《心理科学进展》审稿意见与作者回应

题目：基于大模型的智能体在大学生心理咨询中的应用

作者：郭静，王沛，马胤哲，陈路晰，郭可，胡彦熙，刘荷

第一轮

审稿人 1 意见：

依照心理科学进展投稿指南对文章类型的相关介绍，本文《基于大模型的智能体在大学生心理咨询中的应用》应该是一篇研究构想类文章。文章针对当前高校心理咨询工作中存在的时空限制、服务同质化、追踪机制缺失等局限，提议基于大模型构建“评估-咨询-督导”多智能体协作系统，并提出了“内循环训练-外循环服务”的双循环模式。文章通过有关文献介绍了基于大模型的智能体技术框架，强调了心理咨询垂域知识的重要性，并详细阐述了如何在大学生心理咨询中构建多智能体协作系统。文章具有一定的创新性，有助于推动高校心理健康服务体系的智能化转型。

为了更好地推动心理学与人工智能的相互促进、融合发展，建议作者在以下方面对文章加以修改：

意见 1：

文章简要介绍了基于大模型的智能体技术，建议在此基础上针对大模型应用于心理学尤其是心理咨询领域的实证研究进行文献综述，以体现该技术的应用特点，并验证其有效性。

回应：

根据审稿专家的意见，我们已增补基于大模型的智能体技术在心理咨询领域应用的研究综述。

主要包括：

（1）在“2.3 智能体技术在心理咨询领域的应用现状”中，新增文献综述部分，重点从来访者构建、心理测评或干预以及心理咨询效果评估三个维度梳理现有研究，为本研究提供实证依据。

（2）在“1 引言”中，系统分析传统高校线下咨询、现有在线咨询平台以及心理咨询智能体的局限性，进而体现本研究特点。

意见 2：

文中部分论述缺乏文献支持，建议引用相关研究加强理论构建。例如，文中提到大模型“仍存在生成方向散漫、上下文遗忘、目标驱动力不足等瓶颈”，“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，可有效提升复杂任务的解决效能。”类似描述需要相关文献支持。

回应：

根据审稿专家的意见，我们已为关键结论性论述补充必要的参考文献。

主要包括：

（1）在“2.1.1 单智能体构建”中，由于相关表述已删除，未补充参考文献支持“仍存在生成方向散漫、上下文遗忘、目标驱动力不足等瓶颈”；补充参考文献《A Survey on Large Language Model Based Autonomous Agents》、《From Individual to Society: A Survey on Social Simulation Driven by Large Language Model-based Agents》，支持“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，单智能体能够有效模拟特定个体”；补充参考文献《一本书读懂 AI Agent：技术、应用与商业》，支持“为精准模拟真实个体行为，构建能够准确复制个体特征的智能体架构至关重要，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”；补充参考文献《AI Agent 应用与项目实战》，支持“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”。

（2）在“2.1.2 多智能体协作”中，补充参考文献《动手做 AI Agent》，支持“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，可有效提升复杂任务的解决效能”。

意见 3：

评估一词易产生歧义。评估师在多智能体协作系统中承担的是心理测评的子任务，“评估”的对象是来访者。然而，文中 3.3 同时还介绍了有关心理咨询评估的概念，这里“评估”

的对象是心理咨询的质量。二者概念不同，使用相同的名称不利于读者区分概念，容易产生歧义。

回应：

根据审稿专家的意见，我们已严格区分不同语境下的“评估”概念。

主要包括：

（1）在全文中，将评估师智能体、“评估-咨询-督导”等表述统一修正为测量师智能体、“测量-咨询-督导”等。

（2）在“2.2 心理咨询垂域知识”中，将心理咨询评估相关表述统一修正为心理咨询效果评估。

意见 4：

有关智能体构建方面，建议作者提供更多的技术细节，以说明其必要性和可行性。

构建通用心理咨询智能体是否可行？作者提到“基于概要、记忆、规划与行动四个模块，构建具备跨领域知识储备与专业化交互能力的通用心理咨询智能体”。构建通用智能体需要大量的数据、人力以及硬件支持，请问作者如何保证数据的来源、GPU 等硬件的支持？成本问题是否在作者的考虑范畴之内？既然通用智能体可以通过非参数化提示以及微调等训练方法实现垂域化，是否有必要构建通用智能体？

回应：

根据审稿专家的意见，我们已补充技术细节，并澄清了通用智能体的表述。

主要包括：

（1）在“2 技术框架与知识融入”中，补充可供选择的大模型示例。

（2）在“3.1.1 心理咨询智能体构建”、“3.1.2 大学生智能体构建”中，细化对应智能体的构建方法。

（3）在“2.1.1 单智能体构建”、“3.1.1 心理咨询智能体构建”、“3.1.2 大学生智能体构建”中，原文通用心理咨询智能体意指基于四个模块构建的智能体，与耗费大量人力、物力、财力建立的智能体相区分。为避免误解，删除通用智能体相关表述，仅从智能体的基础架构与构建方法角度阐述。

意见 5：

如何对通用心理咨询智能体进行监督微调（注：监督微调的英文是 supervised

fine-tuning，不是 scalable fine-tuning）？作者提到“心理咨询智能体以心理咨询基础数据、理论数据与实践数据为训练标签，通过监督微调技术优化模型参数”。能否详细介绍如何对心理咨询基础数据、理论数据与实践数据进行打码，以实现监督微调？

回应：

根据审稿专家的意见，我们已细化心理咨询智能体的构建方法。

主要包括：

（1）在“2.1.1 单智能体构建”中，更正监督微调的英文为 Supervised Fine-tuning。

（2）在“3.1.1 心理咨询智能体构建”、“4 总结与展望”中，进一步明确构建方法为非参数化提示与参数化训练相结合。非参数化提示技术应用于心理咨询基础数据与理论数据；参数化训练技术应用于心理咨询实践数据，对测量师、咨询师与督导师智能体进行监督微调，数据打码方式详见正文。

意见 6：

如何确保大学生智能体能够提供高仿真训练环境？越来越多的研究正在使用大模型合成数据作为虚拟用户代理，但是其有效性并没有得到广泛的验证。作者提到智能体参与者需要完成心理测评子任务，是否有研究证实智能体可以参照异质性心理健康问题提供具有充分信效度的心理测评数据？请问作者计划如何设计大学生智能体，并保证其模拟数据的灵活性与适应性？

回应：

根据审稿专家的意见，我们已阐述大学生智能体的有效性。

（1）在“2.1.1 单智能体构建”中，补充说明其基础架构（人类行为的理论抽象）与构建方法（基于数据对齐真实个体）能够共同保障大学生智能体的高保真度。

（2）在“3.1.2 大学生智能体构建”中，删除大模型合成数据表述。补充参考文献《PSYCHE: A Multi-faceted Patient Simulation Framework for Evaluation of Psychiatric Assessment Conversational Agents》、《ProAI: Proactive Multi-Agent Conversational AI with Structured Knowledge Base for Psychiatric Diagnosis》，支持“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，为心理咨询智能体提供高仿真训练环境，从而驱动服务策略的迭代优化”。这些研究证实精神疾病患者智能体能够提供高临床相关性、定量化、且具有充分信效度的心理测评数据。

（3）通过定制四个模块（概要、记忆、规划、行动）的基础架构以及采用参数化训练的构

建方法设计大学生智能体。通过注入覆盖各类心理健康问题的心理咨询实践数据，提升系统灵活性；通过设置拟人化记忆机制与规划模块，提升系统适应性。

意见 7：

请作者具体说明关于短期记忆和长期记忆的设置依据。作者提到大学生智能体“仅保留短期记忆，按时间顺序写入咨询对话历史，建立个性化咨询档案以支持多次咨询追踪。”请问短期记忆是否能够完整保存所需的心理咨询历史，实现多次咨询追踪的目标？短期记忆和长期记忆如何界定，是否会受到智能体上下文窗口的制约？

回应：

根据审稿专家的意见，我们已补充对记忆机制的说明。在“2.1.1 单智能体构建”、“3.1.2 大学生智能体构建”、“4 总结与展望”中，明确短期记忆记录即时互动信息，无法完整保存历次咨询对话历史；长期记忆存储持久全局信息，防止目标偏离。将大学生咨询档案存入各个心理咨询智能体的长期记忆，供智能体在当次咨询中动态检索，支持多次咨询追踪。补充记忆模块的设计旨在克服大模型上下文窗口的制约，模拟人类记忆。

意见 8：

在多智能体协作要素中，作者提到智能体之间的通讯风格为合作性交互，能否解释何为合作性交互，以及如何实现合作性交互？

回应：

根据审稿专家的意见，我们已补充合作性交互的说明。

主要包括：

（1）在“2.1.2 多智能体协作”中，补充合作性交互与竞争性交互的定义。

（2）本研究合作性交互的实现方式为通过函数封装各个心理咨询智能体的多种行为，在主函数中依序调用，推动多智能体协作完成任务。

意见 9：

鉴于智能体在大学生心理咨询中的应用较为欠缺，现阶段无法保证智能体能够完成文中预设的所有任务。建议作者对双循环模式提供更多理论或者实证证据，以证明该模式的有效性。

在心理测评子任务中，作者提到“督导师智能体通过对比大学生智能体的预设概要与测

评结果生成优化建议”，如果测评结果与大学生智能体的预设概要存在差异，该结果是否可能源自大学生智能体的自身缺陷？

回应：

根据审稿专家的意见，我们已补充实证证据，并修正了大学生智能体的构建方法。

主要包括：

（1）在“2.3 智能体技术在心理咨询领域的应用现状”中,新增文献综述，旨在证明智能体完成预设任务（测评、干预、效果评估）的可行性。

（2）在“3.2.2 ‘内循环训练-外循环服务’双循环模式”中，补充参考文献《Agent Hospital: A Simulacrum of Hospital with Evolvable Medical Agents》，支持“‘测量-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”。该研究证明，医生智能体通过内循环（治疗大量患者智能体）实现自主进化，其能力超越 SOTA 模型，为“外循环服务”（服务真实患者）奠定基础。

（3）在“3.1.2 大学生智能体构建”、“4 总结与展望”中，将大学生智能体的构建方法修正为参数化训练技术。因为仅靠提示词工程预设大学生智能体的概要可能会与大模型内部设定冲突，使大学生智能体存在缺陷；而监督微调能够更有效地利用数据，减少模型本身限制。

（4）在“3.2.2 ‘内循环训练-外循环服务’双循环模式”中，补充参考文献《Depression Diagnosis Dialogue Simulation: Self-improving Psychiatrist with Tertiary Memory》，支持“督导师智能体通过对比测评结果与大学生智能体的预设概要生成优化建议，动态更新测量师智能体的长期记忆，推动服务策略迭代”。该研究利用监督者智能体比较诊断结果与患者初始概要，生成反思存储至精神科医生智能体的长期记忆以优化后续决策，消融实验证明了此机制的关键作用。

意见 10：

在心理干预子任务中，作者提到“大学生智能体在干预前后分别填写心理咨询效果评估量表，并于干预后提交心理咨询师评估量表”。请问作者如何保证大学生智能体能够准确地对心理咨询进行评估？有无实证研究确认智能体的量表结果可以准确反映心理咨询的效果？

回应：

根据审稿专家的意见，我们已阐述心理咨询效果评估的有效性。

主要包括：

（1）高保真大学生智能体的建立（通过基础架构与构建方法）是准确评估的基础。

（2）在“3.2.2 ‘内循环训练-外循环服务’双循环模式”中，补充参考文献《Towards a Client-Centered Assessment of LLM Therapists by Client Simulation》，支持“大学生智能体在干预前后分别填写心理咨询效果评估量表，并于干预后提交心理咨询师评估量表”。该研究证实来访者智能体的量表结果在核心维度（会话效果、治疗联盟）上，能够有效评估心理咨询效果。

意见 11：

“大学生咨询档案库”可以视作多智能体协作系统的核心。为了真实反映大学生的心理问题以及应对方案，该档案库是否应该更多地参考现实生活中的大学生心理咨询档案，而不是基于大学生智能体在内循环训练中生成的数据？

回应：

根据审稿专家的意见，我们已细化大学生咨询档案库的设计。

主要包括：

（1）在“3.1.1 心理咨询智能体构建”中，将真实咨询档案纳入心理咨询实践数据。

（2）在“3.2.2 ‘内循环训练-外循环服务’双循环模式”中，明确在真实咨询档案的基础上，构建大学生咨询档案库。

意见 12：

文中提到内循环训练为外循环服务提供了基础。建议作者进一步思考二者是否可以相互促进。例如，外循环服务的真实数据与反馈结果可以为内循环提供心理咨询的实践数据，辅助智能体进一步的优化与微调。

回应：

根据审稿专家的意见，我们已更新双循环模式的相互促进作用。在“3.2.2 ‘内循环训练-外循环服务’双循环模式”中，补充将“外循环服务”产生的真实数据反馈至“内循环训练”，作为心理咨询实践数据用于智能体的持续优化。

意见 13：

文中提到多智能体协作系统只能“在封闭域内执行规范的咨询行动”，但是大学生智能体“通过开放域行动模拟其在真实咨询场景中的多样化情绪表达”。请问作者计划如何设置

封闭域，并如何保证封闭域内的咨询行动可以解决开放域的真实咨询场景？

回应：

根据审稿专家的意见，我们已更正心理咨询智能体的行动域设定。在“3.1.1 心理咨询智能体构建”、“4 总结与展望”中，将心理咨询智能体行动域由封闭域更正为开放域，因为封闭域无法有效应对来访大学生异质性心理健康问题的复杂性。

意见 14：

作者提到“构建方法层面，大学生智能体以心理咨询基础数据为标签进行参数化训练”，以及“利用心理咨询基础数据、理论数据与实践数据参数化训练心理咨询智能体”。请问作者打算从何处获取心理咨询基础数据？是否考虑过有关个人隐私等伦理问题？

回应：

根据审稿专家的意见，我们已明确数据来源并强调伦理处理。

主要包括：

（1）大学生智能体使用的心理咨询基础数据与心理咨询智能体来源相同，均源自政府报告、问卷调查、社交媒体等。

（2）在“3.1.1 心理咨询智能体构建”、“4 总结与展望”中，补充对真实心理咨询垂域数据进行脱敏预处理，强调未来研究可着眼于构建心理咨询智能体专属伦理框架，以保护来访者隐私，确保符合伦理规范。

……

审稿人 2 意见：

意见 1：

引言强调 AI 的自主性、反应性和社会能力，但对大学生特定心理健康问题的讨论（如焦虑、抑郁的流行率）较为笼统，可能削弱研究针对性的说服力。现有 AI 在心理咨询中的应用（如聊天机器人）未详细对比，研究差距未完全明确。建议补充现有 AI 心理咨询工具的文献回顾，明确本研究如何创新。

回应：

根据审稿专家的意见，我们已补充相关实证研究的文献综述。

主要包括：

（1）在“2.3 智能体技术在心理咨询领域的应用现状”中，增加对智能体技术在心理咨询

领域应用的文献回顾，重点围绕来访者构建、心理测评或干预以及心理咨询效果评估三个维度展开。

（2）在“1 引言”中，增加对现有 AI 心理咨询工具应用局限性的分析，明确研究差距及本研究的创新之处。

意见 2：

文章提出的系统近乎一个全自动的“黑箱”，这在现实世界的高校咨询中心如何运用仍需更多阐述。建议详细阐述如何与现有体系的整合以及人类咨询师在其中的作用。

回应：

根据审稿专家的意见，我们已阐述系统的应用路径及其与现有体系的整合方式。

主要包括：

（1）本研究构建的心理咨询智能体定位为学生端的心理健康陪伴助手，旨在辅助而非替代传统高校线下心理咨询师的工作。其有望应用于高校新生心理测评、常态化心理健康筛查、日常心理疗愈支持等场景，有效补充现有高校心理咨询体系。

（2）在“3.2.3 ‘测量-咨询-督导’多智能体协作评估”中，进一步说明真实心理咨询专家在过程中的指导作用与真实人类咨询师的评估作用。

意见 3：

智能体技术框架清晰，但技术细节不足，如具体使用的 LLM（如 GPT-4、Llama 3）及其训练过程、微调方式、数据来源未详细说明，可能影响大家的理解和技术的可重复性。心理咨询方法与智能体设计的结合也较为泛泛，未明确如何将结构化（如认知行为疗法）或非结构化咨询方法融入智能体。建议详细说明训练数据集来源、预处理方法及微调过程；补充心理咨询方法如何具体融入智能体设计，如评估者如何应用标准化心理评估工具；确保方法部分符合心理学研究的可操作性标准，如是否咨询了临床心理学家以验证咨询方法的适用性。

回应：

根据审稿专家的意见，我们已补充技术细节以及心理咨询方法的融合路径。

主要包括：

（1）在“2 技术框架与知识融入”中，补充可供选择的大模型示例。

（2）在“3.1.1 心理咨询智能体构建”中，阐述数据主要来源于心理咨询垂域数据，且需进

行脱敏预处理；明确采用非参数化提示与参数化训练相结合的构建方法，并补充监督微调细节；在规划模块中融入结构化流派（如认知行为疗法）思想。

（3）大学生智能体在干预前后分别填写心理咨询效果评估量表，并于干预后提交心理咨询师评估量表。

（4）在“3.2.3 ‘测量-咨询-督导’多智能体协作评估”中，增加真实心理咨询专家参与验证咨询方法适用性、评估流程合理性的关键步骤，确保方法符合心理学研究的可操作性标准。

意见 4：

关于心理咨询流派的介绍，结构化的流派中，朋辈咨询（Peer Consulting）不属于结构化流派，它不属于任何一种心理流派，而是一种心理咨询递送形式，不是用专业的咨询师来递送，而是使用非专业的朋辈咨询师来递送。

回应：

根据审稿专家的意见，我们已在“2.2.2 心理咨询流派”中，删除将朋辈咨询归类为结构化流派的不当表述。

意见 5：

确保各部分逻辑流畅，考虑添加术语表或简要解释相关的 AI 概念，方便心理学读者理解。

回应：

根据审稿专家的意见，我们已采取以下措施明晰相关 AI 概念。

主要包括：

（1）在“2.1.1 单智能体构建”中，详细阐述单智能体的基础架构与构建方法。

（2）在“2.1.2 多智能体协作”中，补充合作性交互与竞争性交互的定义。

意见 6：

补充经典文献：适当补充一些心理咨询领域的奠基性文献。

回应：

根据审稿专家的意见，我们已在“2.2.2 心理咨询流派”中，增补引用主要流派的奠基性文献。

第二轮

审稿人 1 意见：

意见 1：

单智能体构建部分缺乏技术细节，部分表述过于抽象、难以理解，并存在前后不一致的地方。建议作者提供具体示例，方便读者理解操作细节。

请说明本文中的单智能体构建是否计划采用预训练。

回应：

本研究中的单智能体构建不采用预训练方法。针对心理咨询服务任务，基座大模型的通用能力已能满足需求，仅需进行垂域定制，无需额外耗费资源进行预训练。

意见 2：

请详细说明单智能体四个模块分别如何构建。

回应：

根据审稿专家的意见，我们已在“2.1.1 单智能体构建”中，细化单智能体四个模块的构建细节。这四个模块均可采用非参数化提示方法构建。此外，概要模块还可通过预训练或监督微调技术构建，记忆模块利用数据库技术实现写入、检索与反思操作，规划模块涉及思维链技术，行动模块则可借助强化学习技术进行优化。

意见 3：

按照表 1 的信息，大学生智能体的构建不涉及非参数化提示。请问如何使用参数化训练部署概要模块？

回应：

根据审稿专家的意见，我们已在“3.1 心理咨询领域的单智能体构建”、“3.1.2 大学生智能体构建”中，修正大学生智能体的构建涉及非参数化提示与参数化训练方法。

意见 4：

请问短时记忆与长时记忆的写入、检索和反思在具体构建与应用过程中如何实现？

回应：

根据审稿专家的意见，我们已在“2.1.1 单智能体构建”中，补充记忆操作的细节。具体而言，记忆写入、检索与反思操作分别通过数据库的读写、检索与摘要技术实现。

意见 5：

请提供具体操作示例说明如何实现共情规划和主观规划。共情规划为大学生智能体的规划类型。作者提到共情规划“通过分析环境中其他主体的意图生成解决方案”，请说明大学生智能体需要生成什么类型的解决方案。主观规划“基于预设规划或目标函数直接输出确定性策略”。请解释“预设规划”和“目标函数”的含义。

回应：

根据审稿专家的意见，我们已优化规划模块的表述。

（1）在“3.1 心理咨询领域的单智能体构建”、“3.1.2 大学生智能体构建”中，在主观规划方面，心理咨询智能体基于心理咨询理论制定标准化服务策略。在共情规划方面，心理咨询智能体与大学生智能体均根据对方话语内容，动态调整自身行动方案。

（2）在“2.1.1 单智能体构建”中，为避免歧义，已将“解决方案”修正为“行动方案”。

（3）在“3.1 心理咨询领域的单智能体构建”中，例如，咨询流派的专家指南即为咨询师智能体的预设规则，咨询师智能体将据此制定初步咨询方案。为避免潜在误解，删除“目标函数”的相关表述。

意见 6：

多智能体协作系统部分过于笼统，欠缺具体操作细节。

文中提到心理咨询智能体“整合心理健康工具箱（如番茄时钟、冥想音乐），支持咨询师智能体调用工具辅助干预，增强服务灵活性”。请问如何对大学生智能体使用冥想音乐之类的工具？

回应：

根据审稿专家的意见，我们已在“3.2.1 “测评-咨询-督导”多智能体协作系统”中，补充咨询师智能体调用工具的操作细节。以多模态大模型为基座的大学生智能体，能够感知冥想音乐并合理响应。但由于本研究不强调多模态这一特征，为避免误解，将冥想音乐更换为电子日记。

意见 7：

文中提到“心理咨询智能体实时监测大学生智能体的心理状态，建立闭环反馈机制”。请问心理咨询智能体如何“实时监测”大学生智能体的心理状态？如何建立闭环反馈？请提供相关技术细节。

回应：

根据审稿专家的意见，我们已在“3.2.1 “测评-咨询-督导”多智能体协作系统”中，修正关于状态属性的描述。

（1）为“实时监测”大学生智能体的心理状态，在其每轮回复后，心理咨询智能体均对其回复内容进行文本情感分析。

（2）在线心理咨询平台通常采用固定干预策略，不随来访者反应而调整；而本研究中的心理咨询智能体能够“实时监测”大学生智能体的心理状态，并据此动态调整后续干预策略。原文中关于“建立闭环反馈”的表述不够清晰，且该机制并非本研究重点，为避免歧义，我们已将此句删除。

意见 8：

请通过示例解释心理咨询智能体如何“利用督导师智能体的动态反馈优化咨询策略”。

回应：

根据审稿专家的意见，我们已在“3.2.2 “内循环训练-外循环服务”双循环模式”中，细化督导师智能体的反馈流程。

（1）在心理测评子任务中，督导师智能体对比测量师智能体的测评结果与大学生智能体的预设概要差异，结合双方对话历史，反思测量师智能体测评偏差的成因，进而将生成的测评优化建议写入测量师智能体的长期记忆，推动服务策略迭代。

（2）在心理干预子任务中，督导师智能体整合效果评估量表前后测差值与咨询师评估量表结果，结合咨访双方对话历史，反思咨询师智能体得分较低题项的原因，进而将提炼的干预优化策略同步至咨询师智能体的长期记忆。

意见 9：

请解释多智能体协作系统的组织要素为什么要采用“分层结构”。这一设置的意义是什么？针对的是构建阶段还是应用阶段？

回应：

根据审稿专家的意见，我们已在“3.2.1 “测评-咨询-督导”多智能体协作系统”中，阐明“分层结构”的设置意义。第一层固定大学生智能体，引入不同的咨询师智能体，旨在积累同一心理健康问题的多流派干预经验；第二层引入不同的大学生智能体，每位均重复第一层的干预流程，以增强心理咨询智能体对各类心理健康问题的适应能力。该分层结构仅针

对系统构建（“内循环训练”）阶段。

意见 10：

文中提到“大学生智能体在干预前后分别填写心理咨询效果评估量表”。请问在干预前填写评估量表的意义是什么？

回应：

根据审稿专家的意见，我们已在“2.2.3 心理咨询评估”中，明晰心理咨询效果通常通过心理咨询效果评估量表的前后测差值进行评估，干预前填写该量表的意义在于建立基线数据。

意见 11：

心理咨询效果评估部分存在概念混淆。心理咨询效果评估涵盖“形成评估”、“过程评估”和“效果评估”。请修改以上表述。

回应：

根据审稿专家的意见，我们已在“2.2.3 心理咨询评估”中，将“心理咨询效果评估”修订为“心理咨询评估”。“心理咨询评估”涵盖“形成评估”、“过程评估”和“效果评估”三个部分。

意见 12：

请进一步说明大学生智能体的必要性和可行性。

回应：

根据审稿专家的意见，我们已在“2.3 智能体技术在心理咨询领域的应用现状”中，阐述构建大学生智能体的必要性和可行性。通过模拟测量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，可实现测量师与咨询师智能体服务能力的自主进化。通常利用真实大学生的心理咨询档案与对话历史等数据，构建模拟心理健康问题的大学生智能体。

意见 13：

大学生智能体只具备短期记忆，如何“驱动服务策略的迭代优化”？

回应：

根据审稿专家的意见，我们已在“3.1.2 大学生智能体构建”中，修正大学生智能体通

过虚拟用户代理形式，为测量师与咨询师智能体提供高仿真训练环境，以支持督导师智能体优化服务策略。大学生智能体仅设置短期记忆，“服务策略的迭代优化”职责由督导师智能体承担。

意见 14：

大学生智能体是否能够模拟复杂精神疾病的来访者？请引用相关文献加以论证。

回应：

根据审稿专家的意见，我们已在“2.3 智能体技术在心理咨询领域的应用现状”中，补充关于模拟精神疾病患者的来访者智能体的参考文献《Ψ-Arena Interactive Assessment and Optimization of LLM-based Psychological Counselors with Tripartite Feedback》、《Interactive Agents Simulating Counselor-Client Psychological Counseling via Role-Playing LLM-to-LLM Interactions》，并明晰文中提出的“测评-咨询-督导”多智能体协作系统主要面向遇到心理健康问题的普通大学生，而非精神疾病患者。

意见 15：

请确认大学生智能体是否具有行动模块。

回应：

根据审稿专家的意见，我们已在“3.1.2 大学生智能体构建”中，明晰大学生智能体具有行动模块。

意见 16：

多智能体协作系统的服务效果评估指标未涉及大学生智能体。请说明如何保证大学生智能体的效果。

回应：

根据审稿专家的意见，我们已在“3.2.3 “测评-咨询-督导”多智能体协作评估”中，增加图灵测试以验证大学生智能体的构建效果。

意见 17：

请对比现实心理咨询服务说明督导师智能体的必要性。

回应：

根据审稿专家的意见，我们已在“2.3 智能体技术在心理咨询领域的应用现状”中，进

一步阐释督导师智能体的必要性。该智能体能够以旁观者视角，向测量师、咨询师智能体提供服务优化建议，及时纠正服务方向的偏差。

意见 18：

文中提到“督导师智能体通过对比测评结果与大学生智能体的预设概要生成优化建议，动态更新测量师智能体的长期记忆，推动服务策略迭代”。请问如何获取测评结果？测评结果和预设概要是什么关系？如何动态更新长期记忆？

回应：

根据审稿专家的意见，我们已在“3.2.1 “测评-咨询-督导”多智能体协作系统”中，细化测量师与督导师智能体的协作流程。

（1）测量师智能体无法直接访问大学生智能体的预设概要，仅能通过自然语言对话对大学生智能体实施心理健康测评，并据此生成测评结果；督导师智能体则能够访问大学生智能体的预设概要，并接收测量师智能体传递的测评结果文本信息。

（2）测评结果的维度与预设概要的维度一一对应。督导师智能体通过对比测量师智能体的测评结果与大学生智能体的预设概要之间的差异，结合双方的对话历史，分析造成这些差异的原因。

（3）督导师智能体将生成的测评优化建议写入测量师智能体的长期记忆，驱动后者服务策略的迭代优化。

意见 19：

请详细说明如何“确保高危个案及时转介至人工干预”。除了“自残或自杀倾向”，是否又其他需要“立即转介人工干预”的情况？

回应：

根据审稿专家的意见，我们已优化“转介人工干预”的相关表述。

（1）在“3.2.2 “内循环训练-外循环服务”双循环模式”中，若测量师智能体检测到“自残或自杀倾向”等高危情况，系统将立即向高校心理咨询中心发出警报，并同步启动人工干预转介流程。在技术实现层面，“测评-咨询-督导”多智能体协作系统已集成高校心理咨询中心的联系电话，确保高危个案能够及时转介至人工干预。

（2）在“4 总结与展望”中，需要“转介至人工干预”的适用场景，除高危情况外，还涵盖心理咨询智能体判定来访大学生的问题超出其处理能力范围，来访大学生主动提出人工干

预需求等情形。

意见 20：

建议作者将“测量-咨询-督导”中的测量一词调整为测评，测评含义更为广泛，更为符合当前应用场景。

回应：

根据审稿专家的意见，我们已将全文中的“测量-咨询-督导”统一修订为“测评-咨询-督导”。

意见 21：

建议作者增加对于时效性更强的学术性文献的引用，并尽量避免引用科普类书籍杂志。同时，需要补充介绍智能体以及多智能体协作系统在心理健康领域的应用情况。目前已经众多相关的研究与落地应用，请分析其优劣，突出当前研究的创新性。

回应：

根据审稿专家的意见，我们已对文献综述部分进行如下调整：

（1）在“参考文献”中，移除对科普类书籍杂志的引用，并补充近 5 年学术性文献的引用。

（2）在“2.3 智能体技术在心理咨询领域的应用现状”中，补充介绍智能体智能体以及多智能体协作系统在心理咨询领域的应用情况，分析不同发展阶段研究的优势与局限，进而引出本文的创新之处。

……

审稿人 2 意见：

本文具有较高的创新性，论述了基于大模型的智能体在心理咨询中的运用。暂无其他修改意见。

第三轮

审稿人 1 意见：

感谢作者针对上一轮审稿意见的回复。请作者澄清以下问题：

意见 1：

建议作者将“测量师”更名为“测评师”，以符合多智能体协作系统“测评-咨询-督导”功能的命名体系。

回应：

根据审稿专家的意见，我们已将全文中的“测量师”统一修改为“测评师”。

意见 2：

作者提到心理咨询的效果评估可以基于心理咨询效果评估量表的前后测差值。在心理咨询前填写心理咨询效果评估量表似乎略为不妥。建议作者考虑基于来访者心理健康状况的前后测差值对心理咨询效果进行评估。

回应：

根据审稿专家的意见，我们已在“2.2.3 心理咨询评估”、“3.2.2 ‘内循环训练-外循环服务’双循环模式”与“3.2.3 ‘测评-咨询-督导’多智能体协作评估”中，将心理咨询效果评估量表替换为心理健康状况评估量表。参考中科院心理所朱廷劭老师在《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》一文中的做法，本文同样采用抑郁-焦虑-压力量表的前后测差值评估智能体的心理咨询效果。

意见 3：

作者在本轮修改中增加了以下表述：“大学生智能体每轮回复后，心理咨询智能体均对其回复进行文本情感分析，以实时监测其心理状态”。请问这里提到的心理咨询智能体是测量师、咨询师还是督导师？仅对回复进行文本情感分析是否能够实现对于大学生智能体心理状态的实时监测？咨询师不具备反思功能，如何“依据大学生智能体在对话中的心理状态变化，动态调整行动方案”？

回应：

（1）根据审稿专家的意见，我们已在“3.2.1 ‘测评-咨询-督导’多智能体协作系统”与“4 总结与展望”中，明晰此处所指的心理咨询智能体为督导师智能体。

（2）大学生智能体每轮回复后，督导师智能体均对其回复进行文本分析（如聚类分析、情感分析）。仅通过文本分析难以完全实现对大学生智能体心理状态的实时监测，修正为督导师智能体通过文本分析获取大学生智能体的心理状态变化情况。

（3）我们已在“3.1.1 心理咨询智能体构建”中，进一步阐明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。

意见 4：

心理健康工具箱的调用是非常有创意的举措。是否有研究表明这些工具适用于智能体？如果不适用于大学生智能体，如何保证心理咨询智能体可以有效调用这些工具应用于真实大学生来访者？

回应：

（1）目前尚无研究明确指出心理健康工具适用于大学生智能体。

（2）根据审稿专家的意见，我们已在“3.2.1 ‘测评-咨询-督导’多智能体协作系统”中，补充说明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），来访大学生由此直接进入相应工具界面，并自主完成相关操作（如记录日记、聆听音乐）。

意见 5：

依据表 1 的描述，所有智能体的构建方法均为手工整理。这一点是否现实可行？能否提供更多细节，例如作者打算从哪些维度定义人物身份，需要多少组测量师、咨询师与督导师的背景数据，如何构建大学生智能体的描述式概要，如何保证其“覆盖各类心理健康问题”？

回应：

（1）完全依赖手工整理构建所有智能体确实不具现实可行性。

（2）根据审稿专家的意见，我们已在“3.1.1 心理咨询智能体构建”中，说明心理咨询智能体通过描述式概要，从咨询流派、对话风格等维度定义人物身份（如“您是一位采用认知行为疗法的温和心理咨询师”），结合手工整理真实测评师、咨询师与督导师的背景数据，确保智能体身份描述的准确性。在整理过程中，心理咨询智能体持续归纳已有的背景数据，并不断引入新的真实从业者样本，直至信息达到饱和。

（3）我们已在“3.1 心理咨询领域的单智能体构建”、“3.1.2 大学生智能体构建”与“4 总结与展望”中，修订从真实大学生的背景数据中提取关键特征，并借助大模型的生成能力，基于人格特质、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。

意见 6：

针对多智能体协作系统的外循环服务的实现方式，能否提供更多细节？例如，心理咨询智能体如何与大学生来访者进行交互，通过文字聊天、语音聊天还是视频聊天？在交互过程中，心理咨询智能体以何种形式呈现，是否具有虚拟头像、特定语音以及动画展示？作者是否考虑过相应的预算？

回应：

（1）根据审稿专家的意见，我们已在“3.2.2 ‘内循环训练-外循环服务’双循环模式”中，补充外循环服务的交互细节。

（2）心理咨询智能体通过文字聊天与来访大学生进行交互。

（3）在交互过程中，心理咨询智能体以静态虚拟头像呈现。

（4）选择文字聊天与静态虚拟头像的方式，旨在保证基本用户体验的同时控制成本。若后续预算允许，可逐步扩展至语音聊天、视频聊天与动态虚拟头像等更为丰富的交互形式。

意见 7：

鉴于参考文献中预印本较多，无法保证现有技术的成熟性，建议作者加强对于心理咨询智能体、大学生智能体以及多智能体协作系统的评估，尤其是增加有关阶段性评估以及具体评估指标的内容。

如果大学生智能体无法通过图灵测试，是否有应对策略或者备选方案？

回应：

（1）根据审稿专家的意见，我们已在“参考文献”中，将预印本替换为正式发表于权威期刊或计算机顶会的论文。

（2）我们已在“3.2.3 ‘测评-咨询-督导’多智能体协作评估”中，加强对于相关智能体与多智能体协作系统的评估。

（3）若大学生智能体未能通过图灵测试，我们将招募真实大学生参与系统的迭代优化。但根据清华大学彭凯平老师在《大语言模型模拟区域心理结构的有效性：人格与幸福感的实证检验》一文中的研究，智能体在模拟特定个体方面展现出较强的能力。因此，大学生智能体通过图灵测试的可能性较高。

意见 8：

督导师智能体是本文的创新点之一。如何评估督导师智能体的反思功能，确保其对测量

师和咨询师智能体提供的优化建议是准确有效的？

回应：

由真实督导师对督导师智能体的反思操作进行评估，判断其提供的优化建议是否准确有效。若评估未达预期，真实督导师即介入指导，对督导师智能体进行迭代优化。

意见 9：

作者打算如何评估多智能体协作系统的总体功能，是否有具体的指标用于确认内循环训练完成、已具备外循环服务的能力？

回应：

参考清华大学黄民烈老师在《Ψ-ARENA: Interactive Assessment and Optimization of LLM-based Psychological Counselors with Tripartite Feedback》一文中的方法，由真实督导师采用督导量表（Supervisor 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。量表得分达到 24 分及以上即表明系统通过验证，具备开展外循环服务的能力。

意见 10：

文中提到测量师智能体的测评准确性和咨询师智能体的干预成效由督导师智能体完成，这一做法似乎不太妥当。

回应：

（1）参考上海交通大学朱其立老师在《Depression Diagnosis Dialogue Simulation Self-improving Psychiatrist with Tertiary Memory》一文中的设定，测评师智能体无法直接访问大学生智能体的预设概要，仅能通过自然语言对话对大学生智能体实施心理健康测评，并据此生成测评结果；督导师智能体则能够访问大学生智能体的预设概要，并接收测评师智能体传递的测评结果。督导师智能体通过对比大学生智能体的预设概要与测评师智能体的测评结果之间的差异，结合双方的对话历史，分析造成这些差异的原因，进而生成测评优化建议。这些测评优化建议随后被写入测评师智能体的长期记忆，驱动其服务策略的迭代优化。

（2）咨询师智能体在评估自身干预成效时，可能受到自身垂域知识的局限，导致评估结果偏高。为此，引入作为独立第三方的督导师智能体，由其负责对咨询师智能体的干预成效进行客观、中立的评估。

第四轮

审稿人 1 意见：

意见 1：

心理咨询智能体构建

请举例说明“真实测评师、咨询师与督导师的背景数据”包含哪些数据，并定义何为“信息达到饱和”。

回应：

根据审稿专家的意见，我们已在“3.1.1 心理咨询智能体构建”中，补充心理咨询智能体的构建细节。

（1）真实测评师的背景数据主要包括适用于大学生群体的量表题目与解读方法、信息挖掘与整合技术等，真实咨询师的背景数据涉及其依据的咨询流派专家指南、共情技巧等，真实督导师的背景数据则涵盖督导理论模型、对咨询师的指导策略等。

（2）本研究借鉴质性研究中的“信息达到饱和”概念，当新增的真实从业者样本对现有信息结构不再产生实质性补充时，即达到数据饱和状态。一旦满足该条件，可认定所构建的心理咨询智能体在知识结构上具备充分的深度与广度，能够全面、稳定地代表目标专业群体，此时应终止数据收集工作。

意见 2：

文中提到“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。”请问测评师在对话中的作用是什么，为什么督导师智能体需要将动态调整建议传递给咨询师智能体？

回应：

（1）在对话过程中，测评师智能体主要负责在初始阶段对来访大学生的心理状态进行全面评估。当督导师智能体监测到来访大学生在回复特定内容时表现出低落、痛苦或逃避等反应，将及时提示测评师智能体调整语气或转换测评方向，以更好地契合来访者当前的心理状态。

（2）在咨询师智能体与来访大学生的对话中，常出现“当局者迷”的情形：咨询师智能体因注意力有限，容易固守预设的结构化服务策略，而忽略来访者个性化的情感体验。督导师智能体作为独立的第三方观察者，能够客观捕捉来访者心理状态的动态变化，并将策略调整建议反馈给咨询师智能体，从而提升心理咨询服务的针对性。

意见 3：

大学生智能体构建

请举例说明如何“从真实大学生的被试数据中提取关键特征”。背景数据包含哪些内容，何为关键特征。

回应：

根据审稿专家的意见，我们已在“3.1.2 大学生智能体构建”中，修订大学生智能体的构建细节。

（1）为提高研究效率与数据有效性，在获取真实大学生背景数据时，我们不再进行全维度数据采集，而是聚焦于能够有效反映其心理与行为模式的关键特征。

（2）关键特征的选取依据《中国国民心理健康发展报告（2023~2024）》中《2024 年大学生心理健康状况调查报告》所采用的测评题项。通过专业测评量表，收集真实大学生在多个核心维度上的属性数据，包括人格特质（如大五人格）、心理状态（如抑郁、焦虑、生活满意度）、生活状态（如恋爱状况、体育锻炼情况、睡眠质量）等。

意见 4：

请修改图 2 中测量师的名称。

回应：

根据审稿专家的意见，我们已在“3.2.2 ‘内循环训练-外循环服务’双循环模式”中，修改图 2 中测量师的名称为测评师。

意见 5：

本文对于现有基于大模型的智能体技术过于乐观。最新研究表明与大模型长时间的互动有可能会引发心理问题（例如 AI psychosis，参见 https://doi.org/10.31234/osf.io/cmy7n_v5）。请作者在文中强调有关风险，并阐述如何避免该系统可能带来的负面影响。

回应：

（1）智能体应用的广阔前景

①权威文献支持

核心参考文献：《心理学报》第 57 卷第 11 期《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》指出，采用提示词工程构建的大模型心理咨询系统在我国实际应用中已展现出服务真实来访者的安全性与有效性，为本研究提供了理论与实证依据。

《Nature》、《The Lancet》相关文献：包括综述性文章《Using Large Language Models in Psychology》、《A Scoping Review of Large Language Models for Generative Tasks in Mental Health Care》，以及涉及心理测评与干预的多篇实证研究，如《Using a Fine-tuned Large Language Model for Symptom-based Depression Evaluation》、《Feasibility of Combining Spatial Computing and AI for Mental Health Support in Anxiety and Depression》、《Large Language Models as Mental Health Providers》等，均表明了基于大模型的智能体在心理测评与干预服务中的安全性与有效性。

《心理科学进展》、《心理学报》相关文献：《基于大语言模型的自杀意念文本数据增强与识别技术》、《大语言模型的共情模拟：评估、提升与挑战》等文章，进一步验证了大模型技术在心理咨询领域应用的安全性与有效性。

②实际应用验证

2022 年 7 月，科大讯飞推出“AI 减压星球”，目前已覆盖 4947 所学校，服务 391 万中小学生，有效降低了典型心理问题（如自责倾向、学习焦虑）的检出率。

2022 年 9 月，基于提示词工程技术的“小清心”智能体在清华大学上线，为全校学生提供心理支持服务，迄今已累计服务 49301 人次。

（2）关于“AI psychosis”研究的澄清

①主要观点

“AI psychosis”研究指出，大模型应用于心理咨询服务是大势所趋，需设计合理机制以应对潜在风险。

②研究对象

“AI psychosis”主要针对已患或易患精神疾病的用户，认为大模型可能强化其妄想症状，从而加剧病情。而本研究仅面向“存在一般心理健康问题的普通人群”。具体而言，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。

③研究工具

“AI psychosis”中所指的“大模型”为通用的基座大模型（如 DeepSeek、Qwen），未融入心理咨询垂域知识与数据。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。

④文章来源

“AI psychosis”为预印本（发布于 PsyArXiv），尚未经过同行评审。

⑤数据来源

“AI psychosis”中引用的数据主要来自媒体报道，其准确性未经严格验证。

（3）明确系统辅助定位

在摘要系统有望成为大学生心理咨询的有效辅助工具、引言旨在开发面向大学生群体的智能化心理咨询辅助工具与总结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中强调系统仅为辅助工具，不能替代真实咨询师。

第五轮

编委 1 意见：

感谢作者的投稿及修订工作。稿件在人工智能与心理咨询交叉领域具有一定探索意义，但经编辑部与专家综合评议后，认为当前版本仍存在重大风险与不足，不适宜在本刊正式发表。主要理由如下：

意见 1：

伦理与安全风险未充分评估。最新研究显示，长时间与大语言模型互动可能引发心理风险（如“AI psychosis”现象，参见 https://doi.org/10.31234/osf.io/cmy7n_v5），而稿件尚未提出相应的风险识别与防控机制。

回应：

根据编委专家的意见，我们细化了覆盖系统设计、测试与使用全流程的风险识别与防控机制。

（1）系统设计阶段

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：真实心理咨询专家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）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过度依赖。

（2）系统测试阶段

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；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。

（3）系统使用阶段

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：督导师智能体实时跟踪来访大学生的心理状态变化，一旦检测到持续恶化趋势，立即启动人工服务流程。

意见 2：

缺乏权威指导与伦理保障。目前国际主流心理学专业机构尚未就生成式 AI 用于心理咨询场景发布正式指南，相关伦理审查和合规验证值得商榷。

回应：

根据编委专家的意见，我们补充了生成式 AI 用于心理咨询场景的权威指导。

（1）核心政策导向

2025 年 8 月，国务院发布《关于深入实施“人工智能+”行动的意见》：提出到 2027 年“新一代智能终端、智能体等应用普及率超 70%”，到 2030 年“普及率超 90%”的发展目标。本研究积极推动《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（习近平, 2018）。

2025 年 11 月，国家卫健委发布《关于促进和规范“人工智能+医疗卫生”应用发展的实施意见》：明确要求“强化公众心理问题智能监测服务，以学生为重点，提供心理问题智能筛查、预警推送、干预服务和随访分析。”

（2）国际权威指南

2023 年，美国心理咨询师协会(American Counseling Association，ACA)发布《AI Recommendations for Counselors and Clients》：是全球第一份指南性文件，旨在确保 AI 开发与应用过程中始终将来访者的福祉、偏好与价值观置于首位，同时为心理咨询师、心理教育者与来访者提供 AI 应用的相关指引。

2023 年 5 月，美国心理学会(American Psychological Association，APA)发布《Statement on Artificial Intelligence》：强调 AI 在心理健康领域的定位为“增强而非替代”，AI 应用须遵循伦理与隐私原则。

2023 年 10 月，APA 发布《Ethical Guidance for the Use of Artificial Intelligence in Professional Psychological Practice》：提出应用 AI 时需遵循行善、不伤害、尊重自主、公正、忠诚、透明六项核心伦理原则。

2024 年 8 月，APA 发布《Artificial Intelligence and the Field of Psychology》：是目前最为完整的指导文件，系统探讨了 AI 的应用原则、风险与伦理挑战，并呼吁心理学工作者积极参与 AI 设计，提升其可解释性与文化适应性。本研究以美国心理学会《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。

（3）我国部委指南

2024 年 11 月，国家卫健委发布《卫生健康行业人工智能应用场景参考指引》：定义“智能学生心理健康管理服务”的基本概念为“应用人工智能、大数据技术，对学生开展心理危机筛查、辅助诊断、预警推送、干预服务和随访分析”，其应用场景包括“将互联网、云计算、大数据、人工智能、可穿戴等技术融入学生心理的测评、分析、管理与服务之中”，“建

立集心理危机筛查定级、心理潜能精准开发、心理宣教资料精准推荐、分类远程监督及管理、干预方案智能推送、大数据多维学生发展个体画像输出等智能学生心理健康管理与服务系统。”

2025 年 2 月，国家标准委公示《心理咨询服务 第 4 部分：人工智能技术辅助应用指南》：“给出利用人工智能技术辅助支持开展心理咨询服务的工作指引，包括总体原则和要求、服务人员、服务过程、风险管理和隐私保护、服务评估和改进。”本研究以国家标准《心理咨询服务 第 4 部分：人工智能技术辅助应用指南》(2025)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。

2025 年 4 月，国家标准委发布《网络安全技术 生成式人工智能服务安全基本要求》：“规定了生成式人工智能服务在训练数据安全、模型安全、安全措施等方面的要求”，为服务提供者开展相关应用提供参考。本研究严格遵循《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。

意见 3：

研究设计与验证不足。论文的“咨询代理”方案仍停留在概念层面，缺乏相关系统安全测试、风险评估及人机交互的独立验证。

回应：

请参见“1.伦理与安全风险未充分评估”部分。

编委 2 意见：

我认真研读了贵稿，并充分参考了两位外审专家的评审意见。总体而言，我高度认同外审专家对本文选题前沿性、现实意义与转化价值的积极评价。AI 智能体应用于大学生心理健康场景，确实体现了心理科学与人工智能技术深度融合的趋势，具备显著的学术引领潜力与社会服务价值。考虑到《心理科学进展》对领域内前沿问题的导向作用，我在外审专家最后意见基础上，进一步提出以下三点补充建议，供作者参考与回应：

意见 1：

关于 AI 智能体交互模态的技术细节与评估维度

当前文本对 AI 智能体与大学生用户的交互方式描述较为笼统。建议作者进一步澄清以

下问题：AI 智能体是否仅支持文本输入？若系统已采集语音数据，是否仅用于转写文本，还是进一步提取副语言特征这些特征是否被纳入诊断或风险评估模型？

回应：

根据编委专家的意见，我们补充了心理咨询智能体与大学生用户的交互细节。

在“3.1.1 心理咨询智能体构建”中，补充在对录音、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）。

在“3.2.1 ‘测评-咨询-督导’多智能体协作系统”中，补充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。

在“3.2.2 ‘内循环训练-外循环服务’双循环模式”中，“内循环训练”阶段补充（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。“外循环服务”阶段补充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，以及（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，并根据其心理风险等级进行分流处理。

在“3.2.4 系统伦理安全保障”中，补充（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。

在“4 总结与展望”中，补充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。

意见 2：

关于“AI psychosis”风险的学理回应与概念澄清

外审专家一提出的“AI psychosis”问题，作者已在修订稿中作出回应，显示出对国内外伦理法规（如《人工智能伦理风险管理办法》《生成式 AI 服务管理暂行办法》）的熟悉。然而，从学理层面看，回应仍显防御性有余、建构性不足。建议明确区分“AI psychosis”作为大众媒体隐喻与临床心理科学术语的差异。可参考近期研究（如 Morrin et al., 2025; King’s College London, 2025）提出的“AI-delusional disorder”或“AI-augmented psychotic episode”等更精准表述。干预与熔断机制：除被动告知“AI 非真人”外，是否可设计动态

风险预警系统？例如，当用户连续出现妄想话语、情感过度依附或现实检验能力下降的语言标记时，系统是否可触发人工转介？为降低用户因长期线上交互而诱发“AI 依赖”或现实感下降的风险，建议作者在算法层面增加“线下促进模块”。

回应：

根据编委专家的意见，我们在“3.2.4 系统伦理安全保障”中，对“AI Psychosis”风险进行了学理层面回应，并补充了具体防控措施。

（1）伦理法规补充

补充（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。

（2）核心概念修正

将“AI Psychosis”修正为“AI-delusional Disorder”，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、用户过度依赖以及人工智能妄想障碍(AI-delusional Disorder)等伦理安全风险(Morrin et al., 2025)。

（3）“AI 非真人”

补充（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”。

（4）动态风险预警

补充在使用阶段，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。

（5）“线下促进模块”

为降低长期线上交互可能诱发的“AI 依赖”或现实感下降风险，（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。

意见 3：

关于“信息饱和”表述的术语修正与方法论补充

作者在修订稿中使用“信息饱和”一词来描述 AI 智能体在收集用户数据过程中逐渐达到诊断或评估稳定状态的现象。该术语源自质性研究，强调“继续采集数据不再产生新概念”的定性判断标准。然而，在本文所涉及的机器学习与心理评估算法优化语境中，“信息饱和”既非定量指标，也缺乏统计可操作性，易引发歧义。为此，建议作者摒弃“信息饱和”这一质性术语，转而采用机器学习领域更为精准且可验证的敏感性分析框架，以论证模型在数据增量、参数扰动或样本分布变化下的稳健性（robustness）与收敛性（convergence）。

回应：

根据编委专家的意见，我们删除了原文中“信息饱和”这一表述，转而采用机器学习领域更为精准的评价指标。

在“3.1.1 心理咨询智能体构建”中，补充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。

第六轮

编委 2 意见：

来稿已仔细读完，修改满意，同意发表。

编委 3 意见：

同意发表。

主编意见：

稿件经过多位专家的审阅，作者进行了认真的修改，达到了发表水平，同意发表。