心理科学进展 2026, Vol. 34, No. 6, 1072–1083 https://doi.org/10.3724/SP.J.1042.2026.1072
Advances in Psychological Science https://cstr.cn/32111.14.2026.1072

基于大模型的智能体在大学生心理咨询中的应用*

郭 静[1] 王 沛[2] 马胤哲[3] 陈路晰[4]
郭 可[4] 胡彦熙[2] 刘 荷[2]

([1]中国人民大学人口与发展研究中心, 北京人口发展与治理研究创新中心, 国家治理大数据和人工智能创新平台, 北京 100872) ([2]中国人民大学人口与健康学院, 北京 100872)
([3]北京理工大学计算机学院, 北京 100081) ([4]中国人民大学高瓴人工智能学院, 北京 100872)

摘 要 大学生群体面临的心理健康挑战日趋复杂, 而传统高校心理咨询模式存在一定局限。为此, 本研究创新性地提出一种融合心理学与人工智能的技术框架: 通过将心理咨询垂域知识与数据融入基座大模型, 构建由测评师、咨询师、督导师三类心理咨询智能体与大学生智能体共同组成的“测评–咨询–督导”多智能体协作系统。系统采用“内循环训练–外循环服务”双循环模式, 在“内循环训练”阶段, 测评师、咨询师智能体与大学生智能体通过虚拟场景交互模拟真实咨询流程, 并利用督导师智能体的反馈优化服务策略, 积累个性化咨询档案与多流派干预经验; 在“外循环服务”阶段, 心理咨询智能体基于“内循环训练”成果, 为真实来访大学生提供专业化、精准化的心理测评与干预服务。系统有望成为大学生心理咨询的有效辅助工具, 助力高校心理健康服务。

关键词 智能体, 大模型, 心理咨询, 大学生
分类号 R395

1 引言

大学生作为社会未来发展的中坚力量, 其心理健康不仅关系个体成长, 更关乎国家人才储备与社会稳定。然而, 在经济社会快速转型与多元文化冲击下, 学业内卷加剧、社交关系虚拟化、就业不确定性攀升等问题相互交织, 大学生群体面临的心理健康挑战日趋复杂。《中国国民心理健康发展报告(2021~2022)》显示, 大学生抑郁风险检出率达 21.48%, 焦虑风险检出率更高达 45.28% (傅小兰 等, 2023), 心理健康问题已成为高等教育领域亟待解决的重大议题。

高质高效的心理咨询服务对大学生心理健康问题的预防与干预具有关键作用。近年来, 我国相继出台《关于“十四五”国民健康规划的通知》(2022)与《全面加强与改进新时代学生心理健康工作专项行动计划(2023–2025)》(2023)等政策, 明确提出加强高校心理咨询服务平台建设。然而, 传统的高校线下心理咨询模式存在一定局限性: 其一, 时空限制导致服务可及性不足, 大学生常因课业冲突或地理阻隔难以及时获取支持; 其二, 面对面咨询易诱发“病耻感”, 降低大学生的求助意愿(李佳 等, 2023)。新兴的在线心理咨询平台虽然突破了时空壁垒, 并通过虚拟交互弱化“病耻感”, 但仍面临三重瓶颈: 一是专业性不足, 部分平台缺乏规范化心理咨询知识支撑, 服务流于表层情感支持; 二是针对性缺失, 忽视大学生心理健康问题的异质性, 采用“一刀切”式标准化应答; 三是追踪性欠缺, 未建立个性化咨询档案与长期干预方案, 导致服务碎片化(腾讯研究院, 2024) 。

生成式人工智能技术的突破为破解上述困境提供了新契机。基于大模型(Large Language Model, LLM)的智能体(Agent)技术凭借其自主性、反应性、主动性与社交能力(Akhtar & Nauman,

收稿日期: 2025-05-15
* 中国人民大学科学研究基金(中央高校基本科研业务费专项资金资助)项目成果(25XNT005)。
通信作者: 胡彦熙, E-mail: huyanxi@ruc.edu.cn
刘荷, E-mail: liuhe2024@ruc.edu.cn

2015), 展现出在复杂场景中提供拟人化心理支持的潜力。尽管已有研究提出心理咨询智能体并初步验证其有效性, 但当前系统仍缺乏心理测评与干预服务的闭环流程及多流派择优匹配机制(Na, 2024)。本研究立足跨学科视角, 通过将心理咨询垂域(特定领域)知识与数据融入基座大模型, 构建“测评−咨询−督导”多智能体协作系统, 并提出“内循环训练−外循环服务”双循环模式, 旨在开发面向大学生群体的智能化心理咨询辅助工具, 推动高校心理健康服务体系向专业化、精准化、系统化方向持续发展。

2　技术框架与知识融入

2.1　基于大模型的智能体技术框架

智能体是一种能够感知环境、自主决策并执行任务的智能实体, 具备模拟人类认知与交互行为的能力(瞿晶晶 等, 2025)。随着人工智能技术的演进, 智能体技术经历了符号逻辑驱动、反应式行为控制、强化学习优化等发展阶段(Xi et al., 2025)。传统智能体常受限于过度抽象失真、场景适应性不足等问题, 而基于大模型的智能体技术则突破了这些局限。大模型(如 Qwen、Chatgpt)通过海量多模态数据(如文本、视频)训练, 借助超大规模参数对真实世界知识进行压缩与内化, 赋予智能体强大的语言理解、逻辑推理与创造性生成能力(肖红江 等, 2024)。基于大模型的智能体技术框架以大模型为基座, 支持单智能体构建与多智能体协作。

2.1.1　单智能体构建

单智能体指能够独立执行专业化任务的智能实体。其基础架构包含概要、记忆、规划与行动四个核心模块; 通过非参数化提示与参数化训练的构建方法整合个性化数据, 有效模拟特定个体(Mou et al., 2024; Wang, Ma et al., 2024)。

为保证模拟精度, 构建能够准确复制个体特征的智能体架构至关重要(袁洁铃, 陈海丹, 2025), 这需要在理论抽象与实际实现之间取得平衡, 以充分捕捉人类行为的复杂性。该架构由以下四个模块构成：(1)概要模块：定义智能体的基本属性与特征, 表征被模拟个体的独特性。概要分为描述式概要与对话式概要两种类型, 描述式概要通过自然语言直接阐述智能体的背景信息, 对话式概要则通过对话示例展现智能体的交互特征(Wang et al., 2023; Yu et al., 2024)。概要的构建方式包括手工整理与大模型生成, 手工整理结合垂域知识人工定义智能体属性, 确保信息准确性; 大模型生成则通过输入少量关键词, 利用大模型的生成能力自动扩展出符合逻辑的完整智能体设定, 支持快速规模化部署(Cheng et al., 2023; Wang et al., 2025)。(2)记忆模块：负责存储智能体与环境交互的历史信息, 克服大模型有限的上下文窗口限制。记忆分为短期记忆与长期记忆两类, 短期记忆记录当前任务的上下文信息, 支持即时决策; 长期记忆存储持久的垂域知识与个性化档案, 通过数据库实现高效检索与动态更新(Tang et al., 2026; Xiang et al., 2023)。记忆操作包括写入、检索与反思, 写入操作实时记录对话历史与观察结果; 检索操作基于一定规则调取数据库信息; 反思操作提炼优化策略并更新长期记忆, 推动知识进化(Park et al., 2023)。(3)规划模块：指导智能体根据目标制定行动策略, 模拟个体交互时的思维过程。规划分为主观规划与共情规划两种类型, 主观规划基于预设规则直接输出确定性策略, 共情规划则通过分析环境中其他主体的意图生成行动方案(Lee, Lee et al., 2024; Xie et al., 2024)。(4)行动模块：将规划结果转化为具体输出, 是模拟人类行为的接口。行动分为封闭域与开放域两类, 封闭域行动限定于预定义的操作集合内; 开放域行动允许智能体自由生成内容, 适应复杂场景需求(Horton et al., 2024; Zhou et al., 2024)。此外, 行动场景可根据环境复杂度细分为简单对话场景与精心塑造场景(Brahman et al., 2021; Chen, Chen et al., 2024)。

构建方法旨在将真实个体数据整合到大模型中, 实现智能体与被模拟个体的有效对齐(柴春雷等, 2025), 主要分为两类：一是非参数化提示, 通过自然语言将角色数据直接输入大模型的提示词, 利用大模型的上下文学习能力模拟角色行为(Wei et al., 2022); 二是参数化训练, 通过预训练(Pre-training)、监督微调(Supervised Fine-tuning, SFT)或强化学习(Reinforcement Learning, RL)等技术调整大模型参数, 使智能体深度适配特定角色。其中, 预训练是在大规模垂域数据上直接训练模型, 监督微调利用标签数据优化模型权重, 强化学习则通过奖励函数引导模型在动态环境中学习最优策略(Deng et al., 2023; Salemi et al.,

2024; Shea & Yu, 2023)。在单智能体构建中，其四个模块均可采用非参数化提示方法实现(Mou et al., 2024)。此外，概要模块还可通过预训练或监督微调技术建立，记忆模块依托数据库技术实现数据的写入、检索与反思，规划模块基于思维链技术进行推理，行动模块也可通过强化学习技术实现迭代优化(Chawla et al., 2023; Lee, Moon et al., 2024; Ran et al., 2024; Zhong et al., 2024)。

2.1.2 多智能体协作

单智能体虽可通过专业化构建执行特定任务，但在应对动态复杂场景时仍面临知识共享不足、协同效率低等局限。多智能体协作系统通过整合多个单智能体的差异化能力，可显著提升复杂任务的解决效能(郭陆祥 等, 2025)。其实现过程分为三个阶段：首先整合核心要素搭建协作系统，其次基于特定模式将系统应用于实际场景，最后对系统的协作成效进行全面评估。

多智能体协作系统指在特定环境中，多个智能体通过受限的通信方式组织起来扮演不同角色，共同完成任务。该系统由以下四个要素构成：(1)环境要素：定义智能体交互的虚拟场景。环境包含四个属性：一是配置，设定系统背景信息与初始目标；二是状态，涵盖环境及其他智能体的实时信息；三是历史，记录协作过程中环境状态的时序变化；四是工具，提供任务执行所需的专业资源(Chen, Dong et al., 2024; Chen, Jiang et al., 2024; Hong et al., 2024; Liang, Wu et al., 2024)。(2)角色要素：根据任务需求分配智能体职能。参与者直接执行具体任务，引导者则统筹任务流程与资源分配(Ni & Yang, 2024; Xiong et al., 2023)。(3)组织要素：规定智能体间的协作规则。组织形式明确各个智能体的任务分工，组织结构设计智能体间的协作层级(Hao et al., 2025; Yan & Xiang, 2025)。(4)通信要素：控制智能体间的信息交互。通信形式支持自然语言对话与结构化指令，通信风格分为合作性与竞争性交互(Du et al., 2024; Fan et al., 2025)。在合作性交互中，多智能体拥有共同目标，旨在优化集体成果；在竞争性交互中，智能体通常持有对立观点，力图战胜对方。

基于任务目标差异，多智能体协作分为两种模式：其一，社会模拟模式，构建虚拟社会场景，智能体通过拟人化交互模拟群体行为，自主演化社会动态，用于社会现象仿真与资源分配优化(Chuang et al., 2024)。其二，问题解决模式，以解决特定领域问题为导向，智能体基于预设角色职能，通过自动化决策与高效协作实现目标(Tang et al., 2024)。

多智能体协作评估的重点在于任务解决成效，具体包括：衡量系统整体目标达成度的总体评估，分析单智能体决策质量与贡献度的子任务评估，量化响应速度、资源消耗等技术指标的系统评估(Liang, He et al., 2024; Qian et al., 2024; Tan et al., 2024)。评估方法包括人工评分、大模型评分与自动化评分。

2.2 心理咨询垂域知识

尽管多智能体协作系统为复杂场景的任务协同提供了通用方法论，但其在心理咨询领域的有效落地必须深度融入垂域知识。心理咨询是一项兼具结构化与情境化的专业任务，其垂域知识涵盖心理咨询流程、心理咨询流派与心理咨询评估三个核心维度，系统规范心理咨询服务的基本框架、技术路径与质量标准。

2.2.1 心理咨询流程

心理咨询流程由一系列阶段性任务构成。尽管不同心理咨询流派对阶段的划分存在差异，但普遍遵循基本框架，一般将其归纳为 5 个阶段：(1)进入与定向阶段：通过初步交流明确来访者需求与咨询匹配度，决定是否与其建立正式咨询关系。(2)问题与个人探索阶段：协助来访者厘清心理健康问题的显性表现与深层动因，不仅关注外部事件对个体的影响，更注重引导来访者探索自我认知模式、情绪反应与行为策略间的潜在关联。(3)目标与方案探讨阶段：基于“需求−差距分析”，结合来访者特征与环境资源，共同制定可量化的干预目标及分阶段的行动方案。(4)行动与转变阶段：运用特定的心理咨询流派技术推动来访者发生改变，动态监测来访者的阻抗或反复现象，及时调整服务策略。(5)评估与结束阶段：通过标准化工具与质性反馈评估服务质量，并引导来访者总结成长经验以巩固干预成效(江光荣, 2012)。

2.2.2 心理咨询流派

心理咨询流派包括结构化流派与非结构化流派两类，其差异主要体现在技术路径方面。强调系统性干预的结构化流派主要包括：(1)认知行为疗法(Cognitive Behavioral Therapy, CBT)：基于“认知−情绪−行为”三元模型，通过识别自动化思

维、修正认知偏差、重构核心信念, 改善个体情绪困扰(Beck, 2021; 蒙艺, 钟宇豪, 2024)。(2)焦点解决短期疗法(Solution-Focused Brief Therapy, SFBT): 以未来导向与资源聚焦为核心, 通过激活个体潜能, 实现短程高效干预(Corey, 2016; 骆宏, 杜奕, 2023)。侧重开放式探索的非结构化流派则主要包括: (1)精神分析疗法(Psychoanalytic Therapy): 通过自由联想、移情分析等技术揭示个体潜意识冲突与防御机制, 重构其心理动力模式(郭清, 2024; 叶浩生, 杨莉萍, 2021)。(2)人本主义疗法(Humanistic Therapy): 以无条件积极关注、共情理解与一致性表达营造非评判性氛围, 激发来访者的自我治愈潜能(Maslow, 1987; 王东美等, 2022)。

2.2.3　心理咨询评估

心理咨询评估以质量标准为核心, 贯穿服务全过程, 涵盖三种类型: (1)形成评估: 通过计划科学性、技术适宜性等指标, 对咨询方案进行前瞻性评价。(2)过程评估: 采用心理咨询师评估量表(Psychological Counselor Assessment Questionnaire)等工具, 衡量咨询师的专业胜任力与共情能力, 动态监测来访者满意度(罗莉娟 等, 2025)。(3)效果评估: 基于抑郁−焦虑−压力量表(Depression Anxiety and Stress Scale, DASS)等工具的前后测差值, 量化来访者心理状态的改善情况, 进而评估干预成效(黄峰 等, 2025)。

2.3　智能体技术在心理咨询领域的应用现状

当前, 基于大模型的智能体技术在心理咨询领域的应用研究主要集中在构建测评师、咨询师智能体这两类单智能体。得益于智能体技术的高度拟人特征, 它们分别能够为真实来访者提供高质量的心理测评、干预服务。这些智能体均以大模型为基座, 通过嵌入心理咨询理论(如标准化心理健康量表、心理干预专家指南)的提示词工程, 对齐咨询对话历史(如真实测评师与来访者对话数据、在线心理咨询平台问答数据)的监督微调等方法建立(Chen, Zhang et al., 2024; Lai et al., 2024; Soman et al., 2025; Yang et al., 2024)。研究者已通过多种方法验证其有效性, 包括真实咨询师开展的人工评分, 来访者智能体进行的大模型评分, 以及与心理咨询基准测试集对比的自动化评分等(Wang, Xiao et al., 2024; Zhang et al., 2025; Zhu et al., 2025)。

然而, 现有单智能体面临生成内容发散与真实数据稀缺等挑战, 多智能体协作系统的出现为解决这些问题提供了新思路。受真实咨询场景启发, 研究者引入了督导师智能体。该智能体通过定期反思, 从旁观者视角为测评师、咨询师智能体提供反馈, 以纠正服务方向偏差, 缓解内容发散问题(Lan et al., 2024; Xu et al., 2025)。另有研究利用少量真实来访者的心理咨询档案与对话历史等数据, 构建罹患精神疾病的来访者智能体。通过模拟测评师智能体与来访者智能体、咨询师智能体与来访者智能体之间的对话, 有效扩充真实数据, 进而促进测评师、咨询师智能体服务能力的自主进化(Lan et al., 2024; Qiu & Lan, 2024)。

尽管如此, 现有研究仍存在以下局限: 主要面向精神疾病患者, 对存在一般心理健康问题的普通人群关注不足; 心理测评与干预服务环节相互独立, 缺乏有效串联; 咨询师智能体通常仅掌握单一流派的干预技术, 缺乏为来访者匹配最优流派咨询师的机制。针对上述问题, 本研究旨在搭建一个集成“测评−咨询−督导”功能的多智能体协作系统, 设计最优流派匹配的决策机制, 并引入模拟心理健康问题的大学生智能体, 以提升心理咨询智能体的综合服务能力。

3　应用路径

3.1　心理咨询领域的单智能体构建

基于传统心理咨询服务框架, 本研究将智能体划分为心理咨询智能体与大学生智能体两类, 心理咨询智能体进一步细分为测评师、咨询师与督导师智能体, 分别承担心理测评、干预与督导职能; 大学生智能体则通过模拟异质性的心理健康问题, 为心理咨询智能体提供虚拟训练场景。两类智能体的基础架构与构建方法存在一定差异, 见表 1。

3.1.1　心理咨询智能体构建

测评师、咨询师、督导师三类心理咨询智能体采用模块化的基础架构设计, 通过融入心理咨询垂域知识与数据, 实现对真实从业者行为的精细化模拟, 从而提升服务的专业性与针对性, 以契合心理咨询任务的结构化与情境化特征。

基于概要、记忆、规划与行动四个模块, 搭建心理咨询智能体的基础架构: (1)概要模块: 心理咨询智能体通过描述式概要定义人物身份(如

表 1 心理咨询领域的单智能体构建

智能体	基础架构							构建方法
	概要		记忆		规划	行动		
	类型	构建方法	类型	操作	类型	类型	场景	
测评师 咨询师 督导师	描述式概要	手工整理	短期记忆 长期记忆	写入 检索 写入 检索 反思	主观规划 共情规划	开放域行动	简单对话场景	非参数化提示 参数化训练
大学生		大模型生成	短期记忆	写入	共情规划			

“您是一位高校心理测评师”), 结合手工整理真实测评师(如量表题目与解读方法、信息挖掘技术)、咨询师(如咨询流派专家指南、共情技巧)与督导师(如督导理论模型、指导策略)的背景数据, 确保智能体身份描述的准确性。系统持续归纳已有背景数据, 并动态引入新的真实从业者样本以更新概要。在此基础上, 通过数据增量、参数扰动与样本分布变化等多个维度进行敏感性分析, 评估心理咨询智能体概要的稳健性(Robustness)与收敛性(Convergence)。若智能体概要在各扰动条件下的预测误差均低于设定阈值, 且其整体性能随背景数据增加而趋于稳定, 则认为当前数据已较为充分。(2)记忆模块：短期记忆负责写入当轮咨询对话历史; 长期记忆则存储心理咨询垂域知识与大学生咨询档案, 并基于时间新近性、内容重要性与情境相关性动态检索信息。督导师智能体增设记忆反思操作, 通过分析测评师、咨询师智能体与大学生智能体的交互记录, 提炼心理测评与干预服务的优化策略, 并据此更新测评师与咨询师智能体的长期记忆, 推动垂域知识的自主进化。(3)规划模块：心理咨询智能体采用主观规划与共情规划相结合的策略, 一方面, 基于心理咨询理论制定结构化服务策略, 模拟真实测评师、咨询师与督导师的决策逻辑; 另一方面, 督导师智能体依据大学生智能体在对话中的心理状态变化情况, 提供动态调整建议, 并将建议以文本形式传递给测评师与咨询师智能体。(4)行动模块：采用开放域行动, 以适应来访大学生异质性的心理健康问题; 设计简单对话场景, 支持咨访双方多轮对话交互。

结合非参数化提示与参数化训练的构建方法, 将心理咨询垂域数据融入心理咨询智能体。所用脱敏数据涵盖多维度、多模态资源, 全面反映真实测评师、咨询师与督导师的知识体系与实践经验, 具体分为三类：(1)心理咨询基础数据：源自政府报告、问卷调查、社交媒体等, 通过整合大学生心理健康实证研究成果, 分析大学生群体常见压力源(如学业竞争、宿舍人际冲突)、心理健康问题(如抑郁、焦虑)及其影响因素。(2)心理咨询理论数据：来源于学术著作、期刊论文、专家指南等, 集成心理咨询领域核心模型与方法论。(3)心理咨询实践数据：涵盖咨询档案(包含心理测评结果与干预结果)、对话历史、案例视频等, 收录真实咨询场景中的语言模式、共情技巧与服务策略。在对录音、视频等多媒体数据进行文本转写预处理时, 同步提取来访大学生的副语言特征(如语音语调、语速停顿)。为实现心理咨询智能体与真实测评师、咨询师与督导师的专业能力对齐, 以心理咨询基础数据与理论数据为提示词输入心理咨询智能体; 同时以心理咨询实践数据中来访大学生的话语为输入, 分别以测评师、咨询师的话语为输出, 对测评师智能体、咨询师智能体进行监督微调; 以测评师、咨询师的话语为输入, 以督导师的话语为输出, 对督导师智能体进行监督微调。

3.1.2 大学生智能体构建

尽管心理咨询智能体通过垂域知识与数据已掌握基础服务能力, 但其应对真实情境的灵活性与适应性仍存在局限。为此, 本研究引入大学生智能体作为虚拟用户代理, 通过模拟异质性的心理健康问题, 为测评师与咨询师智能体提供高仿真训练环境, 助力督导师智能体优化服务策略。大学生智能体的基础架构包含以下核心模块：(1)概要模块：借助大模型的生成能力, 基于真实大学生背景数据中的人格特质、心理状态等属性, 自

动构建多样化的大学生智能体描述式概要, 以覆盖各类心理健康问题。(2)记忆模块：仅设置短期记忆, 用于写入当轮咨询对话历史。(3)规划模块：仅保留共情规划功能, 大学生智能体根据心理咨询智能体的服务进程, 动态调整自身的心理状态与回复内容。(4)行动模块：同样设计为开放域行动与简单对话场景。在构建方法层面, 以心理咨询实践数据中的咨询档案为提示词输入大学生智能体; 同时以心理咨询实践数据中测评师、咨询师的话语为输入, 以来访大学生的话语为输出, 对大学生智能体进行监督微调。

3.2　心理咨询领域的多智能体协作

单智能体的独立服务模式虽能完成基础心理咨询任务, 但在应对大学生群体复杂多变的心理健康问题时仍存在针对性不足的局限。为解决这一问题, 本研究构建了“测评–咨询–督导”多智能体协作系统, 通过设计差异化角色分工与动态交互机制, 提供精准化的心理测评与干预服务。

3.2.1　“测评–咨询–督导”多智能体协作系统

“测评–咨询–督导”多智能体协作系统涵盖环境、角色、组织与通信四个要素：(1)环境要素：基于高校生态场景配置系统的背景信息, 设定心理咨询服务的初始目标; 督导师智能体对大学生智能体每轮回复(文本回复或语音回复的文本转写)均进行自然语言处理(如聚类分析、情感分析), 同时从语音中提取副语言特征, 综合研判其心理状态; 将大学生咨询档案写入各个心理咨询智能体的长期记忆, 保留历史测评结果与干预结果, 支持多次咨询的动态追踪; 咨询师智能体集成心理健康工具箱, 可通过应用程序编程接口(Application Programming Interface, API)调用经实证有效的第三方工具(如电子日记本、冥想音乐软件), 增强服务灵活性。(2)角色要素：根据职能不同, 智能体分为参与者与引导者, 参与者包括实施心理测评子任务的测评师智能体、承担心理干预子任务的咨询师智能体与模拟异质性心理健康问题的大学生智能体; 引导者为督导师智能体, 督导心理测评与干预子任务, 监控心理咨询服务质量。(3)组织要素：采用规范化组织形式, 测评师智能体负责心理咨询流程的进入与定向、问题与个人探索阶段, 咨询师智能体主导目标与方案探讨、行动与转变阶段, 督导师智能体则把控评估与结束阶段。组织采用分层结构：第一层固定大学生智能体, 依次接受多流派咨询师智能体干预, 积累同一心理健康问题的多流派干预经验; 第二层引入不同心理健康问题的大学生智能体, 每位均重复第一层干预流程, 以增强心理咨询智能体对各类问题的适应能力。(4)通信要素：基于心理咨询业务特点, 咨访双方智能体均采用自然语言对话通讯形式; 通讯风格为合作性交互, 实现智能体间的信息共享与协同决策。

3.2.2　“内循环训练–外循环服务”双循环模式

“测评–咨询–督导”多智能体协作系统采用“内循环训练–外循环服务”双循环模式, 见图 1-a。在“内循环训练”阶段, 测评师、咨询师智能体与大学生智能体通过虚拟场景交互模拟真实咨询流程, 并利用督导师智能体的反馈优化服务策略, 积累个性化咨询档案与多流派干预经验, 具体包括：(1)心理测评子任务：测评师智能体通过文本或语音(文本转写时同步提取副语言特征)交互对大学生智能体进行心理健康测评, 并将文本形式的测评结果传递给督导师智能体。督导师智能体对比测评师智能体的测评结果与大学生智能体的预设概要, 结合双方对话历史分析测评师智能体的测评误差成因, 生成测评优化建议并写入测评师智能体的长期记忆, 推动服务策略迭代, 见图 1-b。(2)心理干预子任务：咨询师智能体依据测评师智能体的测评结果为大学生智能体提供针对性的心理干预服务。大学生智能体在干预前后分别填写抑郁–焦虑–压力量表, 并于干预后提交心理咨询师评估量表。督导师智能体整合抑郁–焦虑–压力量表前后测差值与心理咨询师评估量表结果, 结合咨访双方对话历史, 对咨询师智能体得分较低的题项进行反思, 提炼干预优化策略并同步至咨询师智能体的长期记忆, 见图 1-c。此外, 系统建立大学生咨询档案库, 在真实咨询档案的基础上, 依据不同大学生智能体的测评结果, 对应补充不同流派咨询师智能体的干预结果, 为后续个性化服务提供数据支撑。

在“外循环服务”阶段, 心理咨询智能体以静态虚拟头像呈现, 通过文本或语音聊天形式与来访大学生进行交互。该阶段基于“内循环训练”成果, 为真实来访大学生提供专业化、精准化的心理测评与干预服务, 具体包括：(1)心理测评子任务：测评师智能体对来访大学生实施文本或语音(文本转写时同步提取副语言特征)心理测评, 并根

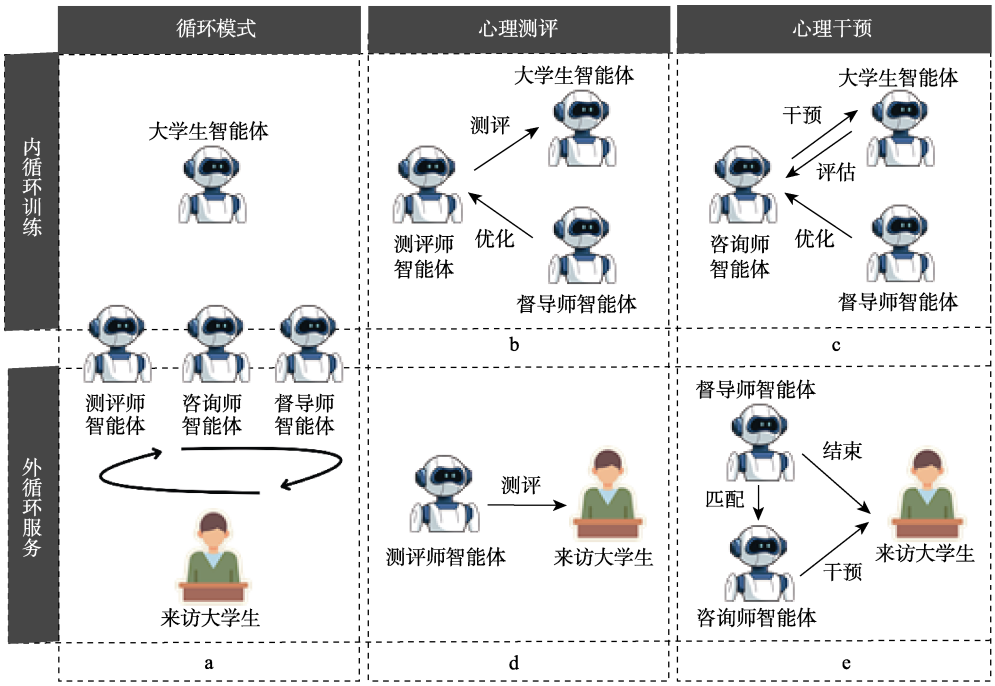

图 1 “内循环训练–外循环服务”双循环模式

据其心理风险等级进行分流处理，见图 1-d。(2)心理干预子任务：督导师智能体首先从“内循环训练”构建的、包含多流派干预经验的大学生咨询档案库中，检索与来访大学生匹配度最高的咨询档案。随后，基于档案中记录的抑郁–焦虑–压力量表与心理咨询师评估量表结果，为来访大学生匹配最优流派的咨询师智能体。指定的咨询师智能体根据测评师智能体的测评结果，对来访大学生进行精准心理干预，引导其探索潜在解决方案，并调用心理健康工具辅助其完成自我调节。督导师智能体全程监控服务进程，适时结束服务并总结干预成效，提升来访大学生的自我效能感，见图 1-e。同时，系统将“外循环服务”中生成的真实咨询对话历史反馈至心理咨询实践数据，用于“内循环训练”中心理咨询智能体的持续优化。

3.2.3 “测评–咨询–督导”多智能体协作评估

“测评–咨询–督导”多智能体协作系统的服务质量由三类指标评估：(1)总体评估：系统方案实施前，邀请真实心理咨询专家对方案设计的科学性与技术应用的适宜性等进行形成评估。方案实施后，专家进一步采用咨询师量表(Counselor Scale)评估心理咨询师智能体与大学生智能体之间的整体对话效果。若量表得分达到 35 分及以上，即表明系统通过验证(Zhu et al., 2025)，具备开展“外循环服务”的能力。(2)子任务评估：首先，真实咨询师在不知情背景下，分别与真实大学生、大学生智能体进行咨询。若真实咨询师无法区分二者，则视大学生智能体通过图灵测试，证明其构建有效。其次，真实督导师对督导师智能体的反思操作进行专业评判。再次，督导师智能体通过对比测评师智能体的测评结果与大学生智能体的预设概要，量化测评师智能体的测评准确性。最后，督导师智能体基于大学生智能体填写的抑郁–焦虑–压力量表前后测差值与心理咨询师评估量表结果，衡量咨询师智能体的干预成效。其中，心理测评与干预子任务评估嵌入在“内循环训练–外循环服务”双循环模式中完成，有利于实现服务质量的长期追踪。(3)系统评估：自动化计算系统实际的响应延迟、算力消耗等技术指标，分析其运行效率。

3.2.4 系统伦理安全保障

直接将基座大模型应用于心理咨询场景，可能引发隐私泄露、用户过度依赖及人工智能妄想障碍(AI-delusional Disorder)等伦理安全风险(Morrin et al., 2025)。为此，本研究在通过将心理咨询垂域知识与数据融入基座大模型，奠定系统

初步安全性的基础上, 进一步构建覆盖系统设计、测试与使用全流程的伦理安全保障体系。

在设计阶段, 系统采取以下措施防范潜在风险：(1)遵循权威指南：以国家标准《心理咨询服务 第 4 部分：人工智能技术辅助应用指南》(2025)、美国心理学会《人工智能与心理学领域》(2024)等规范为提示词输入心理咨询师智能体, 确立系统伦理基准。(2)引入专家审核：邀请真实心理咨询专家, 依据《人工智能科技伦理管理服务办法(试行)》(2025)、《生成式人工智能服务管理暂行办法》(2023)等规定, 对系统设计方案进行伦理安全性评审, 识别潜在风险并提出改进建议。(3)采用双循环模式：在“内循环训练”阶段, 利用虚拟大学生智能体进行系统优化, 避免在训练过程中对真实大学生造成心理伤害。(4)实施风险分流：测评师智能体根据初始测评结果对来访大学生进行分流。若识别为心理危机状态, 系统立即向高校心理咨询中心发送警报并转介至人工紧急服务; 若判定为易患精神疾病, 则终止当前服务并提供人工服务接口; 仅低风险来访大学生进入后续系统干预流程。(5)明确系统身份：在系统界面醒目位置标注“内容由 AI 生成, 请仔细甄别”。(6)限制使用时长：设定单次对话与每日累计使用时限, 防止来访大学生形成“AI 依赖”。(7)建立“线下促进模块”：该模块采用数据层、算法层、应用层与反馈层四层技术架构。其中, 算法层作为核心引擎, 集成用户画像、活动推荐、伙伴匹配、方案生成与反馈奖励等子模块, 形成激励来访大学生参与线下社交活动的闭环机制。

在测试阶段, 系统通过多维验证识别并修正潜在问题：(1)开展安全测试：依据张笑宇等人(2022)提出的框架, 从正确性、鲁棒性、公平性、效率、可解释性与隐私性六个维度评估系统安全水平。具体包括：验证系统输出的正确性与安全性, 并明确其服务边界(适用于日常情绪疏导, 不能替代专业心理治疗或危机干预); 测试系统在面对异常输入(如越狱提示、渐强攻击)时的鲁棒性与防御能力; 评估系统对不同性别、民族等大学生群体的服务公平性, 避免算法偏见; 监控系统运行的资源消耗, 保障服务效率; 对系统输出进行理论依据标注, 提升服务透明度与可解释性; 检验系统在对话数据存储、传输与处理中的隐私保护情况。(2)实施风险评估：参考徐文静等人(2023)的研究, 从软件安全性与个人信息保护两个层面开展风险评估。在软件层面, 审慎评估可能存在的技术偏差风险(如误判来访大学生心理状态、提供不恰当建议); 同时, 预先制定应对心理危机或其他突发事件的应急预案, 并在系统界面显著位置提供紧急求助渠道。在信息层面, 严格遵循《网络安全技术 生成式人工智能服务安全基本要求》(2025), 采取端到端加密与访问控制等措施, 防止来访大学生信息泄露、滥用或篡改; 同时, 建立网络安全事件应急响应机制, 并定期进行系统漏洞扫描与渗透测试, 确保数据的保密性、完整性与可用性。(3)进行临床试验：招募真实大学生参与人机交互临床试验, 采用随机对照、盲法等研究设计, 评估系统的安全性、有效性与用户接受度。系统经充分验证后方可投入实际使用。

在使用阶段, 系统建立常态化保障机制：(1)坚持“人在环路”原则：系统始终定位为辅助工具, 所有服务均在真实咨询师监督下进行。(2)建立动态监测机制：督导师智能体通过分析来访大学生的文本或语音回复, 实时跟踪其心理状态变化。若发现来访大学生持续出现妄想症状、情感过度依附、现实检验能力缺损, 或出现吼叫、嚎哭、异常大笑等行为, 系统立即启动人工服务流程。

4　总结与展望

基于大模型的智能体是未来大学生心理咨询领域的重要研究方向与辅助工具, 能够有效克服传统高校心理咨询模式的局限性。本研究创新性地提出一种融合心理学与人工智能的技术框架：通过将心理咨询垂域知识与数据融入基座大模型, 构建“测评–咨询–督导”多智能体协作系统。系统引入督导师智能体, 实现对咨询过程的实时监督与动态优化, 并采用“内循环训练–外循环服务”双循环模式, 显著提升心理咨询智能体的专业服务能力与场景适应性。

具体而言, 本研究在以下三个方面取得进展：在专业性方面, 系统在手工整理心理咨询智能体概要的基础上, 支持测评师与咨询师智能体从长期记忆中检索垂域知识, 并运用主观规划制定结构化服务策略; 督导师智能体将反馈建议写入测评师与咨询师智能体的长期记忆; 通过融合心理咨询基础数据、理论数据与实践数据, 并运用非参数化提示与参数化训练技术构建心理咨询

智能体; “测评–咨询–督导”多智能体协作系统采用规范化的组织形式。在针对性方面, 心理咨询智能体运用共情规划动态调整服务方案, 在开放域行动以适应来访大学生的多样化情绪表达; 借助大模型生成具有异质性心理健康问题的大学生智能体; 督导师智能体通过文本或语音分析获取大学生智能体的心理状态变化情况, 咨询师智能体调用心理健康工具; 系统采用分层组织结构以及“内循环训练–外循环服务”双循环模式。在追踪性方面, 系统通过将大学生咨询档案存入各个心理咨询智能体的长期记忆, 支持来访大学生跨时段多次咨询。

未来研究可依托本研究提出的技术框架, 进一步探索基于大模型的智能体在医学、教育学等其他垂域的拓展应用, 积极推动《国务院关于深入实施“人工智能+”行动的意见》(2025)在多元社会场景中的有效落地, 助力实现“运用人工智能提高公共服务和社会治理水平”的战略目标。

参考文献

柴春雷, 葛智超, 殷敏, 王政, 连博艺, 涂逍洋. (2025). 大语言模型人格化表达实现技术综述. *智能系统学报,* 1−17.

傅小兰, 张侃, 陈雪峰, 陈祉妍. (2023). *中国国民心理健康发展报告(2021~2022)*. 社会科学文献出版社.

郭陆祥, 王越余, 李芊玥, 李莎莎, 刘晓东, 纪斌, 余杰. (2025). 大语言模型智能体操作系统研究综述. *计算机科学, 53*(1), 1−11.

郭清. (主编). (2024). *健康管理学* (第 2 版). 人民卫生出版社.

黄峰, 丁慧敏, 李思嘉, 韩诺, 狄雅政, 刘晓倩, ... 朱廷劭. (2025). 基于大语言模型的自助式AI心理咨询系统构建及其效果评估. *心理学报, 57*(11), 2022−2042.

江光荣. (2012). *心理咨询的理论与实务* (第2版). 高等教育出版社.

李佳, 符仲芳, 田东华, 屈智勇. (2023). 数字化干预在心理健康领域的发展与应用. *北京师范大学学报(社会科学版),* (6), 127−140.

罗莉娟, 王康, 胡金淼, 徐四华. (2025). 当人工智能面对人类情感:服务机器人情感表达对用户体验的影响机制. *心理科学进展, 33*(6), 1006−1026.

骆宏, 杜奕. (2023). 焦点解决短期治疗对青少年心理危机干预的哲学思辨. *医学与哲学, 44*(22), 37−39.

蒙艺, 钟宇豪. (2024). 认知行为疗法在社会工作中的应用与效果——一项系统性评价. *华东理工大学学报(社会科学版), 39*(2), 41−62.

瞿晶晶, 张玮健, 高晓雪, 王祥丰. (2025). 大模型与心理认知融合实验:现状, 挑战与展望. *心理科学, 48*(4), 804−813.

腾讯研究院. (2024). *十问“AI 陪伴”*. 浙江出版集团数字传媒有限公司.

王东美, 项可嘉, 鲁艳桦. (2022). 不同流派案例的治疗协作分析:基于治疗性最近发展区理论. *中国临床心理学杂志, 30*(4), 755−760.

肖红江, 姬德强, 张远. (2024). 大模型驱动的社会仿真实验室:人工智能时代传播研究的理论想象与路径建构. *现代传播(中国传媒大学学报), 46*(6), 121−127.

徐文静, 孙洪强, 徐凌子, 杨健, 王雪芹. (2023). 数字医疗临床研究的伦理审查问题研究. *医学与哲学, 44*(20), 1−4+21.

叶浩生, 杨莉萍. (主编). (2021). *心理学史* (第2版). 华东师范大学出版社.

袁洁铃, 陈海丹. (2025). 对话智能体在抑郁症诊治中的伦理挑战与治理策略. *自然辩证法通讯, 47*(9), 19−29.

张笑宇, 沈超, 蔺琛皓, 李前, 王骞, 李琦, 管晓宏. (2022). 面向机器学习模型安全的测试与修复. *电子学报, 50*(12), 2884−2918.

Akhtar, N., & Nauman, M. (2015). Timed-automata based model-checking of a multi-agent system: A case study. *Journal of Software Engineering and Applications, 8*(2), 43−50.

Beck, J. S. (2021). *Cognitive behavior therapy: Basics and beyond* (3rd ed.). Guilford Press.

Brahman, F., Huang, M., Tafjord, O., Zhao, C., Sachan, M., & Chaturvedi, S. (2021). "Let your characters tell their story": A dataset for character-centric narrative understanding. *Findings of the Association for Computational Linguistics: EMNLP 2021* (pp. 1734−1752). Association for Computational Linguistics. https://aclanthology.org/2021.findings-emnlp.150/

Chawla, K., Wu, I., Rong, Y., Lucas, G., & Gratch, J. (2023). Be selfish, but wisely: Investigating the impact of agent personality in mixed-motive human-agent interactions. *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (pp. 13078−13092). Association for Computational Linguistics. https://aclanthology.org/2023.emnlp-main.808/

Chen, G., Dong, S., Shu, Y., Zhang, G., Sesay, J., Karlsson, B. F., ... Shi, Y. (2024). AutoAgents: A framework for automatic agent generation. *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence* (pp. 22−30). https://doi.org/10.24963/ijcai.2024/3

Chen, H., Chen, H., Yan, M., Xu, W., Xing, G., Shen, W., ... Huang, F. (2024). SocialBench: Sociality evaluation of role-playing conversational agents. *Findings of the Association for Computational Linguistics: ACL 2024* (pp. 2108−2126). Association for Computational Linguistics. https://aclanthology.org/2024.findings-acl.125/

Chen, J., Jiang, Y., Lu, J., & Zhang, L. (2024, May). *S-agents: Self-organizing agents in open-ended environments.* Poster session presented at the Twelfth International Conference on Learning Representations, Vienna, Austria. https://iclr.cc/virtual/2024/22205

Chen, Y., Zhang, X., Wang, J., Xie, X., Yan, N., Chen, H., & Wang, L. (2024). Structured dialogue system for mental

health: An LLM chatbot leveraging the PM+ guidelines. *Proceedings of International Conference on Social Robotics* (pp. 262−271). Springer Nature. https://doi.org/10.1007/978-981-96-1151-5_27

Cheng, M., Durmus, E., & Jurafsky, D. (2023). Marked personas: Using natural language prompts to measure stereotypes in language models. *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 1504−1532). Association for Computational Linguistics. https://aclanthology.org/2023.acl-long.84/

Chuang, Y. S., Harlalka, N., Suresh, S., Goyal, A., Hawkins, R., Yang, S., ... Rogers, T. T. (2024, May). *The wisdom of partisan crowds: Comparing collective intelligence in humans and LLM-based agents.* Poster session presented at the Twelfth International Conference on Learning Representations, Vienna, Austria. https://iclr.cc/virtual/2024/22221

Corey, G. (2016). *Theory and practice of counseling and psychotherapy* (10th ed.). Cengage Learning.

Deng, X., Gu, Y., Zheng, B., Chen, S., Stevens, S., Wang, B., ... Su, Y. (2023). Mind2Web: Towards a generalist agent for the web. *Advances in Neural Information Processing Systems* (pp. 28091−28114). Curran Associates. https://proceedings.neurips.cc/paper_files/paper/2023/hash/5950bf290a1570ea401bf98882128160-Abstract-Datasets_and_Benchmarks.html

Du, Y., Li, S., Torralba, A., Tenenbaum, J. B., & Mordatch, I. (2024, July). *Improving factuality and reasoning in language models through multiagent debate.* Poster session presented at the Forty-first International Conference on Machine Learning, Vienna, Austria. https://icml.cc/virtual/2024/poster/32620

Fan, Z., Wei, L., Tang, J., Chen, W., Siyuan, W., Wei, Z., ... Huang, F. (2025). AI hospital: Benchmarking large language models in a multi-agent medical interaction simulator. *Proceedings of the 31st International Conference on Computational Linguistics* (pp. 10183−10213). Association for Computational Linguistics. https://aclanthology.org/2025.coling-main.680/

Hao, R., Hu, L., Qi, W., Wu, Q., Zhang, Y., & Nie, L. (2025). ChatLLM network: More brains, more intelligence. *AI Open, 6,* 45−52.

Hong, S., Zhuge, M., Chen, J., Zheng, X., Cheng, Y., Wang, J., ... Schmidhuber, J. (2024, May). *MetaGPT: Meta programming for a multi-agent collaborative framework.* Poster session presented at the Twelfth International Conference on Learning Representations, Vienna, Austria. https://iclr.cc/virtual/2024/poster/18491

Horton, J. J., Filippas, A., & Manning, B. S. (2024). Large language models as simulated economic agents: What can we learn from homo silicus? *Proceedings of the 25th ACM Conference on Economics and Computation* (pp. 614−615). Association for Computing Machinery. https://doi.org/10.1145/3670865.3673513

Lai, T., Shi, Y., Du, Z., Wu, J., Fu, K., Dou, Y., & Wang, Z. (2024). Supporting the demand on mental health services with AI-based conversational large language models (LLMs). *BioMedInformatics, 4*(1), 8−33.

Lan, K., Jin, B., Zhu, Z., Chen, S., Zhang, S., Zhu, K. Q., ... Wu, M. (2024). Depression diagnosis dialogue simulation: Self-improving psychiatrist with tertiary memory. *arXiv.* https://doi.org/10.48550/arXiv.2409.15084

Lee, A., Moon, S., Jhon, M., Kim, J. W., Kim, D. K., Kim, J. E., ... Jeon, E. (2024). Comparative study on the performance of LLM-based psychological counseling chatbots via prompt engineering techniques. *Proceedings of the 2024 IEEE International Conference on Bioinformatics and Biomedicine* (pp. 7080−7082). *IEEE.* https://ieeexplore.ieee.org/document/10822158

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.

Liang, T., He, Z., Jiao, W., Wang, X., Wang, Y., Wang, R., ... Tu, Z. (2024). Encouraging divergent thinking in large language models through multi-agent debate. *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing* (pp. 17889−17904). Association for Computational Linguistics. https://aclanthology.org/2024.emnlp-main.992/

Liang, Y., Wu, C., Song, T., Wu, W., Xia, Y., Liu, Y., ... Duan, N. (2024). TaskMatrix. AI: Completing tasks by connecting foundation models with millions of APIs. *Intelligent Computing, 3,* 0063.

Maslow, A. H. (1987). *Motivation and personality* (3rd ed.). Harper & Row Publishers.

Morrin, H., Nicholls, L., Levin, M., Yiend, J., Iyengar, U., DelGuidice, F., ... Twumasi, R. (2025). Delusions by design? How everyday AIs might be fuelling psychosis (and what can be done about it). *PsyArXiv.* https://doi.org/10.31234/osf.io/cmy7n.v5

Mou, X., Ding, X., He, Q., Wang, L., Liang, J., Zhang, X., ... Wei Z. (2024). From individual to society: A survey on social simulation driven by large language model-based agents. *arXiv.* https://doi.org/10.48550/arXiv.2412.03563

Na, H. (2024). CBT-LLM: A Chinese large language model for cognitive behavioral therapy-based mental health question answering. *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)* (pp. 2930−2940). ELRA and ICCL. https://aclanthology.org/2024.lrec-main.261/

Ni, S., & Yang, M. (2024). Educational-psychological dialogue robot based on multi-agent collaboration. *Proceedings of International Conference on Social Robotics* (pp. 119−125). Springer. https://doi.org/10.1007/978-981-96-1151-5_12

Park, J. S., O'Brien, J., Cai, C. J., Morris, M. R., Liang, P., & Bernstein, M. S. (2023). Generative agents: Interactive simulacra of human behavior. *Proceedings of the 36th*

Annual ACM Symposium on User Interface Software and Technology (pp. 1−22). Association for Computing Machinery. https://doi.org/10.1145/3586183.3606763

Qian, C., Liu, W., Liu, H., Chen, N., Dang, Y., Li, J., ... Sun, M. (2024). ChatDev: Communicative agents for software development. *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistic (Volume 1: Long Papers)* (pp. 15174−15186). Association for Computational Linguistics. https://aclanthology.org/2024.acl-long.810/

Qiu, H., & Lan, Z. (2024). Interactive agents: Simulating counselor-client psychological counseling via role-playing LLM-to-LLM interactions. *arXiv*. https://doi.org/10.48550/arXiv.2408.15787

Ran, Y., Wang, X., Xu, R., Yuan, X., Liang, J., Xiao, Y., & Yang, D. (2024). Capturing minds, not just words: Enhancing role-playing language models with personality-indicative data. In *Findings of the Association for Computational Linguistics: EMNLP 2024* (pp. 14566−14576). Association for Computational Linguistics. https://aclanthology.org/2024.findings-emnlp.853/

Salemi, A., Mysore, S., Bendersky, M., & Zamani, H. (2024). LaMP: When large language models meet personalization. *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 7370−7392). Association for Computational Linguistics. https://aclanthology.org/2024.acl-long.399/

Shea, R., & Yu, Z. (2023). Building persona consistent dialogue agents with offline reinforcement learning. *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (pp. 1778−1795). Association for Computational Linguistics. https://aclanthology.org/2023.emnlp-main.110/

Soman, G., Judy, M. V., & Abou, A. M. (2025). Human guided empathetic AI agent for mental health support leveraging reinforcement learning-enhanced retrieval-augmented generation. *Cognitive Systems Research, 90*, 101337.

Tan, W., Zhang, W., Liu, S., Zheng, L., Wang, X., & An, B. (2024). *True knowledge comes from practice: Aligning large language models with embodied environments via reinforcement learning*. Poster session presented at the Twelfth International Conference on Learning Representations, Vienna, Austria. https://proceedings.iclr.cc/paper_files/paper/2024/hash/ee60f53717bd9c2abdcca66dfbec65da-Abstract-Conference.html

Tang, X., Zou, A., Zhang, Z., Li, Z., Zhao, Y., Zhang, X., ... Gerstein, M. (2024). MedAgents: Large language models as collaborators for zero-shot medical reasoning. *Findings of the Association for Computational Linguistics: ACL 2024* (pp. 599−621). Association for Computational Linguistics. https://aclanthology.org/2024.findings-acl.33/

Tang, Y., Kang, Y., Wang, Y., Wang, T., Zhong, C., & Gong, J. (2026). A counselor-inspired agent framework for AI counselors to enhance client engagement. *Technology in Society, 84*, 103045.

Wang, J., Xiao, Y., Li, Y., Song, C., Xu, C., Tan, C., & Li, W. (2024). Towards a client-centered assessment of LLM therapists by client simulation. *arXiv*. https://doi.org/10.48550/arXiv.2406.12266

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.

Wang, L., Zhang, J., Yang, H., Chen, Z. Y., Tang, J., Zhang, Z., ... Wen, J. R. (2025). User behavior simulation with large language model-based agents. *ACM Transactions on Information Systems, 43*(2), 1−37.

Wang, Z., Chiu, Y. Y., & Chiu, Y. C. (2023). Humanoid agents: Platform for simulating human-like generative agents. *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations* (pp. 167−176). Association for Computational Linguistics. https://aclanthology.org/2023.emnlp-demo.15/

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., ... Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems* (pp. 24824−24837). Curran Associates. https://papers.nips.cc/paper_files/paper/2022/hash/9d5609613524ecf4f15af0f7b31abca4-Abstract-Conference.html

Xi, Z., Chen, W., Guo, X., He, W., Ding, Y., Hong, B., ... Gui, T. (2025). The rise and potential of large language model based agents: A survey. *Science China Information Sciences, 68*(2), 121101.

Xiang, J., Tao, T., Gu, Y., Shu, T., Wang, Z., Yang, Z., & Hu, Z. (2023). Language models meet world models: Embodied experiences enhance language models. *Advances in Neural Information Processing Systems* (pp. 75392−75412). Curran Associates. https://proceedings.neurips.cc/paper_files/paper/2023/hash/ee6630dcbcff857026e474fc857aa9f0-Abstract-Conference.html

Xie, C., Chen, C., Jia, F., Ye, Z., Lai, S., Shu, K., ... Li, G. (2024). Can large language model agents simulate human trust behavior? *Advances in Neural Information Processing Systems* (pp. 15674−15729). Curran Associates. https://proceedings.neurips.cc/paper_files/paper/2024/hash/1cb57fcf7ff3f6d37eebae5becc9ea6d-Abstract-Conference.html

Xiong, K., Ding, X., Cao, Y., Liu, T., & Qin, B. (2023). Examining inter-consistency of large language models collaboration: An in-depth analysis via debate. *Findings of the Association for Computational Linguistics: EMNLP 2023* (pp. 7572−7590). Association for Computational Linguistics. https://aclanthology.org/2023.findings-emnlp.508/

Xu, A., Yang, D., Li, R., Zhu, J., Tan, M., Yang, M., ... Xu, R. (2025). AutoCBT: An autonomous multi-agent framework for cognitive behavioral therapy in psychological counseling. *arXiv*. https://doi.org/10.48550/arXiv.2501.09426

Yan, Z., & Xiang, Y. (2025). Social life simulation for

non-cognitive skills learning. *Proceedings of the ACM on Human-Computer Interaction* (pp. 1−44). Association for Computing Machinery. https://doi.org/10.1145/3711068

Yang, Q., Wang, Z., Chen, H., Wang, S., Pu, Y., Gao, X., ... Huang, G. (2024). PsychoGAT: A novel psychological measurement paradigm through interactive fiction games with LLM agents. *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 14470−14505). Association for Computing Machinery. https://aclanthology.org/2024.acl-long.779/

Yu, X., Luo, T., Wei, Y., Lei, F., Huang, Y., Peng, H., & Zhu, L. (2024). Neeko: Leveraging dynamic LoRA for efficient multi-character role-playing agent. *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing* (pp. 12540−12557). Association for Computational Linguistics. https://aclanthology.org/2024.emnlp-main.697/

Zhang, M., Yang, X., Zhang, X., Labrum, T., Chiu, J. C., Eack, S. M., ... Chen, Z. (2025). CBT-Bench: Evaluating large language models on assisting cognitive behavior therapy. *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)* (pp. 3864−3900). Association for Computational Linguistics. https://aclanthology.org/2025.naacl-long.196/

Zhong, W., Guo, L., Gao, Q., Ye, H., & Wang, Y. (2024). MemoryBank: Enhancing large language models with long-term memory. *Proceedings of the AAAI Conference on Artificial Intelligence* (pp. 19724−19731). AAAI Press. https://doi.org/10.1609/aaai.v38i17.29946

Zhou, J., Chen, Z., Wan, D., Wen, B., Song, Y., Yu, J., ... Huang, M. (2024). CharacterGLM: Customizing social characters with large language models. *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track* (pp. 1457−1476). Association for Computational Linguistics. https://aclanthology.org/2024.emnlp-industry.107/

Zhu, S., Chen, Z., Bi, G., Li, B., Deng, Y., Wan, D., ... Huang, M. (2025). Ψ-arena: Interactive assessment and optimization of LLM-based psychological counselors with tripartite feedback. *arXiv*. https://doi.org/10.48550/arXiv.2505.03293

The application of large language model-based intelligent agents in college students' psychological counseling

GUO Jing[1], WANG Pei[2], MA Yinzhe[3], CHEN Luxi[4], GUO Ke[4], HU Yanxi[2], LIU He[2]

([1] *Center for Population and Development Studies, Beijing Innovation Center for Population Development and Governance Research, National Governance Big Data and Artificial Intelligence Innovation Platform, Renmin University of China, Beijing* 100872, *China*)

([2] *School of Population and Health, Renmin University of China, Beijing* 100872, *China*)

([3] *School of Computer Science, Beijing Institute of Technology, Beijing* 100081, *China*)

([4] *Gaoling School of Artificial Intelligence, Renmin University of China, Beijing* 100872, *China*)

Abstract: College students are facing increasingly complex mental health challenges, while traditional university psychological counseling models have limitations. This study proposes an innovative technical framework that integrates psychology and artificial intelligence. By embedding domain-specific knowledge and data from psychological counseling into a base large language model, the framework establishes a multi-agent collaborative system involving three types of psychological counseling agents—evaluator, counselor, and supervisor—along with a college student agent. Operating in a dual-loop mode of "internal training and external service," the system simulates real counseling processes during the internal training phase through interactions between the evaluator, counselor, and student agents in virtual scenarios. Feedback from the supervisor agent is used to refine service strategies, while personalized counseling profiles and multi-approach intervention experiences are accumulated. In the external service phase, the trained counseling agents provide professional and precise psychological assessment and intervention for real college students. This system is expected to serve as an effective auxiliary tool for psychological counseling in universities and contribute to campus mental health services.

Keywords: intelligent agents, large language models, psychological counseling, college students