第一轮修改

诚挚感谢编辑部老师们的辛勤付出，感谢审稿专家们百忙之中审阅本文，并提出了非常宝贵的修改意见，对于审稿专家提出的意见，文章已作相应修改，所有对应修改内容已在文中用红色字体标注。

审稿专家一：

本文以 ChatGPT-4 为例，从语义响应偏移分析框架角度，探索了人工智能技术面向心理健康应用的回答策略及可能偏差。研究选题具有良好的时效性，研究发现具有一定的潜在应用价值。目前论文文稿还需要较大幅度的调整，具体意见如下。

意见 1：

本文主题是心理健康还是学生心理健康？

回应 1：

感谢审稿专家指出“主题不清晰”的问题。我们已对涉及到的章节内容进行了统一修改，进一步明确了本研究的主题为“心理健康”，而非指“学生心理健康”。具体修改如下：

1.标题与核心内容聚焦于“心理健康教育”，强调研究面向的是广义人群，包括但不限于学校等心理健康教育场景。

2.原提示语方案基于涵盖多群体（如学校、社区）的公开语料制定，故适用群体并不限于大学生。我们的提示语方案则在此基础上结合心理健康教育与心理咨询领域专家们的意见，进行了本土化修订，适用范围也不局限于学生心理健康问题。

3.原实验设计中，用户角色设定存在“学生身份”倾向问题，已统一修订为中性词“中文用户”；在偏差分析中做出相应修改，删除了“学科情境”组。

4.实验数据已重新收集并分析，使生成式 AI 的回答内容符合更具普适性的心理健康教育语境，体现本研究面向的是广义人群的定位。

意见 2：

引言中共情计算概念的提出和讨论似乎与大语言模型没有实际关联，ChatGPT 中并未单独考虑共情计算技术。

回应 2:

非常感谢审稿专家对引言中“共情计算概念”关联性问题的指正。我们充分认同，原稿在实验设计中并未系统考虑或实现共情计算模块。因此，引言中关于共情计算（empathy computing）的表述确有不妥。据此，我们在修订稿中对引言部分进行了重写，删除了涉及“共情计算”相关表述，并补充了大语言模型相关的概念，以避免概念误导。

1.补充了关于大语言模型在心理健康教育对话中应用价值的实质性讨论，重点强调其在多轮对话理解与语义生成等方面的优势。

2.明确界定本研究的理论聚焦，强调研究关注的是基于大语言模型架构的国内外主流生成式 AI 在心理健康教育场景应用中，表现出的语义分布、回答策略以及潜在偏差，而非对人类情感过程的模拟。

以上修改已在修订稿中标注，恳请专家审阅。

意见 3：

本文所提倡的对语义结构和词汇选择的系统性考察，对心理健康应用评估有何特殊的价值和意义？

回应 3:

感谢审稿专家指出的“语义结构和词汇选择的价值与意义”问题，我们在引言与讨论部分进行了补充与澄清。

在针对以往 AI 文本质量的评估研究中，单纯依赖事实准确性或主观满意度评分已难以满足高风险应用场景的安全与可解释性需求。越来越多的证据表明，语义结构与词汇选择是理解文本深层逻辑的关键。其一，语义分析框架揭示了主题衔接与论证层级，能捕捉表面正确但是语言逻辑混乱的现象。其二，细粒度的用词差异反映立场、文化暗示乃至刻板隐喻，量化这一层面有助于及早识别误导性价值倾向。其三，相较于定性内容的评估，语义与词汇指标提供了跨参数、跨模型、跨文化的统一评价尺度，使研究者得以系统地比较每一轮语义结构的稳定性与多样性。其四，在未来实际场景应用中，基于语义分析框架与词汇倾向的实时监测可构建可追溯的审核机制，当回答内容在结构或用词上出现异常波动时提供预警，从而显著提升生成式 AI 的透明度与安全性。这些综合价值共同表明，对语义结构与词汇选择的系统性考察在 AI 评估中至关重要。

具体而言，研究从“语义分析框架结构”与“词汇选用倾向”两个维度，循环生成模型的回答数据，通过将词语向量化，通过语义矩阵相似度与聚类方法，能够观察到类似的词语在空间距离上是比较相近的。通过此类方法，尽管AI在表述上可能采用不同的词或由词组成的句子，但由于含义相近的词向量在空间距离中更相近，我们仍然能够观察到类似的语义空间分布。从而通过回答内容的语言结构评估其语义分布的稳定性与策略选择。

（1）识别语义漂移与AI幻觉。多轮回答中模型可能存在前后矛盾的表达或“语义漂移”等问题，传统句级评价难以识别。我们通过构建词向量语义结构矩阵，并使用Frobenius范数量化不同回答轮次的一致性与多样性，有助于识别潜在的“幻觉”和语义不连贯等问题，为回答

内容的可控性提供解释性机制。（2）识别AI在不同类别的心理健康问题上的回答差异。本研究通过词汇结构与语义分布分析，识别生成式AI在不同类型心理健康问题下所采取的应对策略差异。通过结合具体问题，对回答的语义结构系统分析，可判断模型是否根据问题属性动态调整回答策略。这种策略适配能力在心理健康教育实践中尤为重要，有助于提升模型建议的可操作性与场景贴合度。（3）建立语义输出与心理干预理论的映射关系。本研究通过聚类分析语义结构之间的相似性，识别模型在多轮语义生成中的回答策略类别，将策略类型与心理干预理论建立联系。以认知闭合需求、启发式决策、控制－价值理论等经典心理学理论进行映射与解释，最终归纳出“核心主导型、核心扩展型、双峰扩展型、多元开放型”四类策略。综上，从语义结构与词汇层面的系统分析，不仅填补了现有心理健康AI评估体系在语言机制层面的盲区，也为后续开展生成模型的实用性验证提供了关键方法支撑。

意见 4:

本文提到的当前研究三点不足，如何在研究中回应和体现？进一步的，本文的核心研究问题是什么？目前读起来并不清晰。

回应 4:

感谢审稿专家指出的关于“研究问题不清晰”与“当前研究三点不足如何回应”的关键性意见。我们充分认同，在原稿中对研究动因与结构回应之间的

映射尚不够清晰。为此，我们对引言与讨论、结果部分结构进行了修改与补充，以回应这两方面问题。

一、对当前研究三点不足的回应

我们修改了原先文章中提到不足之处，明确了研究的核心问题并重新表述现有生成式 AI 在心理健康教育研究中的两方面挑战。一是回答内容不稳定（hallucination），二是系统性偏差（systemic bias）。而以往研究在解决这两方面问题时，存在以下三点不足：（1）评估维度单一。缺乏多轮语义一致性量化指标，且常依赖于对单次回答的固定分析，方法外推性不足；（2）模型回答策略生成机制不明。缺乏对其背后规范推理和社会理论基础的充分探讨，难以从计算科学与心理学视角共同揭示模型的可靠性；（3）对社会情境下的偏差分析的视角单一。缺乏定性与定量结合的依据，忽略模型偏差计算背后的社会心理学视角。具体文中方法部分回应如下：

1.缺乏多轮语义一致性的量化手段。在方法中提出了基于 Frobenius 范数的语义结构相似度分析方法，对每个问题的 30 轮回答构建 30×30 相似度矩阵，并通过均值与标准差，量化了模型输出的语义一致性与多样性水平。并采用“主模型+副模型”的验证方式，横向评估了主模型（DeepSeek）不同参数下的表现，纵向验证了在其他国内外主流大模型（ChatGPT，豆包）上的泛化性能。

2.模型回答策略机制不明，缺乏系统归类与解释。在方法中引入 PCA 降维 + K-means 聚类方法，结合心理学理论，识别出现实数据驱动下大模型的回答策略模式。在结果中给出策略类别，并在讨论中结合数据驱动原理与对应策略的生成机制进行心理学解释。

3.社会情境偏差研究单一。在方法中，引入内容分析+非参数检验（Mann–Whitney U 检验），对性别与民族情境下的六项维度得分进行比较。在结果中汇报差异表现，并基于统计显著性与语义差异对偏差进行解释与界定，增强研究的实证基础与分析深度。

二、关于“核心研究问题”的回应与补充

为进一步增强研究的聚焦性与可读性，我们在引言部分明确设定三个核心研究问题，并在文中逐一展开（1）生成式 AI 在多轮心理健康教育对话中，是否具备语义结构的一致性与多样性？（2）生成式 AI 是否在特定问题上存在稳

定的回答策略，这些策略如何分类与解释？以及其在参数化调整下和其他模型上的表现是否存在差异？（3）在不同社会情境下（如性别、民族），生成式AI是否存在语言偏差？如何定性识别，并定量评估其差异？

在结果部分中，文章以此三个核心问题为逻辑主线依次展开：语义分布特征、策略聚类识别、情境偏差分析；在讨论部分中，则围绕这三个研究问题进行综合性分析与理论映射，强化研究主线与结论逻辑的闭环。

意见 5:

研究方法：本文的多轮对话和情境设置如何具体实现？目前的方法细节介绍缺失。

回应 5:

感谢审稿专家提出的宝贵建议。针对“多轮对话与情境设置缺乏具体实现细节”的问题，我们在修订稿中已进行了系统性补充。主要在方法章节 2.1 中，我们详细说明了多轮对话生成的具体流程和参数配置，考虑到文章篇幅限制，我们在附录 1 中补充了更加详细的情境设置过程和技术细节。

意见 6:

仅用单个 AI 模型开展研究，并且没有进行参数化的探索，所得结论是不充分的。如果要得到有价值的结果，需要开展多个对话模型的平行测试，并且对模型进行必要的参数化探索。

回应 6:

感谢审稿专家提出的重要建议。我们充分认同，仅基于单一模型与单一参数设定的研究，在解释力与外推性方面存在不足。考虑这一问题的重要性，我们在修订稿中新增了模型参数化实验，以评估不同生成策略对模型语义稳定性和偏差表现的影响。（1）以 DeepSeek 为主模型，补充了多个对话模型的平行测试，（2）纵向对比了国内其他平台模型的预测效果。此外，我们亦在方法章节 2 和附录 1 中新增参数设定规则，并明确本研究所采用参数均为可控变量，未来亦具备迁移至其他生成模型的技术通用性。

具体而言，我们在主模型的原始参数基础上新增三组设置，分别对应“低温稳定组”（*temperature*=0.3, *top_p*=0.9），“中温均衡组”（*temperature*=0.5, *top_p*=0.95）与“高温发散组”（*temperature*=0.7, *top_p*=1.0），三种参数组合。

在每组设定下，我们对 7 类代表性提示语分别生成 30 轮回答文本，并使用原语义一致性量化指标进行分析对比。已将对应的实验结果补充至章节 3.2.2、3.2.3 中，并可视化呈现参数变化下语义稳定性的变化轨迹结果（见图 9；表 3、表 4）。

意见 7:

图 1 太多技术细节，本文特色关键信息不突出。同时，所有方法需要明确的细节介绍，以读者可以重现为标准进行展开，并给出必要的依据。例如，k-means 基于怎样的准则选择了 3 聚类？其他聚类数下的指标表现如何？等等。

回应 7:

感谢审稿专家对图示设计与方法透明度提出的重要反馈。我们已根据您的建议对图 1 与相关方法部分进行了如下修订与补充。

（1）关于图 1 信息冗余问题

原图 1 中确实包含过多底层技术流程，导致论文整体研究逻辑与关键贡献不够突出。我们重新修订了图 1。保留整体研究框架的“数据来源与生成方式”，“语义建模与策略识别”，“偏差分析”三层结构，聚焦呈现本研究的语义评估思路与偏差检测目标。

（2）关于方法细节展开问题

我们已在对应方法 2.3 和附录 1 中细化了如下技术实现细节，确保方法细节明确、方法可复现。明确指出词向量采用中文预训练词向量库（维度 300），分词工具为 Jieba；相似度计算采用标准化语义矩阵的 Frobenius 范数矩阵的均值与标准差；降维算法分别使用 PCA 与 t-SNE 并行处理，用于可视化与策略聚类；每轮回答均独立设定对话上下文，记录 UUID 与时间戳，确保批次可追溯。在文中呈现了原始技术实现流程，包括词嵌入、相似度矩阵计算、聚类等处理路径，增强可复现性。并在附录 1 中补充了更加详细的技术细节。

（3）关于 K-means 聚类参数选择

对于聚类策略数量（$k = 3$）的确定，我们在原分析基础上新增如下补充说明（附录 1D）。在本研究中，尽管在部分问题中轮廓系数（Silhouette Score）在 $k = 2$ 或 $k = 4$、$k = 5$ 时略有提升，但我们坚持选择 $k = 3$ 作为统一聚类数标准。主要源于策略方案的一致性的需求。本研究核心目标之一是对比 21 个问题

在语义生成策略上的类型分布与差异，若每个问题采用不同的聚类数，结果将无法形成具有可比性、可聚合性的结构映射。在心理健康教育等实际应用场景中，我们更关注模型在多轮对话中是否呈现出稳定的策略模式数量，而不仅仅是指标选择（如最大化轮廓系数）。

具体而言，$k=3$ 在大多数问题中表现出最优或次优的聚类能力。从表格统计结果（附录：表 2）可见，在 21 个问题中，有 16 个问题在 $k=3$ 时的轮廓系数排名为前两位，显示出 $k=3$ 在跨问题情境中的较强适配性与稳定性。尽管部分问题在 $k=2$ 时轮廓系数略高，但多数情况下差异有限（例如问题 1 中 $k=2$ 为 0.6126，$k=3$ 为 0.5802，差距不足 0.04），并不构成对模型结构的实质性支持。相反，$k=2$ 聚类往往会掩盖少数关键策略路径的存在，降低语义层次的辨识度与可解释性。此外，选择 $k=3$ 在兼顾结构稳定性与心理策略多样性方面更具理论与实证基础。我们在语义聚类基础上，进一步将模型生成的回答策略映射策略类别。该策略结构在计算指标和心理学理论中具有平衡的可解释性与操作性。选定 $k=3$ 不仅满足数据聚类结构的稳定性要求，也增强了模型输出与心理机制的映射力。

综上所述，统一使用 $k=3$ 聚类，是在理论可解释性、跨问题一致性和统计适配性的标准下的最优选择，其结果可为后续的策略分类提供基础。

意见 8:

文章的结果和讨论需要重新梳理的问题和研究设计来重新组织。目前版本的感受是，现在的表达不够规范和清晰。

回应 8:

感谢审稿专家指出当前“结果与讨论部分表达不够清晰规范”的问题。我们充分认同文章中存在的逻辑性与研究主线的一致性表达不清晰的问题，并据此对引言，讨论和结果进行了系统性的梳理，重新围绕研究设计进行重新组织。

审稿专家二：

研究选题关注了生成式人工智能在心理健康教育领域的应用，具有很好的前沿性和应用性。文章逻辑清晰、内容翔实，研究方法得当，结果对心理健康教育实践具有较强的指导意义。我有以下几点建议请作者考虑。

意见 1：

作者基于 ChatGPT-4 进行分析，所得出的结论是否适用于国内大语言模型如 Deepseek、豆包、文心一言等？

回应 1：

感谢审稿专家提出的宝贵建议。我们充分认同，本研究原先采用的是基于 ChatGPT-4 的单模型评估，确实存在外推性不足的问题。本研究所提出的语义分析框架及偏差评估流程具有较强的模型通用性，可迁移至其他具备 API 接口或对话交互能力的生成式大语言模型中，包括国内的 Deepseek、豆包、文心一言等。

鉴于不同模型在训练数据来源、对齐策略及语义生成机制上的差异，我们重新收集数据并补充对应实验，并在文章方法、结果和讨论部分进行修改，以 DeepSeek 为主模型，两个国内外主流模型（豆包、ChatGPT）为副模型，系统比较了生成式 AI 模型在心理健康教育对话中的语义稳定性与偏差表现，进一步拓展模型适配性与文化敏感性的评估维度。

意见 2:

作者修订了 Maurya（2023）的心理健康教育问答提示语，但是我看到该研究所发表的期刊并不是领域内的权威期刊或者高水平期刊，JCR 分区是 3 区。这种情况下这个手册的可信度和科学性如何保证？比如“睡觉前我应该什么时候停止使用手机？”这个为何分到了生活方式的类别而不是一般健康类别？作者可以思考一下这些提示语在所属类别划分上有没有其他的分类方法。

回应 2:

感谢审稿专家对提示语来源可靠性与分类逻辑提出的重要建议，我们对 21 条提示语列表进行了逐条比对与重新修订。最终形成一套更贴近中国语境实际使用习惯的 21 条心理健康提示语，并在方法章节 2 进行修改和补充。

（1）关于提示语手册的科学性保证

鉴于 Maurya（2023）所发表期刊影响力有限（JCR 3 区），我们在修订稿中不再仅依赖其提示语体系，而是通过以下方式对 21 条提示语列表进行了再构建与本土化校准，调整提示语表述，使其更贴合中国用户语境。采 Delphi 法进行专家轮询验证：首先，我们组织了 2 轮 Delphi 匿名专家评审，共邀请 6 位具有心理咨询、临床心理干预与人工智能交叉研究背景的专家对提示语的分类准确性、表述清晰性与文化适配性进行匿名评审。分别对提示语的“类别归属”、“表述清晰度”、“文化贴切性”三个维度进行打分与反馈修订意见；其次，对二次修订后的版本进行重新匿名评分，最终达成 *S-CVI/Ave* = 0.952 的一致性共识，确保具有更高的可用性与说服力。

（2）“关于睡觉前我应该什么时候停止使用手机”分类的合理性

对于 Q18 条目的分类，我们在修订稿中修订了提示语表述与归类依据，以回应专家提出的“生活方式 vs 一般健康”的问题。Q18 问题的表述方式的确模糊于“一般健康类别（关注睡眠健康：睡觉前多久停止使用手机有助于改善睡眠）”与“生活方式（关注睡前行为：如何安排睡前使用手机的时间）”之间。故我们结合专家的建议作出调整。将其修改为“睡觉前放下手机的最佳时间是什么时候？”该修订强化了行为导向特征，使其更明确聚焦于个体日常行为习惯与自我调控策略。

分类上，我们保留其在“生活方式”类别中，主要依据如下：研究分类“生活方式”的核心目的在于个体习惯管理与行为调整。本意是基于其与日常行为习惯的密切关联，强调屏幕设备（数字产品）使用时间与日常行为模式的关系。世界卫生组织在其针对 5 岁以下儿童的指南中，已将屏幕时间、久坐行为和睡眠纳入“生活方式医学（Lifestyle Medicine）”的范畴，强调其在健康干预中的行为习惯管理属性（World Health Organization, 2019）。斯坦福大学生活方式医学中心建议，屏幕设备使用时间的干预本质是一种生活方式行为调整，在睡前应限制情绪刺激内容的暴露（Stanford Lifestyle Medicine, 2024）。因此，我们依据上述行为目标，将“睡前手机使用”归入“生活方式”而非“个人健康”类问题，使其更契合心理健康教育中文化可迁移性与干预实操性。

参考文献：

Lifestyle Medicine. (2024). *Screen time and sleep: It's different for adults*. Stanford University School of Medicine.

World Health Organization. (2019). *Guidelines on physical activity, sedentary behaviour and sleep for children under 5 years of age*. Geneva: World Health Organization.

意见 3:

作者有没有进行一些访谈，征集一下中国被试在使用生成式人工智能模型时会使用的相关提示语？

回应 3：

感谢审稿专家关于提示语来源与文化适配性的关键建议。针对这一问题，我们在修订过程中补充并优化了提示语清单的来源逻辑，增强了其代表性与文化适配性。具体而言，我们通过间接征集的方式，邀请了六位国内的心理健康教育领域与心理咨询专家进行评估，重新收集了中国被试在使用生成式 AI（或心理咨询）中关注的心理健康相关主题，使问题归类更合理，表述更口语化，并符合无感知对话。并在文章中进行补充说明（方法 2.3；表 2），以确保提示语设计具备语义自然性、文化适应性与教育可实施性。

意见 4:

Maurya（2023）里面涉及到的宗教/灵性的维度，并不适合中国国情，至少不适用于中国学生群体。针对这一维度的命名和相关内容的表述建议结合国情修改。

回应 4：

感谢审稿专家提出的关于“宗教/灵性”问题分类合理性的建议。我们已对原始分类体系进行了系统审视与调整，考虑到文章定位，将文章中的大学生群体改为适用于广泛群体。原类别“宗教/灵性”容易将深层心理机制误归于宗教或信仰层面。为更贴切反映所涉问题中对生命意义、自我定位、情感哀伤及存在体验的探讨，我们在修订稿中将该类问题归入“意义与存在”（Meaning & Existential Issues）。并对 Q13、Q14、Q15 分别做出相应修订（见文章中表 2）。此修订得到心理学与精神健康文献的充分支持。一方面，Frankl（Frankl, 1959）在“意义疗法（Logotherapy）”中强调：寻找生活意义是人类的基本动机，而非宗教信仰本身；另一方面，Yalom（1980）在其“存在主义心理治疗”理论中指出，个体在心理发展中普遍面临意义缺失、死亡焦虑与孤独等存在性挑战，呼应了我们研究中所包含的提示语内容。此外，Park（2010）等人

的研究亦表明，意义建构与存在体验在心理健康维持中扮演关键角色，其作用超越特定宗教取向，具备跨文化的适应性。因此，我们认为应使用“意义与存在”作为宏观分类替换原有“宗教/灵性”问题。

参考文献：

Frankl, V. E. (1959). *Man's Search for Meaning*. Boston: Beacon Press.

Park, C. L. (2010). Making sense of the meaning literature: An integrative review of meaning making and its effects on adjustment to stressful life events. *Psychological Bulletin, 136*(2), 257–301.

Yalom, I. D. (1980). *Existential Psychotherapy*. New York: Basic books.

意见 5:

表 5 中，相关性的男生得分、伦理考量上的男女得分都是 5，这个为何没有写标准差，另外均值没有保留两位小数？另外 U 检验的值有的写了小数有的没有，同一个表格中的书写风格要保持一致。

回应 5:

感谢审稿专家指出的表 5 中标准差缺失、均值小数位不统一、U 检验数值呈现风格不一致等问题，我们已在修订稿中作出如下修改。对于所有维度的评分项，均补充了标准差（$M \pm SD$）格式，所有均值保留两位小数，与其他表格保持一致；所有 Mann–Whitney U 检验的 U 值与 p 值统一采用带小数点格式，确保整表风格一致。

上述修订已应用于修订稿中结果章节 3.3 表 5，并同步调整了正文引用格式。非常感谢您的指正，帮助我们提升文章的严谨性。

意见 6:

“汉族与傣族、文科与理科情境下各评估标准的差异值均未达到统计学显著性（$p > 0.05$）”和“在其他几个评估标准上，男女无显著差异（$p > 0.05$ ）”这里的 p 应写成 ps，因为是对多个指标的检验。

回应 6:

感谢审稿专家指出关于统计报告中符号使用的规范问题。我们已对文中涉及多指标统计检验结果的表述做出如下修改：

原文中“汉族与傣族、文科与理科情境下各评估标准的差异值均未达到统计学显著性（$p > 0.05$）”已统一更正为（$p_s > 0.05$）；同样，“在其他几个评

估标准上，男女无显著差异（$p > 0.05$）”亦已修正为“（$p_s > 0.05$）”。上述修改已在文中结果章节 3.3 和表 5 中同步完成。非常感谢您的指正，帮助我们提升了统计写作的学术规范性。

再次感谢专家们提出的以上宝贵意见，为本研究的实验设计与方法的优化、文章严谨性等诸多方面提供了重要保障。

第二轮修改

诚挚感谢编辑部老师们的辛勤付出和专家的宝贵意见，下面根据专家建议逐一说明修改之处，所有对应修改内容已在文章中用红色字体标注。

意见 1：

摘要部分中，研究背景和研究目的之间存在一定的逻辑断层。传统心理健康教育的缺陷，如何直接引出本研究针对生成式 AI 在语义稳定性、回答策略及偏差等方面进行细致具体分析的必要性和重要性？该逻辑关联需要进一步清晰呈现和补充阐明。

回应 1：

非常感谢专家对文章中逻辑断层的建议，我们充分认同，摘要部分确实在“传统不足”到“语义分析的必要性”衔接不够恰当，已经对应在摘要、引言中补充对应逻辑链，并在讨论中呼应引言中提出的问题。

意见 2：

作者通过大量前沿或传统技术对研究主旨问题进行了深入分析，但为何能够将此称为“多层次”（纵向）语义分析，而非“多方面”（横向）语义分析？这一表述的依据有待在文中进一步阐明。

回应 2：

非常感谢专家对术语表述精准性的精准建议。我们已认真审视“多层次语义分析”一词在本研究语境中的适用性。我们充分认同，尽管本研究在方法流程中包含若干模块间的执行先后顺序，但整体框架并不构成严格意义上的“层次性结构”，不具备深浅递进逻辑特征。因此“多层次”一词确实不够严谨。

经重新评估，我们认为本研究提出的语义分析框架更符合“综合性”的划分。为提升术语表达的准确性和框架内涵的可理解性，我们已将原文中的“多层次语义分析框架”调整为“**语义行为综合分析框架（Comprehensive Semantic Behavior Analysis Framework, CSBAF）**”。具体而言，语义学是语言学的一个分支，而语义分析技术在处理自然语言和辅助机器学习方面至关重要（Salloum et al., 2020; Li et al., 2018）。本研究采用的“语义行为”一词，是指生成式 AI 在面对相同输入（如提示语）时，在多轮次生成中所表现出的结构分布模式的稳定性与策略变异特征，同时涵盖其回答内容的语义表达在不同社会

情境下的语义差异与不同模型之间所展现出的语义适应性。在言语行为理论视角下，“言语行为”是人类信息表达的形式，是一种社会行为，反映个体在特定环境、任务中的反应策略、语义加工路径与内容组织方式，不仅依赖内部认知加工过程，如语言选择、语法编码（Levelt, 1999），也受社会情境、交互目标的共同作用（Skinner, 1957）。本研究借鉴这一视角，提出“语义行为”作为观测 LLMs 输出内容稳定性与策略倾向的分析单元，并构建相应综合评估框架。因此，“语义行为”不仅指语言生成的语义内容本身，而且是基于“语言行为理论”视角对 AI 的“语言”生成行为在语义层面上的系统性观察与建模分析。该表述既能够涵盖本研究在语义生成建模、结构相似性评估、策略倾向识别中各模块的先后顺序，又能够涵盖偏差分析模块在多个维度上的整合性探索，强调其综合、多视角、跨平台的分析特征。

我们在文章引言部分对该术语进行了重新定义，并对文中所有相关表述进行了对应修改，以帮助读者更清晰地理解其应用边界与理论构成

参考文献：

Levelt, W.J.M. (1999). *Models of word production*. Trends in Cognitive Sciences, 3(6), 223–232.

Li, L., Zhou, C., He, J., Wang, J., Li, X., & Wu, X. (2018). Collective semantic behavior extraction in social networks. Journal of Computational Science, 28, 236–244.

Salloum, S. A., Khan, R., & Shaalan, K. (2020). A survey of semantic analysis approaches. In Proceedings of the International Conference on Artificial Intelligence and Computer Vision (AICV2020) (pp. 61–70). Springer International Publishing.

Skinner, B.F. (1957). *Verbal Behavior*. Copley Publishing Group.

意见 3：

关于“定性内容分析”，通常被认为属于质性研究范畴。文章中的具体部分何以体现采用了定性内容分析方法？以及为何使用该表述，建议在文中进一步予以明确说明。

回应 3:

非常感谢专家对文章表述准确性的建议，我们充分认同文章中关于偏差检验的部分，主要目的在于评估其在不同社会情境下是否存在回答内容上的偏差，而并非通过的质性分析流程重新提取内容中的概念，故严格意义上讲并非

采用定性分析法。具体而言，文章主要对 Maurya 等人（2023）编制的定性内容分析编码框架（见附录 3A）进行本土化修订，作为研究使用的内容评分手册（见附录 3B），随后，采用专家评分法分别从每个组别抽取三组文本进行打分。对此，我们对应修改方法章节 2.3、2.4.5 中涉及到的内容分析方法对应修改和补充为**“专家评分法”**。并对全文中涉及“定性内容分析”的章节及图片进行修改，更正为“基于内容分析编码手册对文本的内容进行评分”。

意见 4：

讨论部分目前主要停留在“就结果论结果”的层面，应进一步深入，将研究发现与生成式 AI 在心理健康教育中的关键议题和发现进行直接对话，从而更加突出研究成果的学术价值和方法论上的潜在贡献。

回应 4：

非常感谢专家对讨论部分的建议。我们已对文章中讨论章节进行修改，缩减对结果的讨论，突出研究成果与生成式 AI 在心理健康教育领域核心议题的对话。同时，明确了本研究在方法论方面的创新贡献，指出 CSBAF 框架在语义行为建模与策略识别方面的实用价值，强化了研究的学术定位与推广潜力，并指出现阶段研究的不足之处和未来方向。

再次感谢专家提出的以上宝贵意见，为本研究的内容质量、文章严谨性等诸多方面提供了重要保障。