tan@siat.ac.cn*; *min.yang@siat.ac.cn*

**Abstract** The application of large language models (LLMs) in the mental health field has emerged as a core research direction at the intersection of artificial intelligence (AI) and clinical psychology. This review systematically synthesizes the research progress in this domain from 3 dimensions: model characteristics and

**收稿日期**：2025-11-02 **修回日期**：2025-12-03
**基金项目**：国家自然科学基金项目 (62406314)；广东省基础与应用基础研究基金项目 (2023A1515110496)；中国博士后科学基金项目 (2023M733654)
**作者简介**：金约汗，本科生，研究方向为多模态情感识别； 谭明环 (通讯作者)，助理研究员，研究方向为大模型的微调和对齐、领域大模型的训练和评测、科学智能，E-mail：mh.tan@siat.ac.cn； 杨敏 (通讯作者)，研究员，研究方向为人工智能、自然语言处理，E-mail：min.yang@siat.ac.cn。

empirical evidence, clinical application scenarios, and technological development pathways. At the level of model characteristics and empirical evidence, this paper analyzes the core traits of LLMs and summarizes the empirical support for their applicability in psychological symptom diagnosis and mental illness intervention. In terms of clinical applications, it systematically summarizes the practical cases and application effectiveness of LLMs in scenarios such as mental illness diagnosis, psychological state assessment, virtual psychotherapy, and clinical decision support. From the perspective of technological development, it focuses on sorting out the key advances in directions including data construction, model capability enhancement, and specialized evaluation methods tailored to the mental health field. Finally, the review explicitly points out the core challenges currently facing the research field—such as the disconnect between diagnostic outcomes and clinical practice, insufficient depth of therapeutic simulation, scarcity of high-quality annotated data, and the lack of clinical translation and validation of technologies—and presents an outlook on the future development directions for clinical application implementation and technological innovation research.

**Keywords** large language models; mental health; symptom diagnosis; therapeutic intervention

**Funding** This work is supported by National Natural Science Foundation of China (62406314), Guangdong Basic and Applied Basic Research Foundation (2023A1515110496), China Postdoctoral Science Foundation (2023M733654)

## 1 引 言

随着人工智能 (artificial intelligence，AI) 技术的飞速发展，尤其是以 Transformer 结构为代表的大语言模型 (large language models，LLMs，以下简称 “大模型”) 的兴起，人工智能技术在各行各业展现出前所未有的通用性与智能化程度。从医疗诊断到客户服务，从金融风控到法律咨询，AI 系统正深刻重塑各行业生态格局。

心理健康问题已成为全球公共卫生领域的突出挑战。据统计，精神障碍已是致残的主要原因之一，抑郁症、焦虑症等常见心理问题持续对社会和经济发展造成沉重负担[1]。传统心理治疗存在资源匮乏、分布不均以及成本高昂等问题，使得社会亟须高效、可扩展且具备智能交互特性的新型辅助工具。大模型凭借其强大的自然语言理解与生成能力，能够分析海量文本数据、模拟人际对话并生成针对性回应，为缓解心理健康资源失衡的现状提供了新的潜在路径。当前，许多研究已经在探索将大模型应用在心理学症状诊断[2]和疾病治疗[3]任务上。然而，大模型仍然存在系统性不足。一方面，大模型可靠性与内容一致性仍然无法得到保证[4]；另一方面，大模型 “黑箱” 特性阻碍其在心理健康这类强调可解释性与伦理性的领域普及，且大模型在该领域的应用缺乏统一的伦理准则和隐私保护机制，容易引发误诊、建议偏差或信息泄露等风险[5]。因此亟须开展系统性综述，从多维视角审视大模型在心理健康中的角色，探讨大模型在心理健康领域应用的现状、挑战和未来发展方向。

已有综述往往从心理学视角和计算机技术视角展开。从心理学视角出发的相关综述主要聚焦大语言模型在心理健康领域应用的潜力与风险，并倾向于按具体应用场景展开分类梳理。Guo 等[6]和 Hua 等[7]梳理了大模型在心理健康临床的具体应用，涉及心理支持、心理咨询和心理治疗

等方面；Omar 等[8]则着重梳理了大模型在精神病学领域的应用；Lawrence 等[9]深入探讨了现有应用背后潜藏的伦理问题；Linardon 等[10]通过对全球各国心理健康研究人员的调研，总结提炼了大模型应用过程中的关键挑战；籍欣萌等[11]系统综述了中文心理健康大模型的应用；涂翠平等[12]则重点分析了大模型在高校心理健康领域的应用场景；王慧等[13]综述了大模型在心理干预中的应用效果、挑战及前景。从计算机技术视角出发的综述聚焦大语言模型相关技术的演进历程，并倾向于依据大模型的任务定义对现有技术体系进行分类梳理。Yuan 等[14]系统梳理了心理健康聊天机器人的技术发展脉络；Na 等[15]综述了大模型在心理治疗中的任务分类及对应的技术脉络；陈旭日等[16]和陈元乐等[17]则总结了大模型在抑郁症检测和识别任务中的技术发展动态。

以上两个视角的综述分别反映了心理学领域 (特别是临床工作) 和计算机领域 (特别是技术研发) 的关注重点，但疏于对两个视角进行一个更综合的梳理，也缺少分析大模型能应用在心理健康领域的实证依据。因此本文从大模型的实证特性、已有的大模型临床场景和大模型技术发展 3 个方向展开综述。首先，介绍大模型可运用在心理健康领域的实证依据；其次，梳理已有大模型技术在心理健康领域的应用现状，归纳当前核心技术难点；最后，回顾大模型在心理领域应用的技术发展史，以及在前述难点上的突破，并为未来临床集成提出建设性建议。本文旨在提供心理学与计算机技术的交叉视角，为大模型在心理健康领域的应用提供支撑。

## 2　大模型可应用在心理健康领域的实证依据

心理健康临床实践的核心任务主要集中在两个方面：一是对个体症状的识别与诊断，二是对精神障碍或相关心理疾病的干预与治疗，如图 1 所示。随着人工智能技术的发展，尤其是大模型在自然语言处理、情感理解和生成式对话等方向上的突破，越来越多的研究表明其在上述两大临床任务中均展现出广阔的应用前景。例如，大模型能在症状识别环节中辅助完成临床量表的自动化解析与语义理解，提高诊断的客观性与一致性；同时，在治疗与干预过程中，其生成的支持

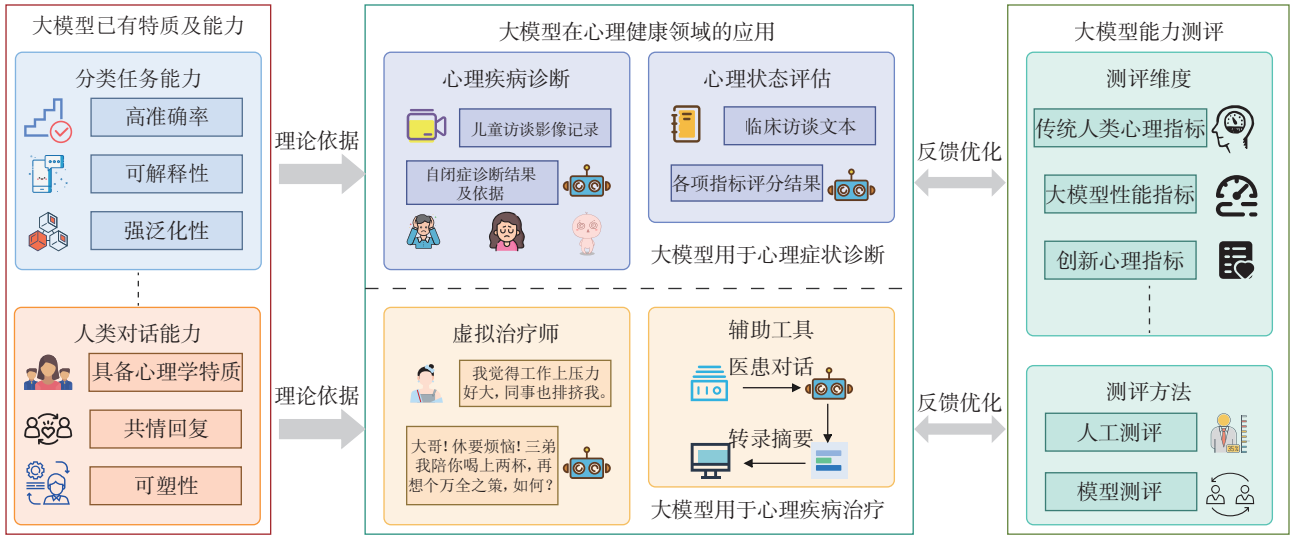


图 1　大模型在心理健康领域的临床应用场景

**Fig. 1　Clinical applications scenarios of large language models in mental health**

性对话与心理教育内容也逐渐显示出潜在的疗效价值。基于此，本节将系统梳理当前相关的实证结论，旨在为大模型在心理健康领域的临床应用提供一定发展支撑。

### 2.1 大模型应用于心理症状诊断的实证依据

心理症状诊断通常涵盖症状筛查、自杀风险评估、情绪状态识别和认知模式分析等环节。传统诊断依赖临床量表与心理学专业人员的面谈评估，既耗时又受制于专业资源的有限性。随着自然语言处理与大模型的发展，基于语言行为和语义特征的自动化诊断方法逐渐成为可能。大模型在分类任务中展现出的能力为其应用于心理症状诊断提供了实证依据。

#### 2.1.1 大模型诊断准确性

大模型应用于诊断任务的首要前提是足够的准确性，不足则会降低大模型的临床价值，并带来误判风险。已有研究表明，大模型在抑郁症文本识别和自杀倾向预测等任务中展现出较高的准确率。因此，从实证上看，大模型可通过学习海量语言样本捕捉心理症状的潜在模式，从而提高诊断的一致性与可扩展性。例如，Jiang 等[18]首次将大语言模型应用于基于真实 ADOS-2 临床场景的自闭症谱系障碍 (ASD) 诊断，提出了 ADOS-Copilot 框架。该研究在包含 28 个临床评估音频样本的数据集上进行了测试。大模型在二元分类 (自闭症谱系与非谱系) 任务中最高达到了 82.14% 的准确率和 81.79% 的 F1 分数；在三元分类 (自闭症、自闭症谱系、非谱系) 任务中达到了 75.00% 的准确率和 78.37% 的 F1 分数。Galatzer-levy 等[19]则聚焦大模型对临床量表分数的预测能力，在包含 145 例抑郁症患者和 115 例创伤后应激障碍患者的临床访谈数据上，Med-PaLM 2[①] 对抑郁症 (PHQ-8 量表) 的评分与人类评估者无显著差异 ($t(1,114)=1.20$；$p=0.23$)[②]，分类准确率在 80%～84%，敏感度为 0.75，特异度为 0.82。

#### 2.1.2 大模型决策可解释性与多任务泛化性

心理健康诊断不仅追求结果的准确，还强调过程的可解释性。临床心理学家需要理解模型为何得出某种判断，以便结合患者的个体背景进行干预。大模型虽然常被批评为“黑箱”，但其生成式语言能力为解释诊断依据提供了新途径，例如，通过生成可读性强的推理路径，展示模型的注意力分布与关联线索。这种“可解释诊断”理论上有助于增强临床信任度，降低误用风险。例如，Lan 等[20]提出了 DORIS 系统，利用大模型分析构建用户的历史情绪轨迹，实现了大模型在抑郁症诊断上的决策可解释性。除此之外，心理症状诊断常涉及多任务与多模态输入，如同时识别情绪状态、评估风险水平和分析认知偏差。传统模型往往需针对单一任务分别训练，而大模型具备显著的迁移学习与多任务泛化能力，能在有限监督下快速适应不同诊断任务。这种能力为其在心理健康领域的综合诊断应用提供了支撑，也为跨领域知识共享 (如从一般心理评估迁移到临床干预) 创造了可能。

### 2.2 大模型应用于心理疾病治疗的实证依据

心理疾病治疗包括认知行为疗法 (cognitive behavioral therapy，CBT) 和精神动力学疗法等传统手段，需要在治疗过程中建立治疗关系、展现共情并对受治疗者提供情绪回应。大模型作为交互式智能体，其价值在于能模拟部分治疗性对话过程，为心理干预提供辅助支持。大模型在人类对话中表现的特质为其应用于心理治疗提供了实证依据。

#### 2.2.1 大模型存在心理学特质

心理疾病治疗强调稳定的“治疗者人格”，包括一致的沟通风格、态度与价值导向。近年

①Google Research 基于 PaLM 2 架构开发的医学领域大语言模型。

②$t$ 值是检验模型评分与人类评分两组数据均值差异的统计量，反映差异的幅度；$p$ 值是差异显著性的判断标准。

来，大量研究表明，大语言模型在经过大规模训练后确实能表现出相对稳定的人格倾向与语言风格，展现出一定心理学特质，为大模型作为心理治疗师提供了支撑。

一类依照心理学现有量表对大模型进行评估的研究表明，现有各类大模型均呈现趋同的心理学特质。Jiang 等[21]基于大五人格理论提出 PersonaLLM 框架，证明大模型在设定人格条件下能稳定生成符合特定人格的语言模式，且在人类评估中具备较高的可感知度。Li 等[22]研究了大五人格等量表，发现已有大模型都呈现不同程度的消极心理学特质，但通过微调的方式可以降低这类心理毒性，以呈现更积极的心理学特质。Huang 等[23]同样基于大五人格量表对多种模型开展跨语言与跨情境对比，发现各类模型存在一致性响应，并指出 ChatGPT 在不同语言和指令条件下均稳定呈现 ENFJ 型人格，未发生显著人格漂移。

另一类研究提出了额外的测评标准，以探索大模型在更多心理学特质上的表现。Huang 等[24]提出了名为 PsychoBench 的大模型心理测评框架，系统评估了主流模型在人格、动机与情绪等心理特征上的表现，发现部分大模型比正常人类的负面心理学特质更多，并有轻微的男性化倾向；更进一步，GPT-4 已在情感智商指标中 (Mean＝121.8，SD＝12.0)[①]展示出与人类相近的测试表现 (Mean＝124.8，SD＝16.5)。

2.2.2 大模型在实证中表现情绪感知与共情回应的能力

情绪识别与共情回应是心理干预的核心。大模型能通过分析语义线索识别用户情绪，并生成带有安慰、鼓励或理解色彩的回应。这种“情绪感知—回应”机制为其在心理治疗中的应用提供了一定支撑。但需注意的是，大模型的“共情能力”与人类临床工作者的共情存在本质区别。前者本质是基于大规模文本数据的语义匹配与模式复刻，依赖语言线索完成情绪识别与回应生成[25-26]；而后者建立在真实互动中的多重非语言线索 (如面部表情、语音语调、肢体动作)、临床经验及对患者个体背景的深度理解之上[27]。现有研究多基于文本或标准化量表评估模型共情表现，缺乏对真实临床场景中多模态交互的考量，因此其接近人类水平的结论仅适用于特定文本任务，不能直接等同于临床场景中的有效共情。

大模型在文本情绪识别任务中表现出可验证的情绪感知能力，且能生成具有共情特征的回应，但与人类水平仍存在显著差异。Schaaff 等[25]利用对话任务和心理学问卷系统评估 GPT-3.5 在多维度共情表现上的能力，发现其在大部分情境中能准确识别情绪，并给出适宜回应。Elyoseph 等[28]则利用情绪意识水平量表系统测试了 ChatGPT 在 20 个情境下的情绪觉察能力，发现其能生成高度契合情境的情绪觉察回应，且性能随交互有所提升。Huang 等[29]基于心理学评估理论构建多场景数据集，并提出 EmotionBench 测试框架，发现大模型在特定情境下能给出恰当回应，展现出基本的情绪反应能力，但与人类的情绪行为存在一定偏差，无法在相似的情境之间建立联系。

随着技术的成熟和参数量的提升，大模型的情绪感知与共情能力也随之增强。Schlegel 等[26]在最新的研究中使用 5 项标准化情绪智能测验题目评估 GPT-4、ChatGPT-o1 和 Gemini 等模型，结果显示，模型平均准确率达 81%，与人类样本 (56%) 具备可比性；但在部分子维度仍低于人类临床工作者。

2.2.3 大模型心理学特质的可塑性

心理治疗过程往往需要根据患者的个体特征进行个性化调整，这对治疗者的心理学特质、语

①Mean 代表平均值，SD 代表标准差。

言风格和干预取向都提出了较高要求。大模型在心理学特质方面具有显著的可塑性，可以通过提示工程、指令微调、人类反馈机制或参数高效微调等方法，在保持通用性能的同时，灵活适配不同的治疗取向，从而弥补传统治疗中个性化不足与资源短缺的困境。

一类研究基于具体的社交角色或者场景数据，从数据驱动的角度训练大模型，以实现人格调控。Zhou 等[30]收集了一个包含不同类别和行为角色的大规模中文语料库，并在此基础上提出了 CharacterGLM 模型，实现了可扩展的多场景角色对话定制。Li 等[31]提出了 BIG5-CHAT 框架，基于大规模人格对话数据集训练模型，使大模型在人格测评中的表现更接近人类，并揭示了特定心理学特质与推理任务表现之间的潜在耦合关系。Li 等[32]提出了无监督构建的个性化词典方法，在解码阶段动态调整预测词概率，无须额外微调即可有效改变模型的人格输出模式。

另一类研究基于现有心理学理论对人格类型进行划分，通过提示词工程等方式微调大模型，以调控人格。Tu 等[33]基于 MBTI 人格理论提出了 CharacterChat 系统，结合人格建模与动态记忆优化，能根据用户人格特征匹配最合适的虚拟支持者。Jiang 等[34]首次系统化验证了基于大五人格理论刻画大模型行为特征的可行性，并通过提示词方法实现模型人格的可控性。Jain 等[35]则通过参数高效微调的方法操控大五人格特质。在表情生成实验中，95% 的人类评委认为大模型生成的表情符号与目标人格应当生成的表情符号一致。

## 3 大模型在心理健康临床场景的应用

心理健康临床实践的最终落脚点在于诊断与治疗。前者主要解决“是什么”的问题，即识别个体是否存在特定的心理障碍或症状，并评估其严重程度；后者则着眼于“怎么办”，即通过干预手段帮助个体缓解或恢复健康状态。随着大模型的迅速发展，其在这两个领域已逐步开始投入临床试验。本节将通过具体案例梳理大模型已在哪些临床领域落地应用，并根据现有临床反馈，归纳大模型在应用中暴露的技术缺陷，以期在后续部分进一步归纳整理相关技术。

### 3.1 大模型在心理症状诊断的应用

心理症状诊断主要可以分为两类，一类是心理疾病诊断，另一类是心理状态评估。前者往往被定义为二分类任务，即判断疾病的有无；后者则常常被定义为多分类任务，涉及多个级别状态的评估。传统深度学习模型很早就在这一领域展开了临床应用，但诊断依据往往只有特定的量表和患者自评。在大模型逐渐兴起后，各类谈话内容和患者的多模态行为信息也可以作为模型的输入，使得大模型在这类任务上有着更高的准确率和可解释性，从而逐渐进入临床的应用当中。

#### 3.1.1 疾病诊断

心理疾病诊断是判定精神障碍和疾病的存在与类型。诊断方式通常包括结构化访谈、量表测评和临床观察。

一类研究主要通过引入临床数据集的方式评估大模型诊断疾病的能力，拓宽诊断疾病的范围。例如，临床中，由于抑郁症患者往往同时患有其他心理疾病，因此 Hengle 等[2]提出了首个面向抑郁—焦虑共病的基准数据集 ANGST，评估了多种大模型的共病诊断性能，发现大模型在单疾病诊断时表现优异，在抑郁、焦虑二分类任务中的 F1 分数分别达到 68.4% 和 78.9%，但在抑郁—焦虑共病的多类别分类任务中，所有模型的最佳 F1 分数均未超过 72%。Lho 等[36]聚焦于自杀风险的识别，利用上千名精神科患者的叙述数据评估大模型与文本嵌入模型，发现两类模型在检测临床相关的抑郁症状和高自杀风险方面均表现良好，在基于自我概念叙述的情境下效果最佳，显示出其在临床风险筛查中的潜力。

另一类研究聚焦于已有临床诊断的缺口，通过各类方法评估或增强模型诊断的性能。Jiang 等[18]最早将大模型用于 ASD 诊断，提出了基于提示词工程的 ADOS-Copilot 框架，二元分类任务中，F1 分数达到 81.79%，三元分类任务中，F1 分数为 78.37%，均优于单一大模型或规则模型。针对传统深度学习模型用于抑郁症诊断时缺乏可解释性的问题，Lan 等[20]提出了 DORIS 系统，结合情绪强度分析构建用户的历史情绪轨迹，再与传统分类器结合，以实现高准确率与可解释性。Xu 等[37]系统评估了各种大模型在中文语境下进行抑郁症检测和自杀检测的准确性，首次为理解大模型在中文语境下心理健康领域的应用能力提供了新视角。

总体而言，利用大模型诊断心理疾病虽然在临床应用的深度和广度已取得一定进展，但仍然存在部分问题，如过于依赖真实世界的完备数据和诊断依据过于依赖文本。

#### 3.1.2 状态评估

心理状态评估涉及个体情绪、认知以及行为模式的量化与分级，用于判定障碍的严重程度及发展趋势，相较于疾病诊断表现为更复杂的分类任务。

现有研究主要通过构建临床数据集的方式探究大模型在状态评估方面的能力。例如，Tu 等[38]构建了由临床医生访谈组成的数据集，并最早提出利用大模型评估自动化创伤后应激障碍(post-traumatic stress disorder，PTSD)。Galatzer-levy 等[19]在临床访谈的基础上进一步引入了病例描述，并作为评估依据，发现大模型能较准确地预测抑郁与 PTSD 等多类精神障碍的临床量表分数。

其余研究则在语言环境、动态评估和互动模式等方面展开创新。例如，Qi 等[39]利用中文社交媒体文本数据集开展自杀风险与认知歪曲的评估，结果表明，大模型在高风险人群早期筛查中具备潜力。针对传统大模型评估过于静态的问题，Hoang 等[40]将大模型应用于动机性访谈，能自动化识别患者的语气改变，为治疗过程中的动态评估提供了可能；为提升评估中的互动性，Yang 等[41]提出了结合大模型与游戏化设计的新型测评框架 PsychoGAT，可在抑郁、认知扭曲及人格特质等多维度实现更具有效性与满意度的状态评估。例如，在人格测试任务中，PsychoGAT 的 Cronbach’s $\alpha$ 信度系数达到 0.97，抑郁测量的 Cronbach’s $\alpha$ 信度系数为 0.77，认知扭曲检测的 Cronbach’s $\alpha$ 信度系数也均在 0.88 以上。

总体而言，状态评估与症状诊断面临类似问题，包括数据依赖性和缺少多模态的评估依据等。

### 3.2 大模型在心理疾病治疗的应用

大模型在心理疾病治疗的应用可主要分为两类：一是作为虚拟治疗师直接为用户提供干预与支持；二是作为辅助工具赋能临床医生或患者本人。传统的数字健康干预往往因对话僵化、缺乏深度共情而效果有限。大模型具有强大的生成式对话、情境化推理与个性化内容生成能力，为突破上述瓶颈提供了新可能。作为虚拟治疗师，大模型能模拟基本的治疗性对话，可提供完整的情感支持与结构化干预；作为辅助工具，则能帮助医生完成耗时的文书工作，以及为患者生成个性化的心理教育材料，从而提升整体治疗效率与质量，逐渐成为临床工作流程中有价值的补充。

#### 3.2.1 虚拟治疗师

虚拟治疗师指大模型通过专业干预手段提供共情式情绪支持，作为直接的治疗代理，为患者提供可及的、即时的心灵陪伴与心理干预。

一部分研究通过严格的实证研究，致力于验证虚拟治疗师在真实世界中的有效性与可靠性。Sharma 等[42]基于 2 000 余名用户的实际研究，首次在大规模样本中验证了大模型生成的重构思维

在“具体性”和“共情性”维度的用户接受度，为其实际效用提供了早期证据。为提供更高级别的循证依据，Heinz 等[43]开展了随机对照试验，有力地证明了基于大模型的聊天机器人 Therabot 能显著改善重度抑郁症和广泛性焦虑症等患者的临床症状。Wang 等[44]深入评估了大模型驱动的聊天机器人，发现其在遵循核心协议的同时，仍存在各类风险，例如，在“理解偏差”方面，专家发现系统多次误解用户语义，如将“尴尬”概述为“艰难”、将“可能”直接当作“是”来回应，这对未来模型的安全部署提出了警示。

另一部分研究聚焦于通过引入结构化心理框架提升干预的专业深度，避免生成空泛的安慰语。例如，Xiao 等[3]开发的 HealMe 模型，针对早期模型在认知重构中逻辑跳跃和缺乏系统性的问题，引入了系统化的心理框架，显著提升了干预步骤的严谨性与效果。类似地，Hu 等[45]提出了 PsyLLM 框架，通过将诊断推理与 CBT、ACT 和心理动力学等多种治疗流派相结合，实现了更具系统性和专业性的临床对话。

总体而言，虚拟治疗师的研究已从早期的功能验证发展到当前的有效性实证与风险探察阶段，并在提升专业性方面取得了显著进展。然而，其长期疗效、对复杂病例的处置能力以及伦理安全边界等问题，仍是未来研究需要攻克的关键挑战。

3.2.2 辅助工具

大模型辅助工具的核心价值在于化解传统心理治疗中的关键瓶颈：医患沟通的效率限制，个性化干预资源的匮乏，以及治疗师培训的高成本与高风险。此类研究旨在将大模型无缝嵌入诊断、治疗与培训各环节，通过自动化和智能化手段提升整体医疗质量与可及性。

一部分研究致力于优化临床工作流程与医患沟通，以提升治疗的连续性与效率。例如，Kim 等[46]提出了 MindfulDiary 应用，旨在解决传统治疗中患者自我记录依从性差以及信息零散的问题，它通过对话式日记自动化记录与结构化患者的日常体验，降低了患者的记录负担，为医生提供了直观的决策支持。为在真实世界中验证这类工具的临床效益，Habicht 等[47]进行了严格的临床实验研究，结果表明，结合大模型的 CBT 辅助工具能显著提高患者的出勤率、依从性和康复率，为其投入实际应用提供了关键的功效证据。

另一部分研究则聚焦于利用大模型大规模生成与扩展个性化干预资源，以突破传统方法在规模与成本上的限制。例如，Maddela 等[48]构建了 PATTERNREFRAME 数据集，直接针对认知重构练习中“非理性想法-理性回应”配对数据稀缺的挑战，为训练能自动生成个性化练习内容的模型奠定了资源基础。Lin 等[49]则进一步将积极心理学理论框架系统性地引入数据构造过程，并提出了积极构建框架及多语种数据集，旨在提升认知扭曲检测与积极重构的准确性、系统性与文化适应性。

此外，大模型在模拟患者以革新临床培训模式方面也展现出独特潜力，旨在为受训治疗师提供安全与可控的练习环境。Cabrera Lozoya 等[50]设计了 Client101 平台，利用大模型高保真地模拟抑郁与焦虑患者的思维与对话模式，从而为受训治疗师提供一个可反复练习且零临床风险的交互环境，有效降低了培训的门槛与成本。

总体而言，作为大模型在心理健康领域落地的重要路径，辅助工具的研究已在提升临床效率、丰富干预资源和创新培训模式 3 个维度展现出明确价值。现有成果初步验证了将大模型用作心理临床治疗中的辅助工具这一“人机协同”模式的可行性。然而，如何确保生成内容的临床精准度，如何与现有医疗系统深度集成并适应多样化的工作流，以及如何评估大模型的辅助功效，仍是大模型作为辅助工具规模化应用于心理临床治疗前必须解决的核心问题。

### 3.3　临床应用与技术发展关系的具体关联

从微观层面，临床应用的落地质量与特定技术的发展成熟度高度相关。大模型在心理疾病诊断与状态评估上的准确率依赖数据构造 (4.1 节) 提供的高质量临床语料，以及能力增强中的检索增强 (4.2.1 节) 补充的专业知识；大模型在心理疾病治疗中的临床适配性依赖数据构造 (4.1 节) 提供的多类型语料，而共情质量与干预效果得益于多智能体长对话建模 (4.2.2 节) 的上下文一致性优化，以及个性化支持 (4.2.3 节) 的人格定制技术。技术演进推动临床应用从功能验证向疗效验证跨越，临床需求则反向驱动技术向高可靠性及高可解释性迭代。

## 4　心理健康领域大模型技术发展

通用大模型在直接处理专业领域任务时存在显著瓶颈：首先，受限于患者隐私与伦理规范，高质量且大规模的真实临床对话数据极度稀缺，严重制约模型的领域适应性训练；其次，模型固有的领域专业知识薄弱、逻辑一致性不足，导致其回应常停留在表层共情，难以提供系统且可靠的心理干预；最后，模型的具体能力缺少规范化的评估，传统的人类心理能力标准不能很好适用于大模型能力的评估。为系统性应对上述挑战，学者从数据构造、能力增强和能力测评 3 个关键技术层面展开了深入探索，旨在夯实大模型应用于心理健康领域的技术基石，如图 2 所示。

### 4.1　数据构造

大模型在心理健康领域的应用离不开高质量数据的支撑。与一般开放域对话不同，心理健康相关对话涉及敏感信息、专业知识和高风险决策，因此，训练与微调必须依赖具有心理学理论依据、场景代表性和高可信度的数据。然而，真实的临床咨询数据受隐私与伦理限制而极为稀缺，同时人工标注与设计语料成本高昂，对标注者的心理学素养也有较高要求。这些因素使得传统依赖大规模真实数据的训练路径在心理健康场景下难以直接适用。为解决上述困境，研究者提出了多种数据构造方法，包括对话补全、案例重构、数据合成和混合策略，以在保护隐私和降低


图 2　心理健康领域大模型技术的发展脉络

**Fig. 2　The technical evolution of large language models in mental health**

成本的前提下，构建兼具专业性与规模的数据资源，从而支持大模型在诊断、干预与支持性对话等任务中的应用。4 种数据构造方法的对比如表 1 所示。数据构造方法的演进呈现“从质量优先到质效平衡”的趋势。早期以对话补全为主，核心目标是基于已有文本进行扩展补充；中期转向案例重构[51]与数据合成，重点在于建立标准化基准以及解决规模不足问题[52]；近期则以混合策略为主，通过多方法融合弥补单一方法缺陷[53]，其演进动力源于临床应用对“大规模高质量数据”的迫切需求。

**表 1 4 种数据构造方法对比**

**Table 1 A comparison of 4 types of data construction methods**

| 数据构造方法 | 核心优势 | 主要局限 | 适用场景 |
|---|---|---|---|
| 对话补全 | 依托现有文本、真实性较高 | 依赖高质量原始报告 | 个性化对话生成 |
| 案例重构 | 专业性强、符合心理学理论 | 成本高、规模有限 | 基准数据集构建 |
| 数据合成 | 成本低、场景覆盖广 | 易产生逻辑偏差、真实性不足 | 大规模预训练 |
| 混合策略 | 兼顾专业性与规模性 | 流程复杂、需多源数据融合 | 临床级模型微调、复杂场景 |

4.1.1 对话补全

对话补全方法以一个已有的且信息丰富的文本为核心，扩展和补全成连贯的多轮对话。核心是从一到多的扩展。

前期的研究重点在于验证从结构化报告生成高质量且可评估的多轮对话的可行性，并建立专业基准。Chen 等[54]提出的 SoulChat 通过扩大数据规模提升模型性能，利用大模型生成多轮对话的方式重构了 200 多万样本数据集。Xie 等[55]提出的 PsyDT 框架代表了这一方向上的最新进展。该方法不再满足于生成通用的咨询对话，而是旨在为特定咨询师构建其专属的“数字孪生”，通过采用 GPT-4 动态学习个体咨询师的独特风格，并基于单轮问答生成具有风格一致性的多轮对话，该工作显著提升了生成对话的真实感，为实现高度个性化的咨询模拟和培训资源生成开辟了新路径。

4.1.2 案例重构

案例重构是解决心理健康领域数据稀缺的基础性路径，其核心是基于心理学理论、专业报告或真实场景，通过人工设计、标注与严格的质控“重建”高质量的专业对话。案例重构与其他数据构造方法的核心差异在于人工干预程度。其依赖专业人员基于心理学理论进行设计与标注，因此在专业性与准确性上显著优于数据合成，但成本远高于数据合成；与对话补全相比，案例重构无须依赖现有文本，可从零构建特定场景对话，但规模扩展能力较弱，适用于基准构建，而非大规模训练。

早期的开创性工作侧重于为情感支持对话任务定义基础框架与标注体系。例如，Zhang 等[56]提出一个基于报告的多轮对话重构和评估框架 CPsyCoun，提供一种创新的两阶段方法，利用诊断报告中的指导信息产生高质量的对话。后续研究则致力于扩展数据的规模与对话的复杂性，以训练更具实用性的模型。例如，Liu 等[51]首次提出了情感支持对话 (ESC) 任务，并基于帮助技能理论构建了 ESConv 数据集。这项研究的首要动机是为该领域建立一个具有清晰心理学理论支撑和丰富策略标注的基准。其采用的求助者—支持者模式与策略标注体系为此后的一系列研究提供了至关重要的任务定义与数据范式。最新的研究前沿开始转向利用更强大的基座模型，实现从“生成通用对话”到“复现特定咨询师风格”的

跨越，追求极致的真实性与个性化。该工作标志着研究重心从“定义任务”向“规模化训练”的演进，为训练出具备更强心理支持能力的对话系统提供更充分的数据基础。

#### 4.1.3 数据合成

数据合成的核心是让大模型自身作为数据工厂，通过自我扮演或角色模拟批量生成对话数据，从而最大限度地减少对稀缺人工资源的依赖。

早期，数据合成研究的主要目标是建立可靠的生成框架，以保障合成数据的多样性与基本质量。例如，Zheng 等[52]提出的 Self-chats 框架，其核心动机是解决早期合成数据存在回复单一和场景覆盖有限的问题。他们通过设计“教师—学生”迭代框架与多样化响应修补 (DRI) 机制，系统性地指导模型生成覆盖多场景的对话，这一工作为后续研究提供了如何利用模型自身能力规模化生产高质量且多样化训练数据的方法论示范。随着基础生成框架的成熟，数据合成的研究重点开始转向通过更复杂的模拟机制提升合成数据的交互真实性与策略细腻度。Ye 等[57]提出的 SweetieChat 代表了这一方向上的前沿探索。该工作不再满足于生成通用的支持性对话，而是旨在模拟更贴近真实咨询生态的多角色、多策略互动。通过构建求助者、策略顾问与支持者 3 类角色，并在多样化场景中进行策略增强的角色扮演，该研究显著提升了合成数据在情感支持任务上的针对性与细腻度，推动了数据合成从“量产”向“优质”和“高保真”的跨越。

#### 4.1.4 混合策略

此类方法结合案例重构、补全与合成，实现多类型心理咨询语料的生成。例如，Wang 等[53]提出了 STAMPsy，收集了 5 000 多条混合类心理咨询对话，并将其与时空状态信息关联，通过迭代自我反馈生成多类心理咨询对话，构建了时空感知的混合型咨询数据集。Hu 等[58]提出了 PsycoLLM，通过多轮问答生成、证据判定与对话精炼生成多轮对话等方式，构建了覆盖问答、多轮对话与知识引导的多维心理学语料，验证了高质量、多类型心理数据的构建与生成能显著提升模型在心理健康场景中的表现。

### 4.2 能力增强

在心理健康场景中，大模型仅依赖内部知识往往难以满足临床对话的需求。首先，模型缺乏心理学理论和专业知识的支持，回应容易停留在表层安慰，难以提供系统干预；其次，长程对话中常出现遗忘或上下文不一致，难以保持连贯性与真实性；最后，不同文化背景和个体差异对心理支持的敏感度要求极高，而大模型在这方面的适配性仍显不足。为突破这些限制，研究者提出了多种能力增强方法，主要包括检索增强与知识融入、多智能体与长对话和多文化与个性化支持。

#### 4.2.1 检索增强与知识融入

检索增强与知识融入方法旨在通过引入外部知识源弥补大模型在专业领域的知识盲区，其技术发展脉络清晰地体现了从增强单一情绪理解到构建复杂知识管理系统的深化过程。

早期研究侧重于将外部知识图谱与情绪理论结合，为大模型提供基础的情绪与共情推理能力。例如，Tu 等[59]提出的 MISC 模型，其核心动机是解决大模型对用户细粒度情绪状态识别不足的问题，通过利用 COMET 知识模型进行情绪推理，再结合策略机制生成回应，首次实现了从情绪识别到干预策略的端到端动态结合。同样，Zhou 等[60]的工作致力于让模型的共情回应更具理论依据，他们通过构建常识图谱与情绪图谱的映射，使大模型的生成过程与心理学上的共情机制对齐，提升了回应的合理性与深度。桑晨扬等[61]则提出了情绪支持对话推理框架 CoES，将端到端的生成问题转化为分阶段的推理问题，针对性地设计了不同的外部知识增强策略，改善了大模

型在心理状态挖掘及支持策略选择过程中的推理效果。

在验证了基础知识融入的有效性后，研究重点转向构建更系统化且规模化的知识融合框架，以支持更专业和复杂的对话生成。Hu 等[62]提出的 APTNESS 框架代表了从“点状”知识注入到“体系化”理论支撑的迈进。该工作基于情绪评价理论构建了一套完整的“情绪调色板”与外部共情数据库，系统地增强了大模型在认知与情感两个维度的共情表现。在融合了完整检索增强与情感支持策略后，LLaMA2-7B 模型在共情、识别与安慰等关键维度上实现了全面进步。Qian 等[63]的研究则探索了如何通过提示工程与交互式检索，在不微调大模型的情况下激发其共情能力，为低成本应用提供了可行路径。

最新的进展则致力于构建高度结构化的知识管理流程，实现多源知识的协同与迭代优化。Yang 等[64]提出的 CascadeRCG 旨在解决海量异构外部知识的筛选、整合与一致性问题。其设计的级联流程实现了从“知识检索”到“知识精炼”的跨越，显著提升了生成内容的专业性与知识密度。

4.2.2 多智能体与长对话

该方法通过设计多个智能体协同工作机制，或增强模型对长上下文的理解能力，模拟更复杂的交互场景，以提升任务的真实性和复杂性。与以外部知识为核心目标的检索增强与知识融入不同，多智能体通过分工协作 (如辩论、建议、评判) 实现复杂临床问题的多维度推理，更侧重动态推理的协同优化[65]，长对话则需通过记忆机制优化实现跨会话上下文的连贯衔接。总体而言，检索增强主要提升模型诊断与干预的专业性和准确性，而多智能体与长对话更能模拟真实临床中的动态交互过程。

多智能体常用于提升模型的推理能力。Xiao 等[66]首次提出一个用于情绪障碍诊断的专业多智能体框架，将临床评估的粒度分析与结构化验证过程相结合，更准确地解释了复杂的精神病学数据。Xu 等[67]提出的 AutoCBT 框架是面向 CBT 的自主多智能体框架，以咨询师智能体为核心交互接口，搭配多个分别对应 CBT 核心原则 (共情验证、认知扭曲识别、反思挑战等) 的监督者智能体，整合动态路由策略与长短时记忆机制，实现了咨询对话的自主优化与上下文连贯性维持。其回应质量在共情、认知扭曲识别、策略实用性等核心维度显著优于单一提示词方法，在超过 70% 的心理问题咨询样本中被心理学专家评为最佳回答系统，有效验证了多智能体协作在提升咨询推理严谨性、长对话情景适配性方面的价值。

长对话主要通过提升模型的情景延续能力完成。Wang 等[68]提出了 AnnaAgent，一种融合情绪调节器与诉求激发器的动态代理系统，使代理能模拟求助者的动态心理状态；本文还设计了三级记忆机制，有效整合短期与长期记忆，从而支持跨会话的情境延续。这类研究展示了从“单轮优化”向“长程交互”演进的趋势，更接近真实的心理咨询场景。

4.2.3 多文化与个性化支持

该方法专注于提升大模型在提供心理健康支持时的文化敏感性、适应性和公平性，确保其回应符合不同文化背景用户和特定用户的价值观、信仰和交流习惯。同检索增强与知识融入、多智能体与长对话相比，该方法以“消除文化隔阂与个体适配偏差”为目标，技术上则依赖多语言和跨文化语料构建、文化标注适配策略，以及提示工程、参数高效微调等人格定制技术，重点增强大模型的跨文化语义理解、个体人格适配度，以及不同背景用户对干预内容的接受度与依从性。3 类技术形成互补协同关系，检索增强与知识融入为文化适配与个性化干预提供专业知识支撑，多智能体与长对话保障跨文化、个性化交互的连

贯性，多文化与个性化支持则让专业、连贯的干预服务更贴合多元用户需求，拓展模型应用的覆盖范围与实际效用。

多文化支持方面，主要通过构造不同语言的数据集提升大模型。例如：Qi 等[69]提出了基于训练有素咨询师角色扮演的数据集构建方法，通过模拟咨询师生动互动构建了包含 6 589 条长文本对话的日语心理咨询数据集 KokoroChat；Kim 等[70]则聚焦韩语语境中动机访谈 (motivational interviewing，MI) 相关公开数据集匮乏的痛点，提出了融合专业治疗师经验的 MI 会话模拟框架，最终构建了首个基于 MI 理论的韩语合成数据集 KMI，包含 1 000 条高质量动机访谈对话；Liu 等[71]构建了 CultureCare 数据集，通过文化标注与 4 种适配策略引导主流模型生成更具文化敏感性的回应，提升了大模型在跨文化语境中的回应质量。

个性化定制方面，主要通过微调等方式调整大模型心理学特质，以满足患者的个性化需求。例如，Zhou 等[30]提出了 CharacterGLM，通过构建多样化角色类别及行为特征数据集支持多场景角色对话定制。Jain 等[35]通过参数高效微调操控大模型人格，采用量化低秩适配对大五人格各维度进行精细化调控，实现个性化人格定制。王润斯等[72]基于 DeepSeek 模型和检索增强技术构建了人工智能辅导员架构，系统集成学生日常事务管理、心理危机干预与职业规划等核心功能模块。

### 4.3　能力测评

如何科学并系统地评估大模型在心理健康领域的专业能力与临床可靠性，已成为关乎其能否安全有效投入应用的关键环节。与通用领域的评测相比，心理健康领域的大模型评估面临三重独特挑战：首先，大模型输出的专业性、安全性与共情质量等维度难以通过简单的自动化指标进行衡量；其次，大模型需在遵守既定安全伦理边界的同时展现出贴近人类心理特质的对话风格；最后，评估体系还需能甄别大模型在复杂敏感临床情境下的推理能力与潜在风险。为应对这些挑战，构建多维度的能力测评体系至关重要。

#### 4.3.1　人类心理能力测评指标

在人类心理学研究与临床实践中，已有多种标准化测评工具被广泛应用于情绪、人格与社会认知等不同维度的能力评估。

在人格领域，大五人格量表[73]以开放性、责任心、外向性、宜人性和神经质 5 个维度为核心，成为研究人格差异与行为模式的基础工具。在社会情绪与同理心领域，多伦多述情障碍量表[74]主要衡量个体在情绪识别和情感表达方面的困难，为研究情绪调节障碍与心理健康的关系提供了重要工具。而同理心商数量表[75]则侧重于认知同理与情感同理两方面的能力测量，广泛用于探讨自闭症谱系障碍、社会适应与亲社会行为的心理机制。

在心理治疗师与医师等专业群体的能力评估中，共情和关系建构被视为核心指标。常见的测评工具涵盖自评、他评和行为观察多种形式。医师共情量表[76]是目前应用最广泛的医师自评量表，从情感理解、以患者为中心的照护取向和视角采择 3 方面考察医患互动中的“移情式理解”，已被广泛用于医学生筛选、课程效果评估及职业发展考核。与之互补的关系共情量表[27]则由患者完成评分，直接反映就诊体验中的关系性共情 (如倾听、情感确认与共同决策)，在初级保健与专科门诊均展现出良好的信效度。

在多维度共情测量方面，人际反应指针量表[77]将共情划分为视角采择、移情关怀、幻想投入与个人痛苦 4 个维度，既用于治疗师候选者与学员的基线特质评估，也被广泛用于解释沟通训练成效。与量表相比，行为编码体系能更客观地捕捉临床互动。罗特互动分析系统[78]通过语用功能标注量化医患互动中的情感回应与共情线索捕捉，是评估沟通培训和服务质量的黄金标准之一；四

习惯编码体系[79]则根据“建立融洽—探索视角—展示同理—达成共识”步骤对医师沟通行为进行量化，直接对接临床沟通技能培训与认证。

4.3.2 大模型能力测评指标

与传统的人类心理能力测评不同，大模型的心理能力测评不仅关注模型是否表现出类似人类的心理特质，更强调其在具体任务中的表现与潜在风险。人类的测评多以量表或行为观察为核心，侧重评估个体在共情、情绪识别与人际互动等方面的稳定特质；大模型的测评则更侧重功能性与任务导向，既要考察其在知识掌握、推理判断和对话生成等细分任务上的表现，也要评估其在诊断、干预及临床沟通模拟等高风险场景下的可靠性与安全性。

针对上述需求，研究者提出了多种评估框架和基准，其中一类是评估大模型在特定心理健康任务中的性能指标。Jin 等[80]提出的 PsyEval 是首个面向心理健康领域的大模型综合评估基准，涵盖知识、诊断与治疗三大维度下的 6 个子任务，通过简洁提示系统性地测量模型在心理健康任务中的表现，并借助人工评测与专家审查确保结果可靠性。Zhang 等[81]提出的 ConceptPsy 和 Zhao 等[82]提出的 CPsyExam 则聚焦于心理学知识，共同构建了中文心理学领域的大模型评估基准，覆盖本科阶段核心知识点，强调在知识点分布、概念覆盖与偏倚控制上的均衡性，从而揭示大模型在不同知识结构下的优势与不足。

除了任务导向的测评外，部分研究还尝试引入心理测量学方法，以提升科学性与解释力。Wang 等[83]提出的构念导向评估框架将情绪识别、同理心和推理等潜在心理构念作为评估核心，强调跨任务预测与解释模型表现的能力，从而突破单一任务准确率的局限。

部分研究已专门针对临床应用的需求做出回应，通过扩充评价维度，提出了创新心理指标。Tam 等[84]提出的 QUEST 框架聚焦医疗场景，提出涵盖规划、实施与评分 3 个阶段的人工评测流程，并提出信息质量、推理能力、安全性与信任度等核心维度，旨在系统量化医疗大模型的能力与风险。而 Liu 等[85]提出的 ChatCounselor 框架针对心理健康支持任务，基于真实咨询数据构建 Psych8k 数据集，并通过 7 项心理咨询指标评估大模型在专业咨询场景下的表现，验证了高质量领域数据在能力提升与测评中的价值。Wang 等[86]针对角色扮演类大模型的扮演能力展开了研究，提出 InCharacter 框架，通过心理量表访谈的方式评估角色扮演智能体的人格一致性，评估大模型在模拟复杂人格特征方面的能力。

此外，一些方法基于人类心理学评估指标，探索自评与人机结合的测评方式。Eberhardt 等[87]提出了一种基于模型自评的量表方法，利用 DISCOVER 框架从心理治疗会话转录中提取多模态信息，并由大模型生成心理构念评分，验证该框架在信度与效度上的可行性，显示了自动化测评的潜力。Szymanski 等[88]则系统比较了模型自评与专家评审的一致性，指出在专业知识任务中，大模型独立评估难以替代人类专家，强调未来测评需结合人工与大模型优势，确保评估的深度与可靠性。

### 4.4 技术发展与临床应用关系的内在关联

从宏观层面，大模型在心理健康领域的临床落地并非单一技术的孤立应用，而是通过数据构造、模型能力增强和模型效果测评的三阶闭环，系统性应用于临床两大场景 (心理疾病诊断和心理疾病治疗)。以大模型作为抑郁症虚拟治疗师的临床应用为例，首先，需通过数据构造方法为大模型的定向微调、检索库构建提供基本语料；其次，为提升治疗效果，需使用检索增强、智能体等方法，提升大模型在理解、共情回复等方面的能力；最后，需多维评测大模型的治疗效果，以实现反馈机制，更好地统筹并发展数据构建、能力增强等涉及的一系列技术。这不仅是前文所

述三大类技术的内在联系，也是目前大模型在心理健康领域临床应用的核心范式。

## 5　总结与展望

### 5.1　现存挑战

5.1.1　诊断依据与临床脱节

大模型虽然在心理症状识别方面展现出一定潜力，但其诊断依据往往过度依赖社交媒体文本或有限的语言线索。这种做法容易与临床实践中的标准化量表和结构化面谈等方法重复，缺乏额外的应用价值。例如：Eriksen 等[89]提出大模型虽然在部分下游任务上表现良好，但与标准化临床评估脱节；Goh 等[90]发现向医师提供大模型辅助并未显著改善诊断推理，表明实验室准确性并不意味着临床改进；Obradovich 等[91]同样强调了电子记录这类数据与标准化临床评估在信息粒度和诊断可靠性上的不一致性。

5.1.2　治疗模拟深度不足

大模型能模仿 CBT 和情绪支持对话等方式，但其干预逻辑通常停留在表层安慰或套路化的认知重构，难以实现临床心理师在真实治疗中的动态调整和长期疗效追踪。Gabriel 等[92]指出大模型虽能生成表层的共情与建议，但在治疗目标导向与长期干预策略方面仍不足；Li 等[93]指出，利用人工智能进行治疗常在短期内观察到患者改善，但缺少长期的治疗效果观察。大模型治疗师接近专业治疗师能力的证据不足；Stade 等[94]同样强调当大模型在连续性、个案适配与疗效追踪方面与真实临床治疗师仍有差距。

5.1.3　高质量心理数据的稀缺与不均衡

临床咨询数据因隐私与伦理限制极难获取，而现有数据集多为模拟生成或人工改写，其专业性、真实性与文化代表性仍存不足。Arora 等[95]强调高质量心理健康数据集标准化的重要性，指出当前许多研究受限于样本偏差与不均衡数据；Zhu 等[96]提出更广泛的训练数据可能会增强模型在不同医疗场景中泛化和有效执行的能力；Guo 等[6]同样指出临床咨询级别的数据稀少且不公开，导致训练与评估的偏差与代表性风险。

5.1.4　技术缺少临床验证

大多数与心理健康相关的大模型技术仍停留在技术验证阶段，缺乏大规模临床验证；而已有的测评框架往往以功能表现为主，缺少与临床效果和患者结局直接挂钩的指标。例如，共情能力测评多基于问卷或专家评审，而非实际疗效数据。Asgari 等[97]强调当前技术在临床部署之前需要专门的临床试验与风险评估指标；Goh 等[90]的研究结果表明，实验室准确性并不意味着临床改进，同样表明当前技术缺乏大规模临床验证。

5.1.5　大模型的可塑性暂未在临床上得到很好体现

虽然技术上可以通过微调和提示工程等方式定制大模型，定制后的大模型在某些测评指标上也展示出了更好的共情能力，但临床上尚未有明确证据表明定制化大模型与通用模型在治疗效果上存在显著差异。Hager 等[98]认为微调虽能改善某些任务表现，但仍存在幻觉、过拟合与鲁棒性差等问题，临床疗效层面的证据仍不足。

5.1.6　繁杂的心理能力评测指标缺乏合理性

为展示方法的先进性，许多研究往往提出新的大模型评估指标，但这些指标是否真正与临床疗效相关仍存疑。例如：Gabriel 等[92]指出很多现有自动指标 (BLEU、ROUGE 和流畅性分数等) 无法反映治疗效果；Farquhar 等[4]认为当前很多自动化评测忽视不确定性与幻觉检测，提出需要更严谨的评测指标，以贴近临床安全性，而非仅展示大模型先进性。

5.1.7　大模型在心理健康领域的应用面临严峻的责任归属困境

心理健康领域的高风险特性导致大模型应用陷入“临床试验悖论”：一方面，大模型需要通过大规模临床试验来证明疗效；另一方面，大模

型一旦在治疗中产生失误，现有体系下难以明确责任承担，导致研究机构和医疗系统不敢贸然开展临床试验。大模型在心理健康领域的应用面临严峻的责任归属困境，这一问题源于其“人机协同”的干预模式与现有医疗责任体系的不兼容。当模型出现误诊、过度建议或隐私泄露时，责任划分涉及开发者、医疗机构、临床医师等多主体：模型开发者的责任边界在于算法安全性与数据合规性，但现有技术难以完全规避“黑箱”导致的决策偏差，且缺乏明确的技术标准界定“合理缺陷”；医疗机构作为应用主体，需承担大模型临床适配与风险管控责任，但多数机构尚未建立针对 AI 工具的准入评估与监控机制；Goh 等[90]发现，临床医师作为最终决策者，面临“过度依赖大模型输出”与“完全忽视技术支持”的两难，其对大模型结果的判读责任缺乏明确规范；此外，Obradovich 等[91]也指出现有法律体系未针对 AI 辅助心理治疗制定专门的责任认定条款，导致事故发生后难以追溯核心责任方，这一困境已成为大规模临床验证的主要障碍。

5.1.8 临床医师缺乏将大模型应用在对应领域的技术能力

尽管大模型在症状识别任务中展现了良好的泛化能力，但临床心理师通常缺乏必要的技术背景，难以对大模型进行微调、提示优化或数据更新。Goh 等[90]指出除技术可用性外，医师的培训、工作流整合和对大模型输出的判读能力也是决定实际效益的关键因素；Roustan 等[99]则指出当前临床医生普遍缺乏对大模型关键技术(如提示词优化、风险评估)的理解与应用能力，强调教育与跨学科合作的必要性；Zhu 等[100]也提出，由于临床医师缺乏对大模型的提示优化、风险评估等关键技术能力，难以在特定语境下借助现有方法准确识别患者心理健康问题的具体表现，进而导致基于大模型的支持与干预策略难以充分发挥效用。

## 5.2 未来方向

3 条主线相互支撑、有机统一：可信数据体系为临床整合提供基础保障，临床整合需求为伦理监管明确实践导向，伦理监管体系为数据与临床应用划定安全边界。未来研究需跨学科协同推进 3 条主线的突破，最终实现大模型在心理健康领域的安全、有效、可及应用，缓解全球心理健康资源失衡的困境。

5.2.1 构建可信数据体系

数据是大模型性能的核心支撑，需突破稀缺和低质的恶性循环。一方面，应推进结构化合成数据技术的发展，融合心理学理论标注与专家知识，构建兼具专业性、真实性与规模性的混合数据集，弥补真实临床数据的不足；另一方面，建立多模态数据整合框架，纳入语音韵律、行为时序等非文本线索，丰富诊断与干预的信息维度，解决当前数据单一化问题。同时，制定心理健康领域数据质量标准，规范数据标注流程与隐私脱敏技术，确保数据使用的合规性与可靠性，为模型训练与评估提供统一基准，从源头规避数据偏差导致的诊断与干预风险。

5.2.2 深化临床整合应用

由于临床落地是技术的最终目标，因此需破解技术与临床脱节的难题。首先，构建人机协同诊疗框架，通过随机对照试验验证 AI 辅助诊疗与传统诊疗在长期疗效、患者依从性等核心结局指标上的差异，明确大模型在诊断筛查、干预支持中的定位与边界，为临床决策提供循证依据；其次，推进个性化动态干预大模型研发，融合记忆机制与因果推理，实现跨会话情境延续与干预策略的个体化调整，突破当前治疗模拟深度不足的瓶颈，提升干预的精准性与持续性；最后，优化临床工作流集成设计，开发直观化交互界面，将模型无缝嵌入电子病历、临床评估等现有流程，并加强临床医师的技术判读与风险管控培训，实现技术赋能而非替代，提升临床适配性与

工作效率。

5.2.3 完善伦理监管体系

由于伦理与监管是技术可持续发展的前提，因此需建立“全流程、多主体”的治理框架。在责任划分方面，明确开发者、医疗机构与临床医师的权责边界，通过立法与行业规范破解“临床试验悖论”，为大规模临床验证扫清障碍。在评测体系方面，构建以临床结局为导向的多维度评估框架，摒弃单纯追求性能指标的倾向，将疗效追踪、安全性检测与可解释性纳入核心评测维度，确保指标与临床价值直接关联。在伦理规范方面，制定心理健康大模型专属伦理准则，强化隐私保护与公平性设计，防范数据泄露与算法偏见；同时建立动态监管机制，实现技术应用的全生命周期风险管控，平衡创新发展与安全底线。

## 参 考 文 献


[1] GBD 2019 Mental Disorders Collaborators. Global, regional, and national burden of 12 mental disorders in 204 countries and territories, 1990–2019: a systematic analysis for the Global Burden of Disease Study 2019 [J]. The Lancet Psychiatry, 2022, 9(2): 137-150.

[2] Hengle A, Kulkarni A, Patankar SD, et al. Still not quite there! Evaluating large language models for comorbid mental health diagnosis [C] // Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, 2024: 16698-16721.

[3] Xiao MX, Xie QQ, Kuang ZY, et al. HealMe: harnessing cognitive reframing in large language models for psychotherapy [C] // Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics, 2024: 1707-1725.

[4] Farquhar S, Kossen J, Kuhn L, et al. Detecting hallucinations in large language models using semantic entropy [J]. Nature, 2024, 630(8017): 625-630.

[5] Ning YL, Teixayavong S, Shang YQ, et al. Generative artificial intelligence and ethical considerations in health care: a scoping review and ethics checklist [J]. The Lancet Digital Health, 2024, 6(11): e848-e856.

[6] Guo ZJ, Lai A, Thygesen JH, et al. Large language models for mental health applications: systematic review [J]. JMIR Mental Health, 2024, 11(1): e57400.

[7] Hua YN, Na HB, Li ZH, et al. A scoping review of large language models for generative tasks in mental health care [J]. npj Digital Medicine, 2025, 8(1): 230.

[8] Omar M, Soffer S, Charney AW, et al. Applications of large language models in psychiatry: a systematic review [J]. Frontiers in Psychiatry, 2024, 15: 1422807.

[9] Lawrence HR, Schneider RA, Rubin SB, et al. The opportunities and risks of large language models in mental health [J]. JMIR Mental Health, 2024, 11(1): e59479.

[10] Linardon J, Messer M, Anderson C, et al. Role of large language models in mental health research: an international survey of researchers' practices and perspectives [J]. BMJ Mental Health, 2025, 28(1): e301787.

[11] 籍欣萌, 咎红英, 崔婷婷, 等. 中文医疗大模型综述: 进展、评估与挑战 [J]. 中文信息学报, 2024, 38(11): 1-12.
Ji XM, Zan HY, Cui TT, et al. A survey of Chinese large language models in medicine: progress, evaluation and challenges [J]. Journal of Chinese Information Processing, 2024, 38(11): 1-12.

[12] 涂翠平, 张兰鸽, 张平. 人工智能在高校心理健康领域应用的实践探索 [J]. 北京教育 (德育), 2025, (9): 76-79.
Tu CP, Zhang LG, Zhang P. Practical exploration

of artificial intelligence applications in university mental health services [J]. Beijing Education, 2025, (9): 76-79.

[13] 王慧, 胡银环, 冯显东, 等. 人工智能在心理干预中的应用: 效果、挑战与前景 [J]. 中国全科医学, 2025, 28(25): 3209-3216.
Wang H, Hu YH, Feng XD, et al. The application of artificial intelligence in psychological interventions: effectiveness, challenges, and prospects [J]. Chinese General Practice, 2025, 28(25): 3209-3216.

[14] Yuan AJ, Garcia Colato E, Pescosolido B, et al. Improving workplace well-being in modern organizations: a review of large language model-based mental health chatbots [J]. ACM Transactions on Management Information Systems, 2025, 16(1): 1-26.

[15] Na HB, Hua YN, Wang ZM, et al. A survey of large language models in psychotherapy: current landscape and future directions [Z/OL]. arXiv Preprint, arXiv: 2502.11095, 2025.

[16] 陈旭日, 沈莹. 抑郁状态自动检测技术演进及发展 [J]. 心理月刊, 2025, 20(19): 218-221.
Chen XR, Shen Y. The evolution and development of automated depression detection technology [J]. Psychologies Magazine, 2025, 20(19): 218-221.

[17] 陈元乐, 刘荣勋, 秦士森, 等. 人工智能技术在学生群体抑郁焦虑识别中的研究进展 [J]. 临床精神医学杂志, 2025, 35(4): 333-336.
Chen YL, Liu RX, Qin SS, et al. Artificial intelligence assisted tools for the detection of depression and anxiety among students: a review [J]. Journal of Clinical Psychiatry, 2025, 35(4): 333-336.

[18] Jiang Y, Shen QY, Lai SZ, et al. Copiloting diagnosis of autism in real clinical scenarios via LLMs [Z/OL]. arXiv Preprint, arXiv: 2410.05684, 2024.

[19] Galatzer-levy I, Mcduff D, Malgaroli M. The capability of large language models to measure and differentiate psychiatric conditions through O-Shot learning [J]. Biological Psychiatry, 2025, 97(9): S62.

[20] Lan XZ, Han ZJ, Cheng YM, et al. Depression detection on social media with large language models [Z/OL]. arXiv Preprint, arXiv: 2403.10750, 2025.

[21] Jiang H, Zhang XJ, Cao XB, et al. PersonaLLM: investigating the ability of large language models to express personality traits [C] // Proceedings of the Findings of the Association for Computational Linguistics: NAACL 2024, 2024: 3605-3627.

[22] Li XX, Li YT, Qiu L, et al. Evaluating psychological safety of large language models [C] // Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, 2024: 1826-1843.

[23] Huang JT, Jiao WX, Lam MH, et al. On the reliability of psychological scales on large language models [C] // Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, 2024: 6152-6173.

[24] Huang JT, Wang WX, Li EJ, et al. On the humanity of conversational AI: evaluating the psychological portrayal of LLMs [C] // Proceedings of the Twelfth International Conference on Learning Representations, 2024: 1-24.

[25] Schaaff K, Reinig C, Schlippe T. Exploring ChatGPT's empathic abilities [C] // Proceedings of the 2023 11th International Conference on Affective Computing and Intelligent Interaction, 2023: 1-8.

[26] Schlegel K, Sommer NR, Mortillaro M. Large language models are proficient in solving and creating emotional intelligence tests [J]. Communications Psychology, 2025, 3(1): 80.

[27] Mercer SW, Maxwell M, Heaney D, et al. The consultation and relational empathy (CARE) measure: development and preliminary validation and reliability of an empathy-based consultation process measure [J]. Family Practice, 2004, 21(6):

699-705.

[28] Elyoseph Z, Hadar-shoval D, Asraf K, et al. ChatGPT outperforms humans in emotional awareness evaluations [J]. Frontiers in Psychology, 2023, 14: 1-7.

[29] Huang JT, Lam MH, Li EJ, et al. Apathetic or empathetic? Evaluating LLMs' emotional alignments with humans [C] // Proceedings of the Advances in Neural Information Processing Systems, 2024: 97053-97087.

[30] Zhou JF, Chen Z, Wan DZ, et al. CharacterGLM: customizing social characters with large language models [C] // Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track, 2024: 1457-1476.

[31] Li WK, Liu JR, Liu A, et al. BIG5-CHAT: shaping LLM personalities through training on human-grounded data [C] // Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics, 2025: 20434-20471.

[32] Li TL, Zheng XQ, Huang XJ, et al. Tailoring personality traits in large language models via unsupervisedly built personalized lexicon [Z/OL]. arXiv Preprint, arXiv: 2310.16582, 2023.

[33] Tu Q, Chen CQ, Li JP, et al. CharacterChat: learning towards conversational AI with personalized social support [Z/OL]. arXiv Preprint, arXiv: 2308.10278, 2023.

[34] Jiang GY, Xu MJ, Zhu SC, et al. Evaluating and inducing personality in pre-trained language models [C] // Proceedings of the 37th International Conference on Neural Information Processing Systems, 2023: 10622-10643.

[35] Jain N, Wu ZK, Villalobos CEM, et al. From text to emoji: how PEFT-driven personality manipulation unleashes the emoji potential in LLMs [C] // Proceedings of the Findings of the Association for Computational Linguistics: NAACL 2025, 2025: 4687-4723.

[36] Lho SK, Park SC, Lee H, et al. Large language models and text embeddings for detecting depression and suicide in patient narratives [J]. JAMA Network Open, 2025, 8(5): e2511922.

[37] Xu YJ, Fang ZX, Lin WN, et al. Evaluation of large language models on mental health: from knowledge test to illness diagnosis [J]. Frontiers in Psychiatry, 2025, 16: 2025.

[38] Tu SC, Powers A, Merrill N, et al. Automating PTSD diagnostics in clinical interviews: leveraging large language models for trauma assessments [C] // Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue, 2024: 644-663.

[39] Qi HZ, Fu GH, Li JQ, et al. Supervised learning and large language model benchmarks on mental health datasets: cognitive distortions and suicidal risks in Chinese social media [J]. Bioengineering, 2025, 12(8): 882.

[40] Hoang V, Rogers E, Ross R. How can client motivational language inform psychotherapy agents? [C] // Proceedings of the 9th Workshop on Computational Linguistics and Clinical Psychology, 2024: 23-40.

[41] Yang QS, Wang ZK, Chen HH, et al. PsychoGAT: a novel psychological measurement paradigm through interactive fiction games with LLM agents [C] // Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics, 2024: 14470-14505.

[42] Sharma A, Rushton K, Lin I, et al. Cognitive reframing of negative thoughts through human-language model interaction [C] // Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics, 2023: 9977-10000.

[43] Heinz MV, Mackin DM, Trudeau BM, et al. Randomized trial of a generative AI chatbot for mental health treatment [J]. NEJM AI, 2025, 2(4): AIoa2400802.

[44] Wang YZ, Wang YM, Xiao Y, et al. Evaluating an LLM-powered chatbot for cognitive restructuring:

insights from mental health professionals [Z/OL]. arXiv Preprint, arXiv: 2501.15599, 2025.

[45] Hu H, Zhou YC, Si JZ, et al. Beyond empathy: integrating diagnostic and therapeutic reasoning with large language models for mental health counseling [Z/OL]. arXiv Preprint, arXiv: 2505.15715, 2025.

[46] Kim T, Bae S, Kim HA, et al. MindfulDiary: harnessing large language model to support psychiatric patients' journaling [C] // Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems, 2024: 1-20.

[47] Habicht J, Dina LM, Mcfadyen J, et al. Generative AI-enabled therapy support tool for improved clinical outcomes and patient engagement in group therapy: real-world observational study [J]. Journal of Medical Internet Research, 2025, 27: e60435.

[48] Maddela M, Ung M, Xu J, et al. Training models to generate, recognize, and reframe unhelpful thoughts [C] // Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics, 2023: 13641-13660.

[49] Lin SY, Wang YX, Dong J, et al. Detection and positive reconstruction of cognitive distortion sentences: Mandarin dataset and evaluation [C] // Proceedings of the Findings of the Association for Computational Linguistics: ACL 2024, 2024: 6686-6701.

[50] Cabrera Lozoya D, Conway M, Sebastiano De Duro E, et al. Leveraging large language models for simulated psychotherapy client interactions: development and usability study of Client101 [J]. JMIR Medical Education, 2025, 11: e68056.

[51] Liu SY, Zheng CJ, Demasi O, et al. Towards emotional support dialog systems [C] // Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, 2021: 3469-3483.

[52] Zheng ZH, Liao LZ, Deng Y, et al. Self-chats from large language models make small emotional support chatbot better [C] // Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics, 2024: 11325-11345.

[53] Wang JY, Huang Y, Liu ZM, et al. STAMPsy: towards spatiotemporal-aware mixed-type dialogues for psychological counseling [C] // Proceedings of the AAAI Conference on Artificial Intelligence, 2025: 25371-25379.

[54] Chen YR, Xing XF, Lin JK, et al. SoulChat: improving LLMs' empathy, listening, and comfort abilities through fine-tuning with multi-turn empathy conversations [C] // Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2023, 2023: 1170-1183.

[55] Xie HJ, Chen YR, Xing XF, et al. PsyDT: using LLMs to construct the digital twin of psychological counselor with personalized counseling style for psychological counseling [C] // Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics, 2025: 1081-1115.

[56] Zhang CH, Li RH, Tan MH, et al. CPsyCoun: a report-based multi-turn dialogue reconstruction and evaluation framework for Chinese psychological counseling [C] // Proceedings of the Findings of the Association for Computational Linguistics: ACL 2024, 2024: 13947-13966.

[57] Ye J, Xiang L, Zhang YP, et al. SweetieChat: a strategy-enhanced role-playing framework for diverse scenarios handling emotional support agent [C] // Proceedings of the 31st International Conference on Computational Linguistics, 2025: 4646-4669.

[58] Hu JP, Dong TT, Luo G, et al. PsycoLLM: enhancing LLM for psychological understanding and evaluation [J]. IEEE Transactions on Computational Social Systems, 2025, 12(2): 539-551.

[59] Tu Q, Li YR, Cui JW, et al. MISC: a mixed strategy-aware model integrating COMET for

emotional support conversation [C] // Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, 2022: 308-319.

[60] Zhou JF, Zheng CJ, Wang B, et al. CASE: aligning coarse-to-fine cognition and affection for empathetic response generation [C] // Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics, 2023: 8223-8237.

[61] 桑晨扬, 马廷淮, 谢欣彤, 等. 基于大语言模型多阶段推理的情绪支持对话生成方法 [J]. 计算机科学与探索, 2024, 18(11): 2925-2939.
Sang CY, Ma TH, Xie XT, et al. Multi-stage reasoning method for emotional support dialogue generation based on large language models [J]. Journal of Frontiers of Computer Science & Technology, 2024, 18(11): 2925-2939.

[62] Hu YX, Tan MH, Zhang CW, et al. APTNESS: incorporating appraisal theory and emotion support strategies for empathetic response generation [C] // Proceedings of the 33rd ACM International Conference on Information and Knowledge Management, 2024: 900-909.

[63] Qian YS, Zhang WN, Liu T. Harnessing the power of large language models for empathetic response generation: empirical investigations and improvements [C] // Proceedings of the findings of the Association for Computational Linguistics: EMNLP 2023, 2023: 6516-6528.

[64] Yang D, Zhu JW, Wu HH, et al. CascadeRCG: retrieval-augmented generation for enhancing professionalism and knowledgeability in online mental health support [C] // Proceedings of the ACM on Web Conference 2025, 2025: 1465-1469.

[65] Li RH, Tan MH, Wong DF, et al. CoEvol: constructing better responses for instruction finetuning through multi-agent cooperation [C] // Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, 2024: 4703-4721.

[66] Xiao MX, Ye M, Liu B, et al. A retrieval-augmented multi-agent framework for psychiatry diagnosis [Z/OL]. arXiv Preprint, arXiv: 2506.03750, 2025.

[67] Xu AC, Yang D, Li RH, et al. AutoCBT: an autonomous multi-agent framework for cognitive behavioral therapy in psychological counseling [Z/OL]. arXiv Preprint, arXiv: 2501.09426, 2025.

[68] Wang M, Wang PD, Wu L, et al. AnnaAgent: dynamic evolution agent system with multisession memory for realistic seeker simulation [C] // Proceedings of the Findings of the Association for Computational Linguistics: ACL 2025, 2025: 23221-23235.

[69] Qi ZY, Kaneko T, Takamizo K, et al. KokoroChat: a Japanese psychological counseling dialogue dataset collected via role-playing by trained counselors [Z/OL]. arXiv Preprint, arXiv: 2506.01357, 2025.

[70] Kim H, Lee S, Cho Y, et al. KMI: a dataset of Korean motivational interviewing dialogues for psychotherapy [Z/OL]. arXiv Preprint, arXiv: 2502.05651, 2025.

[71] Liu CC, Arnaout H, Kovačić N, et al. Tailored emotional LLM-supporter: enhancing cultural sensitivity [Z/OL]. arXiv Preprint, arXiv: 2508.07902, 2025.

[72] 王润斯, 赵革, 胡晓龙. 基于 DeepSeek-RAG 的人工智能辅导员系统框架研究 [J]. 南京开放大学学报, 2025, (2): 1-6.
Wang RS, Zhao G, Hu XL. On the framework of artificial intelligence counselor system based on DeepSeek-RAG [J]. Journal of Nanjing Open University, 2025, (2): 1-6.

[73] John OP, Donahue EM, Kentle RL. Big five inventory [J]. Journal of Personality and Social Psychology, 1991.

[74] Bagby RM, Parker JDA, Taylor GJ. The twenty-

item Toronto Alexithymia scale—i. item selection and cross-validation of the factor structure [J]. Journal of Psychosomatic Research, 1994, 38(1): 23-32.

[75] Baron-Cohen S, Wheelwright S. The empathy quotient: an investigation of adults with Asperger syndrome or high functioning autism, and normal sex differences [J]. Journal of Autism and Developmental Disorders, 2004, 34: 163-175.

[76] Hojat M, Gonnella JS, Nasca TJ, et al. The Jefferson scale of physician empathy: further psychometric data and differences by gender and specialty at item level [J]. Academic Medicine : Journal of the Association of American Medical Colleges, 2002, 77: 58-60.

[77] Davis MH. Measuring individual differences in empathy: evidence for a multidimensional approach [J]. Journal of Personality and Social Psychology, 1983, 44(1): 113-126.

[78] Roter D, Larson S. The Roter interaction analysis system (RIAS): utility and flexibility for analysis of medical interactions [J]. Patient Education and Counseling, 2002, 46(4): 243-251.

[79] Krupat E, Frankel R, Stein T, et al. The Four Habits Coding Scheme: validation of an instrument to assess clinicians' communication behavior [J]. Patient Education and Counseling, 2006, 62(1): 38-45.

[80] Jin HA, Chen SY, Dilixiati D, et al. PsyEval: a suite of mental health related tasks for evaluating large language models [Z/OL]. arXiv Preprint, arXiv: 2311.09189, 2024.

[81] Zhang JL, He HL, Song NR, et al. ConceptPsy: a benchmark suite with conceptual comprehensiveness in psychology [Z/OL]. arXiv Preprint, arXiv: 2311.09861, 2024.

[82] Zhao JH, Zhu JW, Tan MH, et al. CPsyExam: a Chinese benchmark for evaluating psychology using examinations [C] // Proceedings of the 31st International Conference on Computational Linguistics, 2025: 11248-11260.

[83] Wang XT, Jiang LM, Hernandez-Orallo J, et al. Evaluating general-purpose AI with psychometrics [Z/OL]. arXiv Preprint, arXiv: 2310.16379, 2023.

[84] Tam TYC, Sivarajkumar S, Kapoor S, et al. A framework for human evaluation of large language models in healthcare derived from literature review [J]. npj Digital Medicine, 2024, 7(1): 258.

[85] Liu JM, Li DH, Cao H, et al. ChatCounselor: a large language models for mental health support [Z/OL]. arXiv Preprint, arXiv: 2309.15461, 2023.

[86] Wang XT, Xiao YZ, Huang JT, et al. InCharacter: evaluating personality fidelity in role-playing agents through psychological interviews [C] // Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics, 2024: 1840-1873.

[87] Eberhardt ST, Vehlen A, Schaffrath J, et al. Development and validation of large language model rating scales for automatically transcribed psychological therapy sessions [J]. Scientific Reports, 2025, 15(1): 29541.

[88] Szymanski A, Ziems N, Eicher-miller HA, et al. Limitations of the LLM-as-a-judge approach for evaluating LLM outputs in expert knowledge tasks [C] // Proceedings of the 30th International Conference on Intelligent User Interfaces, 2025: 952-966.

[89] Eriksen AV, Möller S, Ryg J. Use of GPT-4 to diagnose complex clinical cases [J]. NEJM AI, 2024, 1(1): AIp2300031.

[90] Goh E, Gallo R, Hom J, et al. Large language model influence on diagnostic reasoning: a randomized clinical trial [J]. JAMA Network Open, 2024, 7(10): e2440969.

[91] Obradovich N, Khalsa SS, Khan WU, et al. Opportunities and risks of large language models in psychiatry [J]. NPP —Digital Psychiatry and Neuroscience, 2024, 2(1): 8.

[92] Gabriel S, Puri I, Xu XH, et al. Can AI relate: testing large language model response for mental health support [Z/OL]. arXiv Preprint, arXiv: 2405.12021, 2024.

[93] Li H, Zhang RW, Lee YC, et al. Systematic review and meta-analysis of AI-based conversational agents for promoting mental health and well-being [J]. npj Digital Medicine, 2023, 6(1): 236.

[94] Stade EC, Stirman SW, Ungar LH, et al. Large language models could change the future of behavioral healthcare: a proposal for responsible development and evaluation [J]. npj Mental Health Research, 2024, 3(1): 12.

[95] Arora A, Alderman JE, Palmer J, et al. The value of standards for health datasets in artificial intelligence-based applications [J]. Nature Medicine, 2023, 29(11): 2929-2938.

[96] Zhu JW, Tan MH, Yang M, et al. CollectiveSFT: scaling large language models for Chinese medical benchmark with collective instructions in healthcare [C] // Proceedings of the International Conference on Social Robotics, 2024: 51-60.

[97] Asgari E, Montaña-brown N, Dubois M, et al. A framework to assess clinical safety and hallucination rates of LLMs for medical text summarisation [J]. npj Digital Medicine, 2025, 8(1): 274.

[98] Hager P, Jungmann F, Holland R, et al. Evaluation and mitigation of the limitations of large language models in clinical decision-making [J]. Nature Medicine, 2024, 30(9): 2613-2622.

[99] Roustan D, Bastardot F. The clinicians' guide to large language models: a general perspective with a focus on hallucinations [J]. Interactive Journal of Medical Research, 2025, 14(1): e59823.

[100] Zhu JW, Xu AC, Tan MH, et al. XinHai@ CLPsych 2024 shared task: prompting healthcare-oriented LLMs for evidence highlighting in posts with suicide risk [C] // Proceedings of the 9th Workshop on Computational Linguistics and Clinical Psychology, 2024: 238-246.