新闻与传播评论 JCR
Journalism & Communication Review
第 79 卷 第 1 期 2026 年 1 月
Vol. 79 No. 1 Jan. 2026. 025-039

DOI:10. 14086/j. cnki. xwycbpl. 2026. 01. 003

谁更像“人”？模型类型与人格设定对大语言模型复刻传播学实验准确率的影响

曾秀芹　陈珂璐

摘要:以 ChatGPT-4o、DeepSeek-R1、豆包-1. 5、Kimi-K1. 5 四种主流大语言模型为对象,采用 2(有 vs. 无大五人格设定)×4(模型类型)实验设计,构建虚拟被试资料以复刻新闻传播学实验。结果显示,各模型拟合表现存在差异:DeepSeek-R1 在模拟真人平均趋势与行为变异性方面最优;ChatGPT-4o 的总体方差拟合偏差较大,但主效应、间接效应复刻的准确性与稳定性较为突出。人格设定的影响方面,无大五人格组描述性统计更贴近真人,而有大五人格组因果效应复刻更稳定,唯一完整复刻两个主效应的模型即来自该组。中介效应复刻成功率偏低,但人格设定可在一定程度上缓解模型输出的方向与效应偏离趋势。此外,研究基于 ChatGPT-4o 进一步发现,实验对象类型(真人组 vs. 无大五人格组 vs. 有部分大五人格组 vs. 有全部大五人格组)对主效应与中介机制部分产生显著调节作用,其中大五人格设定可一定程度抑制模型极端响应。研究实现多模型横向比较与人格设定控制下的复刻实验,验证了大五人格设定对提升模型模拟精度的积极作用,同时指出模型复刻复杂心理机制的局限性,推动传播学实验向“人机共演”的新范式转变,也拓展“媒介即延伸”在智能传播语境下的现实外延。

关键词:大语言模型;大五人格;复刻研究;硅基被试

中图分类号:G206　　**文献标志码**:A　　**文章编号**:2096-5443(2026)01-0025-15

基金项目:国家社会科学基金重点项目(22FXWA001)

人工智能的飞速发展正深刻重塑社会科学研究范式,尤其是大语言模型(Large Language Model,简称“LLM”)的迅猛进步,为学术界带来前所未有的机遇与挑战。LLM 已从工具性存在,逐步迈向能够模拟人类认知、情感与行为的智能体角色,成为社会科学研究的独特助力。其在内容分析、文本编码方面的能力可辅助研究与创作,推动研究方法革新,还能模拟个体与群体行为、完成问卷填答等任务[1-2],在实验复刻及决策预测中发挥深层作用。然而,LLM 能否真正重现人类认知机制仍是关键难题。

这些对于新闻传播学研究也具有重要意义。首先,当前学科正面临严峻的方法论危机:被试招募难、样本代表性不足导致结论难以推广,共同方法偏差影响研究效度,且传统实验耗时耗力、成本居高不下。“硅基被试”可能在降低成本、提升实验可重复性、优化样本代表性等方面展现出重要价值。其次,针对难以预测对目标受众的影响的核心难题,可以通过为 LLM 注入目标受众的人格设定,模拟不同框架、情绪与叙事下的受众反应,构建可预测传播影响的模型,助力传播从业者前置评估传播效果,进而优化内容与策略。同时,传播学研究对象始终与技术演进紧密相连,LLM 不仅是新工具,更在某种程度上实现了传播主体的再生产与延展。

本研究选用 ChatGPT-4o、DeepSeek-R1、豆包-1. 5 和 Kimi-K1. 5 四种主流的 LLM,对一项传播学领域的典型实验进行复刻,探索 LLM 作为“硅基被试”的可行性,推动人机共演的传播学实验方法的范式更新,进而为破解学科方法论困境提供新思路与解决方案。

一、文献综述及研究问题提出

（一）大语言模型模拟人类认知与行为的潜力

随着LLM的快速发展，社会科学研究迎来了新的工具与方法，相关研究不断揭示其在多个认知领域的潜力。Argyle等证明GPT-3能准确地模拟人类社会行为、政治态度与人口分层结构，为社会科学开启了用LLM替代人类样本的新研究范式[3]。Suri等发现，GPT-3.5在锚定效应等决策偏误上的表现与人类非常接近[4]。另有学者基于Psych-101数据集构建了Centaur模型，通过160个心理实验任务成功预测了人类认知模式与偏差[5]，首次大范围地显示出其与人脑活动对齐的潜力。这些研究共同表明，LLM已初步具备模拟人类受众反应的能力，展现出复刻社会科学实验的潜力。

（二）科研领域内大语言模型复刻实验的主要研究发现

由于LLM已初步具备类人类认知与行为模式的潜力，过去两年LLM被广泛应用于教育、科研、医疗和经济金融等领域[6]。以经济金融领域为例，Shapira等让LLM参与语言提示情境下的经济博弈，发现当训练数据量充足时，使用LLM生成的数据来训练预测模型，其准确率甚至高于由真人训练所得的数据集[7]。在市场营销领域，Yeykelis等复刻了*Journal of Marketing Research*上的133个实验，主效应复刻率达76%，但交互效应复刻成功率较低，仅为27%[8]；Li等验证了LLM在品牌感知分析中的有效性，结果显示LLM的结果与人类调研的结果高度拟合，且与人工调查发生行为的一致性超过75%[9]。政治领域也已开始借助LLM进行创新研究，Bachmann等利用GPT-4.0生成瑞士政治问卷中的虚拟选民数据，并将其用于预训练自适应问卷模型，结果表明在政治调研中，LLM有助于显著缓解冷启动问题，并提高早期预测的准确度[10]。在心理学研究中，有学者用GPT-4.0大规模复刻了156个心理学实验，发现72.7%的主效应与45.7%的交互效应被成功复刻[11]；Cho等提出与验证Doppelgänger模型可以通过学习人类对话来模拟问卷回答，显示出用LLM替代部分人类调查的潜力[12]。总体而言，LLM被视为构建“虚拟实验室”的重要工具，但在处理复杂的法律与心理学任务时仍存在显著不足，尤其在涉及种族、民族等问题与交互效应时复刻成功率偏低[8,11,13]。

（三）人格设定与大语言模型个性化模拟

诸多研究证实，人口统计学特征、性格特质等角色变量会影响LLM的表现。有学者构建了320个人格设定各异的LLM，证实ChatGPT-3.5与4.0版本在人格测验和写作风格中能体现差异化人格特质，真人可通过LLM生成的故事在一定程度上识别出其表现的人格特征[14]；Yuta等也验证，大五人格设定会影响LLM的检索查询生成，体现在查询长度与词汇选择上[15]。由于传统实验被试兼具多元人口统计学特征与人格特质等，研究者愈发关注LLM能否实现个性化实验模拟。Hu等发现，人格变量可解释不超过10%的标注方差，且其与人类标注的相关性越高，LLM基于人格提示的预测准确度也越高[16]。Petrov等则发现不同LLM有不同的“个性”，可通过微调改变[17]。然而，关于注入大五人格设定能否提升LLM复刻传播学实验的表现，尤其在模拟人类的复杂心理机制方面，仍缺乏系统实证检验。

（四）问题的提出

尽管近年LLM已被用于市场营销调查模拟[9,15]、政治学民意调查[10]、经济金融实验[7]及心理学研究[4-5,11]，但系统性探索其在传播学研究中的应用表现仍较为缺乏。同时，虽有研究显示LLM能在一定程度上模拟大五人格的总体特质分布，但其生成的人格标记难以反映现实人群的高度异质性。有学者发现使用通用人格比输入真实人口统计信息的硅基人格表现更优，却均无法模拟个体层面真实细腻的心理特质[17]，且LLM人格标记也高度依赖提示词设计，表达方式稍有不同便可能导致测量结果迥异，影响其稳定性与可靠性。此外，尽管LLM复刻社科实验已成研究热点，但相关研究多聚焦单一模型检验，鲜少系统比较国内外不同LLM在相同实验任务中的表现差异，而各模型在训练数据、架构及语言理解能力上的不同，可能使其复刻结果在准确度、稳定性上存在差异。

基于此，本研究拟使用当前最主流的大模型ChatGPT-4o以及在*Top100 Gen AI Consumer Apps*报

告中国内大模型排名长居前三的 DeepSeek-R1、豆包-1.5 和 Kimi-K1.5 四种 LLM,横向对比这四种 LLM 在复刻同一新闻传播学实验下的表现差异,同时结合大五人格提示,以探索被试人格设定和 LLM 类型对模拟真实人类实验的影响,及其模拟人类认知与行为的边界条件与潜能,为社会科学合理运用 LLM 助力科研提供参考。具体研究问题如下:

问题 1:四种主流的大语言模型(ChatGPT-4o、DeepSeek-R1、豆包-1.5 和 Kimi-K1.5)在复刻传播学实验时,哪一个更贴近真人的实验结果?

问题 2:为大语言模型注入大五人格特质信息,是否能够提升复刻人类实验结果的准确性?

二、研究设计

本研究拟按照以下三个步骤开展:第一步为被试资料编写与测试;第二步为复刻实验以获得实验数据;第三步为结果分析。接下来详细介绍上述步骤。

(一)被试资料编写与测试

本研究复刻的实验来自 2024 年《新闻大学》中的一篇论文[18],论文探讨了社交媒体广告的信息呈现方式对分享意愿的作用机制,并探究第三人效果在其中的中介作用。选择该研究基于三点:一是实验涵盖传播学经典理论、传播心理等核心要素,且具备典型实验范式特征,能有效检验 LLM 在复刻传播学实验中的拟人化表现。二是该实验由本研究团队成员主持完成,研究者既对原始设计逻辑与实验材料的语境条件有充分了解,又可获得所有被试的人口统计学资料和详细实验结果,能保证复刻研究与原实验在程序、材料、语义层面的一致性,也便于与硅基被试资料对齐。三是发表的期刊在新闻传播学科中具有良好的声誉,为研究质量提供了支撑。

基于原始实验所收集的被试资料,研究团队编制了两份略有不同的实验被试资料:其中一份仅包括原始资料中的人口统计学信息;另一份则在人口统计学信息基础上,额外附加完整的大五人格量表(NEO-FFI)分数,该量表共 60 题[19]117-122,并且被试的 NEO-FFI 分数根据其个人特征分布于中国大学生常模区间内[20]。每一组的复刻样本量均与原始研究未剔除无效样本时的样本量相同(N=200)。这样 2(无大五人格设定 vs. 有大五人格设定)×4(ChatGPT-4o vs. DeepSeek-R1 vs. 豆包-1.5 vs. Kimi-K1.5)一共生成 8 组复刻实验。为了更细致地探测人格设定对复刻效果的影响,同时减少复刻实验的巨大工作量,就中介效应的整体模型进行探索性研究,从 2 个操纵水平增加到 3 个操纵水平,分别为无大五人格、部分大五人格和完整大五人格。具体模型假设图参见附录中的图 1a 和图 1b①。

在正式复刻实验开始之前,本研究对实验材料进行了测试和必要的调整。在预实验中,发现 LLM 对一些词汇的理解具有偏差,因此在尽可能保留原始实验材料的基础上,通过参考 LLM 的内部隐藏状态补充了一些解释性说明[21]。此外,从 LLM 的问卷回答中可以观察到,其在整个作答过程中始终基于预设的人口统计学资料和人格特征来生成反应。

(二)实验步骤

研究拟选用 ChatGPT-4o、DeepSeek-R1、豆包-1.5 和 Kimi-K1.5 四种大语言模型,其原因在于:首先,2025 年 3 月相关报告显示,DeepSeek、豆包及 Kimi 稳居中国区前三名,而 ChatGPT 作为国际参照对象,其综合表现居于全球大语言模型领先地位[22];其次,这四种 LLM 分别采用差异化的技术路线与训练策略,系统覆盖通用应用、专业推理、办公辅助及垂直行业等主要应用方向,以确保研究样本具有充分的代表性;最后,这四种模型均开放调用接口或提供测试平台,可访问性与运行稳定性高,适用于大规模实验验证与研究复刻。

正式数据收集与实验于 2025 年 5 月 23 日—6 月 7 日开展,研究采用逐个 AI 进行实验的方式,数据收集周期均为 3~5 天。实验期间,LLM 的温度参数为默认值,以兼顾响应的多样性、连贯性和

①模型假设图、流程图等图表均见附录,获取链接为:https://osf.io/d3q7r/overview?view_only=84ffc43993b84cf4904e915a0ac60635。复制获取链接到浏览器中时,请注意删除字母符号间的多余空格。

相关性[11]，并统一采用网页版，且无深度思考与联网设置，以控制因配置与外部工具导致的偏差。实验步骤为：首先，在每一轮调查前均新开一个对话框，在对话框内分别输入包含完整大五人格设定、部分大五人格设定和无大五人格设定的资料，三组被试独立且轮流完整完成对应任务，避免并行测试的资源干扰与策略混淆，同时保证其间无版本更新与架构迭代，减少系统性偏差；随后，仿照所复刻的实验流程发送实验刺激物材料及辅助理解的提示词以减少干扰变量；最后，发送问卷要求 LLM 依据实验材料反应、分析，仔细填写并进行数据收集。因篇幅所限，具体流程见附录中的图 2 与图 3。

（三）结果分析

复刻实验在每个平台分别完成有全部大五人格被试资料和无大五人格被试资料的实验各 200 次，四种 LLM 共 1600 次；后采用部分大五人格被试资料进行实验 200 次，累计完成 1800 次被试资料的输入与输出，最终获得了 1800 份有效问卷。其中本研究复刻了原始实验的 2 个主效应、2 个中介效应，8 组复刻实验共统计分析了 16 个主效应与中介效应。对复刻实验所得数据的分析，完全遵循原始实验的分析步骤、使用的统计工具及流程，并以显著性、效应方向、效应量与置信区间作为评估维度[23-24]，在保持对原始实验结果最大程度忠实的前提下来衡量复刻实验的准确性。

三、复刻实验结果分析

（一）测量变量的描述性统计对比分析

1. 四种大语言模型复刻平均值（标准差）与真人实验对比分析

如表 1 所示，在均值层面，除 Kimi-K1. 5 的第三人效果外，真人组在第三人效果（$M_{真人}$ = -0. 16）与分享意愿（$M_{真人}$ = 3. 15）两个变量上的得分均低于各大语言模型组。DeepSeek-R1 模型在两项变量综合上与真人组均值最为接近（平均值差值为 0. 29，分别为 0. 20 与 0. 09）；其次是豆包-1. 5（平均值差值为 0. 68，分别为 0. 57 和 0. 11）和 Kimi-K1. 5（平均值差值为 1. 22，分别为 0. 33 和 0. 89）；ChatGPT-4o 则偏离最大（平均值差值为 1. 55，分别为 0. 36 与 1. 19）。在标准差方面，真人组在第三人效果（$SD_{真人}$ = 1. 85）与分享意愿（$SD_{真人}$ = 1. 61）上均呈现出显著更高的个体差异，而大语言模型组内部波动较小。从两个变量的标准差差值来看，DeepSeek-R1 在两项变量综合上（标准差差值为 1. 27，分别为 0. 31 和 0. 76）最接近真人；其次是豆包-1. 5（标准差差值为 1. 38，分别为 0. 63 和 0. 75）；最后是 Kimi-K1. 5（标准差差值为 1. 71，分别为 1. 03 和 0. 68）与 ChatGPT-4o（标准差差值为 1. 72，分别为 0. 72 和 1. 00）在两个变量的综合值上表现出更低的标准差，显示其对个体差异的模拟能力最弱。整体而言，大语言模型在标准差层面普遍低估了人类行为的变异性。

表 1　四种大语言模型复刻平均值（标准差）与真人实验对比结果

测量变量	ChatGPT-4o		DeepSeek-R1		豆包-1. 5		Kimi-K1. 5		真人实验	
	M	SD	M	SD	M	SD	M	SD	M	SD
第三人效果	0. 20	1. 13	0. 04	2. 16	0. 41	1. 22	-0. 49	0. 82	-0. 16	1. 85
分享意愿	4. 34	0. 61	3. 24	0. 85	3. 26	0. 86	4. 04	0. 93	3. 15	1. 61

总体来看，DeepSeek-R1 在模拟平均值和标准差的表现与真人水平最贴近，其次是豆包-1. 5 和 Kimi-K1. 5 模型，ChatGPT-4o 偏离真人水平最显著。

2. 有 vs. 无大五人格设定复刻平均值（标准差）与真人实验对比分析

表 2 综合显示，在均值层面，无大五人格组在两个变量综合上更贴近真人表现，其平均值差值亦略低于有大五人格组。在标准差维度上，无大五人格组在两个变量的总体上更接近真人的行为变异性（标准差差值分别为 0. 35 和 0. 62）。总而言之，无大五人格设定的 LLM 在整体均值和标准差两个维度上更贴近真人组表现。

表 2　有 vs. 无大五人格设定复刻平均值(标准差)与真人实验对比结果

测量变量	有大五人格		无大五人格		真人实验	
	M	SD	M	SD	M	SD
第三人效果	0.13	1.41	−0.05	1.50	−0.16	1.85
分享意愿	3.69	0.92	3.74	0.99	3.15	1.61

(二)主效应检验的对比分析

1. 复刻成功率

本研究复刻了原始研究的两个有明确方向的主效应,共用 4 种 LLM 将被试资料输入成有和无大五人格进行 16 次实验。将复刻结果的显著性与效应方向与原实验对比,若结果一致,则表明实验复刻成功。当考虑显著性($p<0.05$)且效应方向与原实验一致时,31.25%(16 例中的 5 例)的结果显示主效应被成功复刻。其中,有大五人格组的 8 个主效应中,ChatGPT-4o 的框架效应、效益目标与第三人效果均被成功复刻,成功率为 25%(8 例中的 2 例);无大五人格组的 8 个主效应中,ChatGPT-4o、Deepseek-R1、豆包-1.5 的效益目标与第三人效果被成功复刻,成功率为 37.5%(8 例中的 3 例)。仅考虑效应方向时,复刻实验主效应与原始研究方向相同的结果占 43.75%(有大五人格组 3 例、无大五人格组 4 例);仅考虑显著性($p<0.05$)时,62.5%的结果显著(两组各 5 例)。结果如表 3 所示:

表 3　各大语言模型组与真人组在信息框架和效益目标上对第三人效果影响的方差分析表

类型	变量	组别	M	SD	F	df	p	η^2
真人组	信息框架	积极框架	−0.25	1.91	8.17	1	0.008	0.038
		消极框架	0.19	1.72				
	效益目标	效益自己	−0.72	1.88	21.20	1	0.000	0.095
		效益他人	0.41	1.64				
ChatGPT-4o 有大五人格组	信息框架	积极框架	−0.08	1.14	4.33	1	0.039	0.021
		消极框架	0.26	1.17				
	效益目标	效益自己	−0.41	1.05	44.92	1	0.000	0.185
		效益他人	0.59	1.06				
DeepSeek-R1 有大五人格组	信息框架	积极框架	0.24	1.72	0.46	1	0.497	0.002
		消极框架	0.06	2.00				
	效益目标	效益自己	−0.10	2.01	3.63	1	0.058	0.018
		效益他人	0.40	1.69				
豆包-1.5 有大五人格组	信息框架	积极框架	0.59	1.15	3.88	1	0.050	0.019
		消极框架	0.95	1.42				
	效益目标	效益自己	0.52	1.38	7.62	1	0.006	0.037
		效益他人	1.02	1.17				
Kimi-K1.5 有大五人格组	信息框架	积极框架	−0.36	0.85	4.97	1	0.027	0.024
		消极框架	−0.62	0.80				
	效益目标	效益自己	−0.80	0.57	32.04	1	0.000	0.139
		效益他人	−0.18	0.94				

续表

类型	变量	组别	M	SD	F	df	p	η^2
ChatGPT-4o 无大五人格组	信息框架	积极框架	0.53	1.02	8.44	1	0.004	0.041
		消极框架	0.09	1.12				
	效益目标	效益自己	−0.35	0.99	115.26	1	0.000	0.368
		效益他人	0.97	0.73				
DeepSeek-R1 无大五人格组	信息框架	积极框架	0.04	2.24	0.49	1	0.485	0.002
		消极框架	−0.20	2.60				
	效益目标	效益自己	−1.28	2.47	64.89	1	0.000	0.247
		效益他人	1.12	1.67				
豆包-1.5 无大五人格组	信息框架	积极框架	−0.03	1.07	1.24	1	0.267	0.006
		消极框架	0.13	0.96				
	效益目标	效益自己	−0.20	0.94	12.82	1	0.000	0.061
		效益他人	0.30	1.03				
Kimi-K1.5 无大五人格组	信息框架	积极框架	−0.49	0.72	0.01	1	0.930	0.000
		消极框架	−0.48	0.88				
	效益目标	效益自己	−0.71	0.66	17.01	1	0.000	0.079
		效益他人	−0.26	0.87				

2. p 值

关于主效应的 p 值方面，LLM 复刻的实验除 ChatGPT-4o 无大五人格组外，框架效应对第三人效果的影响的 p 值均大于真人组的 p 值（$p_{\text{ChatGPT-4o无大五人格组}} = 0.004$，$p_{\text{真人组}} = 0.008$）；除 DeepSeek-R1、豆包-1.5 有大五人格组外，效益目标对第三人效果的影响的 p 值均相同（$p<0.001$）。这表明 LLM 复刻实验观察到的框架效应对第三人效果的影响较弱，但在另一个自变量效益目标上，复刻出来的实验结果与原始实验在显著性上是较为一致的，具体数据请见附录中的表 1。

3. 效应量与置信区间

研究比较了 LLM 在复刻框架效应与第三人效应以及效益目标与第三人效应的效应量的表现。根据表 3 中的数据，在框架效应与第三人效应上，所有 LLM 产生的 η^2 值普遍较小，与真人组（$\eta^2 = 0.038$）接近，显示 LLM 在该效应拟合上较为稳定（$M_{\eta^2\text{框架效应与第三人效应}} = 0.014$）。然而在效益目标与第三人效应上，LLM 的 η^2 值波动较大，其中有 50%组别的 η^2 值大于真人组，因此在该效应上存在更大的系统性偏离，部分模型呈现明显放大效应的倾向（$M_{\eta^2\text{效益目标与第三人效应}} = 0.142$，$\eta_{\text{真人组}}{}^2 = 0.095$）。进一步对不同 LLM 在复刻两项主效应时的表现进行分析，发现各 LLM 在效应值上与真人组之间均存在一定程度的差距，但差距大小因模型而异。对比图详见附录中的图 4。根据附录中的表 2 所示，ChatGPT-4o 有大五人格组是唯一在两个主效应上均被成功复刻，两项平均差值的均值为 0.054，显示出较为稳定且贴近真人结果的复刻表现。相比之下，其余 LLM 仅在效益目标或第三人效应中出现一次，其中以豆包-1.5 无大五人格组的表现最佳，绝对差值仅为 0.034，是所有单项结果中最接近真人的数据。其他无大五人格模型的差距相对较大，如 ChatGPT-4o 无大五人格组与真人组的差距高达 0.273，而 DeepSeek-R1 无大五人格组亦达到 0.152，偏离更明显。从整体来看，将无大五人格组三个模型的差值取平均，其平均差值为 0.153，显著高于有大五人格组的差值 0.054。虽然个别无大五人格模型在显著性次数多于有大五人格，但有大五人格量表设定的 ChatGPT-4o 在复刻实验效应时更具全面性和准确性。根据附录中的图 5a、图 5b 可知，在置信区间方面，原始实验的两个主效应的

效应量落在复刻实验95%的置信区间的占比为43.75%(7例),其中有大五人格组有4例(50%),分别为框架效应与第三人效应上的ChatGPT-4o、豆包-1.5、Kimi-K1.5与效益目标与第三人效应上的Kimi-K1.5;无大五人格组有3例(37.5%),分别为框架效应与第三人效应上的ChatGPT-4o与效益目标与第三人效应上的豆包-1.5、Kimi-K1.5。原始效应量高于复现95%的置信区间的上限占比为37.5%(6例),低于复现95%的置信区间的下限的占比为18.75%(3例)。同时,复刻实验有68.75%(11例)的置信区间比原始研究的置信区间较窄。

(三)中介效应检验

1. 复制成功率

按照复刻研究的既定规范,当效应方向和显著性与原始研究中真人组一致时认为复刻成功。总体来看,中介效应有12.5%的复刻成功率(Kimi-K1.5有大五人格组、ChatGPT-4o无大五人格组在路径1上的中介效应被成功复刻)。其中有大五人格组和无大五人格组每组的8个效应中都仅有1组复刻成功,复刻成功率均为12.5%。但不考虑统计显著性时,复刻实验中介效应与原始研究方向相同的结果占18.75%(有大五人格组2例、无大五人格组1例);仅考虑复刻显著性($p<0.05$)时,50%的结果显著(有大五人格组6例、无大五人格组2例)。结果如表4所示:

表4　各大语言模型组与真人组在各间接效应的路径对比检验

组别	路径	Coeff	SE	BC Bootstrap95%置信区间		DE(p)
				下限	上限	
真人组	X1→M→Y	0.12**	0.07	0.009	0.278	0.036
	X2→M→Y	0.16**	0.09	0.008	0.357	0.662
ChatGPT-4o有大五人格组	X1→M→Y	−0.04**	0.02	−0.095	−0.003	0.004
	X2→M→Y	−0.15**	0.05	−0.258	−0.070	0.637
DeepSeek-R1有大五人格组	X1→M→Y	0.02	0.03	−0.040	0.083	0.000
	X2→M→Y	−0.05	0.03	−0.125	0.001	0.658
豆包-1.5有大五人格组	X1→M→Y	−0.07**	0.05	−0.185	−0.001	0.000
	X2→M→Y	−0.12**	0.05	−0.229	−0.029	0.545
Kimi-K1.5有大五人格组	X1→M→Y	0.07**	0.03	0.007	0.143	0.001
	X2→M→Y	−0.12**	0.06	−0.238	−0.016	0.496
ChatGPT-4o无大五人格组	X1→M→Y	0.04**	0.02	0.006	0.096	0.046
	X2→M→Y	−0.11	0.07	−0.237	0.021	0.881
DeepSeek-R1无大五人格组	X1→M→Y	−0.01	0.01	−0.038	0.017	0.000
	X2→M→Y	−0.04	0.08	−0.193	0.119	0.0004
豆包-1.5无大五人格组	X1→M→Y	−0.02	0.02	−0.075	0.018	0.439
	X2→M→Y	−0.07**	0.04	−0.163	−0.007	0.671
Kimi-K1.5无大五人格组	X1→M→Y	−0.002	0.03	−0.058	0.046	0.683
	X2→M→Y	−0.07	0.04	−0.164	0.014	0.314

***、**、*分别表示在0.001、0.01、0.05的水平下显著

2. 效应量与置信区间

整体而言,真人组在路径1(框架效应X_1、第三人效应M、分享意愿Y)上呈现显著且正向的中介效应,相较之下,Kimi-K1.5有大五人格组在该路径上呈现出显著且较贴近原始研究的正向效应值,

与真人效应方向一致,其他 LLM 多在该路径上产生相反或极小的效应量。在路径 2(效益目标 X_2、第三人效应 M、分享意愿 Y)上,真人组同样呈现出显著正向的中介效应,但几乎所有 LLM 在该路径上的效应方向均为相反方向的负数。对比图详见附录中的图 6。

根据附录中的图 7a、图 7b 的对比图可知,原始效应量落在复刻实验 95%的置信区间的只有 Kimi-K1.5 有大五人格组中的路径 1。原始效应量高于复现 95%的置信区间的上限占比为 93.75%(15 例)。同时,所有复刻实验的中介效应的置信区间比原始研究的置信区间都窄。

(四)人格设定对研究模型复刻效果影响的探索性研究

为检验实验对象类型(人格设定)在路径 1(框架效应 X_1、第三人效应 M、分享意愿 Y)的复刻效果上的调节作用,本研究采用 PROCESS 4.3 的 Model8 进行数据分析。实验对象分为四类:真人、无大五人格的 LLM、具备部分大五人格的(仅神经质 N、开放性 O 维度)LLM、具备完整大五人格的 LLM。选择 N、O 维度的依据为二者分属稳定性与可塑性中阶元因素,且相关性在大五维度中最低[25]。同时参考过往文献[16]及本研究复刻中介效应结果最优的 ChatGPT-4o 模型,基于该模型完成 200 次实验,设定为部分大五人格组。分析数据见表 5。

表 5　路径 1(框架效应 X_1、第三人效应 M、分享意愿 Y)上的有调节的中介模型检验

回归方程(N=1788)		拟合指标			系数显著性		
结果变量	预测变量	R	R^2	F	B	SE	t
第三人效果		0.73	0.55	340.17***			
	框架效应				−1.22	0.15	−8.09***
	W1(无大五人格)				0.11	0.08	1.28
	W2(有部分大五人格)				1.92	0.11	18.37***
	W3(有大五人格)				0.29	0.84	3.44***
	框架效应 * 无大五人格				3.46	0.17	20.67***
	框架效应 * 有部分大五人格				1.24	0.21	5.88***
	框架效应 * 有全部大五人格				3.34	0.17	19.91***
分享意愿		0.56	0.31	112.16***			
	框架效应				−0.56	0.15	−3.85***
	第三人效果				−0.16	0.02	−7.34***
	W1(无大五人格)				0.60	0.08	7.54***
	W2(有部分大五人格)				−1.24	0.11	−11.44***
	W3(有全部大五人格)				0.58	0.08	7.30***
	框架效应 * 无大五人格				1.07	0.18	6.06***
	框架效应 * 有部分大五人格				0.51	0.20	2.51**
	框架效应 * 有全部大五人格				0.83	0.17	4.77***

注:"***""**""*"分别表示在 0.001、0.01、0.05 的水平下显著

在第三人效应回归方程中,框架效应与实验对象类型的交互项预测作用显著(框架效应 * 无大五人格:$B_{第三人效果}=3.46$, $t_{第三人效果}=20.67$, $p<0.001$;框架效应 * 有部分大五人格:$B_{第三人效果}=1.24$, $t_{第三人效果}=5.88$, $p<0.001$;框架效应 * 有全部大五人格:$B_{第三人效果}=3.34$, $t_{第三人效果}=19.91$, $p<0.001$),表明其可调节框架效应对第三人效果的影响。简单斜率分析显示(见附录图 8a 与表 3):真人组中框架效应对第三人效应呈显著负向预测(simple slope=−2.44, t=−8.09, $p<0.001$);无大五人格 LLM 组

(simple slope=4.49,t=30.69,p<0.001)与全部大五人格 LLM 组(simple slope=4.23,t=28.94,p<0.001)中均呈显著正向预测,即消极框架增强 LLM 中的第三人效应;而部分大五人格 LLM 组中该直接效应不显著(t=0.11,p>0.05)。

在分享意愿的回归模型中,框架效应与实验对象类型的交互项同样显著(框架效应 * 无大五人格:$B_{分享意愿}=1.07$,$t_{分享意愿}=6.06$,$p<0.001$;框架效应 * 有部分大五人格:$B_{分享意愿}=0.51$,$t_{分享意愿}=2.51$,$p<0.01$;框架效应 * 有全部大五人格:$B_{分享意愿}=0.83$,$t_{分享意愿}=4.77$,$p<0.001$)。简单斜率分析进一步显示(见附录图 8b 与表 3),在真人组中,框架效应对分享意愿的预测为负向且显著(simple slope=-1.12,t=-3.85,p<0.001),而在有全部大五人格 LLM(simple slope=0.54,t=3.27,p<0.01)以及无大五人格 LLMs(simple slope=1.01,t=5.97,p<0.001)组中则为正向预测作用,表明有全部大五人格和无大五人格组对于真人组,在框架效应对分享意愿的路径上同样表现正向调节作用,但仅有部分大五人格的 LLM 组中框架效应对分享意愿的直接效应不显著(t=-0.40,p>0.05)。

因此,在第三人效应与分享意愿回归模型中,无大五人格的正向调节效应均较有大五人格的高。此外,除有部分大五人格 LLM 组外,框架效应对分享意愿的直接效应及第三人效果的中介作用 95%置信区间的上下限皆未包含 0,具体见表 6。这表明其余两组的直接效应与中介效应皆显著。

表 6　路径 1(框架效应 X_1、第三人效应 M、分享意愿 Y)的不同组别上的直接效应和中介效应

	实验对象类型(真人 vs. 无大五人格 vs. 有部分大五人格 vs. 有全部大五人格)	效应值	SE	Boot CI 下限	Boot CI 上限
直接作用	真人	-0.57	0.15	-0.85	-0.28
	无大五人格 LLM	0.51	0.08	0.34	0.67
	有部分大五人格 LLM	-0.06	0.14	-0.33	0.22
	有全部大五人格 LLM	0.27	0.08	0.11	0.43
第三人效果的中介作用	真人	0.20	0.06	0.10	0.32
	无大五人格 LLM	-0.35	0.07	-0.49	-0.22
	有部分大五人格 LLM	0.00	0.01	-0.24	0.02
	有全部大五人格 LLM	-0.33	0.06	-0.46	-0.21

表 7　路径 2(效益目标 X_2、第三人效应 M、分享意愿 Y)上的有调节的中介模型检验

回归方程(N=1788)		拟合指标			系数显著性		
结果变量	预测变量	R	R^2	F	B	SE	t
第三人效果		0.45	0.20	72.14***			
	效益目标				-0.59	0.20	-2.95**
	W1(无大五人格)				0.12	0.11	1.05
	W2(有部分大五人格)				1.94	0.14	13.93***
	W3(有大五人格)				0.30	0.11	2.68**
	效益目标 * 无大五人格				1.48	0.22	6.65***
	效益目标 * 有部分大五人格				0.53	0.28	1.90
	效益目标 * 有全部大五人格				1.59	0.22	7.16***
分享意愿		0.57	0.32	117.17***			

续表

回归方程(N = 1788)		拟合指标			系数显著性		
	效益目标				-0.42	0.14	-2.93**
	第三人效果				-0.10	0.02	-6.37***
	W1(无大五人格)				0.60	0.08	7.58***
	W2(有部分大五人格)				-1.34	0.10	-12.86***
	W3(有大五人格)				0.57	0.08	7.21***
	效益目标 * 无大五人格				0.97	0.16	6.07***
	效益目标 * 有部分大五人格				0.56	0.20	2.83**
	效益目标 * 有全部大五人格				0.61	0.16	3.83***

注:"***""**""*"分别表示在 0.001、0.01、0.05 的水平下显著

研究进一步检验实验对象类型在路径 2(效益目标 X_2、第三人效应 M、分享意愿 Y)的复刻效果中的调节作用,结果如表 7 所示,在第三人效果的回归方程中,除部分大五人格组外,效益目标与其他实验对象类型的交互项预测显著(框架效应 * 无大五人格:$B_{第三人效果} = 1.48$,$t_{第三人效果} = 6.65$,$p<0.001$;框架效应 * 有部分大五人格:$B_{第三人效果} = 0.53$,$t_{第三人效果} = 1.90$,$p>0.05$;框架效应 * 有全部大五人格:$B_{第三人效果} = 1.59$,$t_{第三人效果} = 7.16$,$p<0.001$)。简单斜率分析表明(见附录图 9a 与表 4),真人组中效益目标对第三人效应呈显著负向预测作用(simple slope = -1.18,t = -2.95,p<0.01);全部大五人格 LLM 组(simple slope = 2.00,t = 10.33,p<0.001)与无大五人格 LLM 组(simple slope = 1.78,t = 9.17,p<0.001)中均呈显著正向预测;部分大五人格 LLM 组在该直接效应中也不显著(t = -0.31,p>0.05)。在分享意愿回归模型中,效益目标与实验对象类型的交互项均显著(框架效应 * 无大五人格:$B_{分享意愿} = 0.97$,$t_{分享意愿} = 6.07$,$p<0.001$;框架效应 * 有部分大五人格:$B_{分享意愿} = 0.56$,$t_{分享意愿} = 2.83$,$p<0.01$;框架效应 * 有全部大五人格:$B_{分享意愿} = 0.61$,$t_{分享意愿} = 3.83$,$p<0.001$)。简单斜率分析表明(见附录图 9b 与表 4),真人组中效益目标对分享意愿呈显著负向预测(simple slope = -0.84,t = -2.93,p<0.01);全部大五人格 LLM 组(simple slope = 0.39,t = 2.76,p<0.01)与无大五人格 LLM 组(simple slope = 1.11,t = 7.85,p<0.001)中均呈正向预测,即 LLM 相较于真人组,对该路径表现方向相反的调节效应;部分大五人格 LLM 组直接效应仍不显著(t = 1.04,p>0.05)。

有全部大五人格对第三人效应的回归模型的调节作用强于无大五人格,而无大五人格对分享意愿的回归模型的调节作用强于有全部大五人格。此外,除有部分大五人格 LLM 组外,效益目标对分享意愿的直接效应及第三人效果的中介作用 95% 置信区间的上下限皆未包含 0,具体见表 8。这表明其余两组的直接效应与中介效应皆显著。

表 8　路径 2(效益目标 X_2、第三人效应 M、分享意愿 Y)的不同组别上的直接效应和中介效应

	实验对象类型(真人 vs. 有大五人格 vs. 有部分大五人格 vs. 无大五人格)	效应值	SE	Boot CI 下限	Boot CI 上限
直接作用	真人	-0.42	0.14	-0.70	-0.14
	无大五人格 LLM	0.55	0.07	0.42	0.69
	有部分大五人格 LLM	0.14	0.14	-0.13	0.42
	有全部大五人格 LLM	0.20	0.07	0.06	0.34

续表

	实验对象类型(真人 vs. 有大五人格 vs. 有部分大五人格 vs. 无大五人格)	效应值	SE	Boot CI 下限	Boot CI 上限
第三人效果的中介作用	真人	0.06	0.03	0.01	0.13
	无大五人格 LLM	-0.09	0.02	-0.13	-0.05
	有部分大五人格 LLM	0.01	0.01	-0.01	0.02
	有全部大五人格 LLM	-0.10	0.02	-0.15	-0.06

值得注意的是,在除效益目标与第三人效应外的路径上,无大五人格的 LLM 展现出比有全部大五人格组更强的调节效应,且方向更偏离真人。这意味着虽引入大五人格参数却未必能完全复刻人类的心理反应方向和效应值,但可能在一定程度上抑制了复刻过程中产生的极端或放大效应。

四、总结与讨论

本研究系统复刻曾秀芹等的实验,包含两个主效应和两个中介效应。基于以往对大语言模型参与社会科学调查与实验的研究,探讨了四种主流的 LLM 在复刻传播学实验中的表现,并探讨了人格设定对复刻结果的影响。研究结果表明,在均值与标准差复刻上,DeepSeek-R1 模型与真人组最接近,且无大五人格设定比有大五人格设定更贴近真人实验。在主效应与中介效应复刻上,ChatGPT-4o 模型表现相对更优。引入大五人格设定的实验组,其综合复刻准确性在各评估指标上表现出优于无大五人格设定组的结果。在中介效应复刻中,当仅注入部分大五人格(神经质与开放性)两个维度设定时,模型复刻效果最差。

(一)本研究结果讨论

1. 在变量平均值与标准差上的复刻

不同模型的拟真能力存在显著分化。横向对比四种模型,DeepSeek-R1 在均值与标准差上均最接近真人组,其还原人类行为平均趋势与个体差异的结构能力更优。反观 ChatGPT-4o,均值与标准差均偏离真人最远,尤其是标准差呈高度收敛式,显著低估了人类行为的离散性。这种过度一致性构成社会行为复刻的结构性偏误,反映了模型对人类多样性的表达仍受限。就人格标注设定而言,无大五人格组在均值与标准差上更接近真人组,模拟人类反应多样性稍占优势;而有大五人格组则出现“均值上移,标准差下偏”的现象。从机制层面看,大五人格设定通过注入稳定特质向量、固定人格倾向的提示词调控输出,虽有助于引导生成内容风格,但经人类偏好训练的模型易出现“谄媚”现象,当提示塑造社会期许时,模型更倾向生成符合期许的谄媚性回答[26],导致均值向更肯定或更高方向偏移,在无意中压缩了模型内部的响应空间;同时,尽管提示词尽量贴合真实人格分布,但人工智能模型需假设一定同质性,通过借鉴多维空间中的类似情况进行推断,不可避免会忽略个体层面的残差变异性,使模型输出回归到方差减小的平均模式[27],这也解释了有大五人格组标准差为何整体低于无大五人格组,反映其表达复杂社会心理变异性的能力受限。

2. 主效应复刻成功率检验

就主效应的方向与显著性复刻而言,LLM 整体成功率偏低,仅有少数模型能够与原始实验结果保持一致。横向对比四种模型时,发现 ChatGPT-4o 是复刻成功率最高的模型,其次是豆包-1.5 和 DeepSeek-R1,最差的是 Kimi-K1.5,在主效应中无一次复刻成功。在对比有无大五人格对主效应的影响时发现 ChatGPT-4o 有大五人格组表现最突出,是唯一在两个主效应上均成功复刻方向与显著性的模型,显示出较强的因果判断能力与复刻稳定性。豆包- 1.5 无大五人格组虽在效益目标与第三人效应路径上最接近真人,但整体波动性大,且仅一路径复刻成功,缺乏一致性。虽然无大五人格

组复刻成功次数略高于有大五人格组，但后者在效应量与置信区间拟合上更优，更易覆盖原始效应量，拟合精度更高。不过多数 LLM 的置信区间宽度小于原始实验，反映其模拟个体差异、情境波动等自然变异性的能力有限，可能由于 LLM 生成机制本身的高度结构化与模式预测特性，使其更倾向于产生集中、低噪声的结果。综合来看，本研究所设定的复刻任务中，ChatGPT-4o 有大五人格组在主效应层面最接近真人表现，整体复刻效果最佳。同时人格设定对复刻结果确实产生了影响，有大五人格组在主效应拟合与置信区间与效应量表现上更具优势。

3. 中介效应检验

在中介效应检验中，四种主流 LLM 复刻中介效应的表现有限，整体成功率较低，复刻效应方向与显著性一致的情况较少。仅 Kimi-K1. 5 有大五人格组与 ChatGPT-4o 无大五人格组在路径 1（框架效应 X_1→第三人效应 M→分享意愿 Y）上成功复刻真人实验方向且中介效应显著；DeepSeek-R1 有大五人格组效应方向与真人一致，但中介效应不显著。总体而言，有大五人格组的复刻结果更稳定，方向偏离小，具有建构性优势。

4. 人格设定对复刻效果的调节

本研究进一步检验了实验对象类型（真人 vs. 无大五人格组 vs. 有部分大五人格组 vs. 有全部大五人格组）对复刻效果的调节作用，发现除有部分大五人格组，其他实验对象类型在两条路径中均表现出显著的调节作用。简单斜率分析表明，真人组在两条路径上均呈现负向效应，然而 LLM 无论是否搭载大五人格设定，多数在这两条路径上表现出显著的正向效应；从调节效应强度来看，在除效益目标与第三人效应路径，其他路径上无大五人格组的调节作用相比于全部大五人格组均更为强烈且方向更偏离真人，意味着虽然引入大五人格设定未必能完全复刻人类的心理反应方向，但它可能在一定程度上抑制了 LLM 产生极端或放大的效应。缺乏人格设定带来的“约束”，则导致更剧烈，且与真人心理规律不符的变化。部分大五人格（仅神经质与开放性）组整体表现劣于完整大五人格组与无大五人格组。相较于无大五人格组，赋予模型部分人格特质会人为强化特定角色，可能引发潜在偏见性输出，因为过度强调某些特质易放大训练数据中的刻板印象与偏见[28]。即模型仅被提示两个维度人格时，生成行为更易偏向两个维度的极端表现，对人类真实人格的模拟反而不及无大五人格组的平均化输出。此外，由于提示文本越短越容易受到提示词影响[29]，部分大五人格提示因信息量有限，反而对模型即时生成的引导性更强；而完整大五人格设定在多维度作用下更均衡，不易过度放大单一特质。因此，部分人格设定不仅未能提升模型的表现，反而削弱了模型的稳定性与真实性。

然而，无论有无大五人格设定，LLM 在模拟变量间关系、主效应以及中介机制等复杂心理与社会过程时，仍存在明显的局限性，尚难完全替代真人实验。

（二）本研究的理论意义与实践启示

1. 理论意义

本研究的理论贡献主要体现在四个方面：其一，以往关于 LLM 在复刻社会科学实验的应用研究，多聚焦单一模型或同一模型的不同代际的特定场景表现[14,30]，缺乏多模型之间系统的横向对比。本研究首次引入国内外四种主流的 LLM，在统一实验框架下开展横向复刻，全面评估其在均值、标准差、主效应、中介效应等层面的拟真能力，揭示不同模型生成机制与复刻表现的显著差异。其二，本研究将大五人格量表分数作为提示词引入实验复刻，验证了完整大五人格设定在一定程度上可增强模型对人类心理结构的拟合能力，尤其表现在提升主效应的稳定性与中介效应的一致性方面。该策略为优化 LLM 模拟个体、群体反应的精度提供了新路径，丰富了虚拟受试者研究的构念框架，亦验证了人格设定能在一定程度上提升模型输出与人类心理结构的契合度，在一定程度上增强大五人格相关心理特征的模拟准确性[31]。其三，相较其他学科，LLM 复刻传播学实验存在独特价值。拉斯韦尔 5W 模式中“Who”与“Whom”传统指向个人或组织，而 LLM 以拟人化被试身份参与实验时，不再仅是工具，更成为具备认知、人格与社会表达能力的智能传播受体，使“Whom”拓展为人机共存的智

能系统。这一创新延展了麦克卢汉“媒介即人类延伸”的命题:LLM 以人格体参与时,既延伸了人的表达能力,也一定程度上复制了人的社会心理逻辑,预示传播学研究将从人如何用媒介拓展至智能媒介如何模拟人。由此,传播主体结构被重构,传播行为与信息接收成为人机共演的过程,这为未来探讨智能媒介内容生成逻辑、社会传播角色定位等全新议题奠定基础。同时,本研究通过多模型在不同人格设定情境下复刻第三人效果理论,证实传播学经典理论在硅基样本中的重现的可能,为传播学采用硅基抽样开展理论研究与实践提供了重要支撑。其四,以往传播学的效应研究建立在人类认知与心理反应的经验模型之上,而 LLM 的复刻结果显示,传播效应的发生并非完全依赖人类主体,也可以通过算法化的语义加工过程得到模拟,其在智能传播时代下既是心理现象,也是算法现象,既存在于人,也存在于模型。由此,传播学的效应研究正从人类心理范式扩展为人机混合的认知范式。

2. 实践启示

本研究为传播学研究提供了多重价值。首先,为 LLM 注入大五人格设定可构建具有人格特质的“硅基被试”,无需招募真人即可开展实验,这一“硅基抽样”方法为资源、时间或经费受限的研究提供了新工具。传播学常聚焦老年人、儿童、少数群体等难获取的样本,或特殊地理、文化的语境,而可控、可复制、可规模化的硅基智能体样本源能有效突破样本稀缺与语境受限的困境。其次,在大规模问卷或多情境实验中,基于 LLM 的人格微调可快速搭建多版本、多人格、多任务实验框架,助力研究者短时间内获取初步反馈,既提升研究效率,也为传统实验前提供低成本的预实验、假设生成与行为预测途径。同时,这一方法支持实验条件重复、变量精确操控与跨模型结果比较,推动学科向可重复、可验证的科学化方向发展。最后,如何精准预测传播内容对受众的影响,一直是传播学研究中的一个核心难题。通过智能体模拟传播过程中的影响路径,从而构建一个可预测新闻传播影响的模型。LLM 可以根据不同的新闻框架、情绪表达以及叙事方式,模拟其对受众态度、信任度以及情感的即时反应。传播从业者可以通过为 LLM 注入符合目标受众的特定的人格设定,先进行传播效果的前置评估与预判,再根据评估结果优化内容创作方向、调整传播策略细节,最终实现内容与目标受众的深度共鸣,提升传播活动的精准度与实际转化效果。

(三)研究展望

本研究探索了不同 LLM 及大五人格设定在传播学实验中的应用,未来可以就以下方面深化研究。第一,研究复刻了单一传播实验任务,未来需向多维度媒介环境、议题语境与受众特征的实验设计扩展,构建跨情境框架以检验模型在不同传播结构下的表现。第二,应进一步优化提示词设计对机器语言的适配性,建立更为规范化的提示工程准则,提升输入一致性与输出稳定性,同时深挖大语言模型内部表征与外显行为的关联,判断其是否具备类人认知结构,通过人格扰动、变量校正等方式提升模拟灵活性,建立人机对照框架量化偏差来源,为实验可解释性与再现性提供更多依据。第三,由于原始实验未收集大五人格数据,仅参考 NEO-FFI 常模设定模拟分布,因此在反映个体异质性上可做进一步探索。未来研究计划设计人格与传播行为同步采集的实验,将真人样本的人格画像与 LLM 人格设定相对应,以便在后续的复刻中实现更高程度的心理一致性。

参考文献:

[1] BAIL C A. Can generative AI improve social science? Proceedings of the national academy of sciences of the united states of America,2024,121(21):e2314021121. [2025-07-21]. https://doi.org/10.1073/pnas.2314021121.

[2] WANG L,MA C,FENG X,et al. A survey on large language model based autonomous agents. (2025-03-02)[2025-07-21]. https://doi.org/10.48550/arXiv.2308.11432.

[3] ARGYLE L P ,BUSBY E C ,FULDA N ,et al. Out of one,many:using language models to simulate human samples. Political analysis,2023,31(3):337-351.

[4] SURI G ,SLATER L R ,ZIAEE A ,et al. Do large language models show decision heuristics similar to humans? A case study using gpt-3. 5. Journal of experimental psychology:general,2024,153(4):1066-1075.

[5] BINZ M,AKATA E,BETHGE M,et al. A foundation model to predict and capture human cognition. Nature,2025,644:1002-1009.

[6] YIGIT G,BAYRAKTAR R. Chatbot development strategies:a review of current studies and applications. Knowledge and information systems,2025,67(9):7319-7354.

[7] SHAPIRA E,MADMON O,REICHART R,et al. Can LLMs replace economic choice prediction labs? The case of language-based persuasion games. (2024-01-30)[2025-07-21]. https://doi. org/10. 48550/arXiv. 2401. 17435.

[8] YEYKELIS L,PICHAI K,CUMMINGS J J,et al. Using large language models to create AI personas for replication,generalization and prediction of media effects:an empirical test of 133 published experimental research findings. (2024-08-28)[2025-07-21]. https://doi. org/10. 48550/arXiv. 2408. 16073.

[9] LI P,CASTELO N,KATONA Z,SARVARY M. Frontiers:determining the validity of large language models for automated perceptual analysis. Marketing science,2024,43(2):254-266.

[10] BACHMANN F,VAN DER WEIJIDEN D,HEITZ L,et al. Adaptive political surveys and GPT-4:tackling the cold start problem with simulated user interactions. (2025-03-12)[2025-07-21]. https://doi. org/10. 48550/arXiv. 2503. 09311.

[11] CUI Z,LI N,ZHOU H. Can large language models replace human subjects? A large-scale replication of scenario-based experiments in psychology and management. (2025-06-20)[2025-07-03]. https://doi. org/10. 48550/arXiv. 2409. 00128.

[12] CHO S,KIM J,KIM J H. LLM-Based Doppelgänger models:leveraging ynthetic data for human-like responses in survey simulations. IEEE access, 2024, 12: 178917-178927. [2025-07-21]. https://doi. org/10. 1109/ACCESS. 2024. 3502219.

[13] GUHA N,NYARKO J,HO D E,et al. LegalBench:a collaboratively built benchmark for measuring legal reasoning in large language models. (2023-08-20)[2025-07-21]. https://doi. org/10. 48550/arXiv. 2308. 11462.

[14] JIANG H,ZHANG X,CAO X,et al. PersonaLLM:investigating the ability of large language models to express personality traits. (2024-04-02)[2025-07-23]. https://doi. org/10. 48550/arXiv. 2305. 02547.

[15] YUTA I,HIDEO J. Effect of LLM's personality traits on query generation. Proceedings of the 2024 Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region (SIGIR-AP 2024),Tokyo,2024:249-258. (2024-12-08)[2025-07-21]. https://dl. acm. org/doi/10. 1145/3673791. 3698433.

[16] HU T,COLLIER N. Quantifying the persona effect in LLM simulations. (2024-06-17)[2025-07-21]. https://doi. org/10. 48550/arXiv. 2402. 10811.

[17] PETROV N B,SERAPIO-GARCÍA G,RENTFROW J. Limited ability of LLMs to simulate human psychological behaviours:a psychometric analysis. (2024-05-12)[2025-07-21]. https://doi. org/10. 48550/arXiv. 2405. 07248.

[18] 曾秀芹,陈敏,刘旭阳. 社交媒体广告信息呈现策略对分享意愿的作用机制研究. 新闻大学,2024(6):102-117+123.

[19] COSTA P T,MCCRAE R R. Revised NEO personality inventory and NEO Five-Factor inventory. Odessa:Psychological Assessment Resources,1992.

[20] 姚若松,梁乐瑶. 大五人格量表简化版(NEO-FFI)在大学生人群的应用分析. 中国临床心理学杂志,2010,18(4):457-459.

[21] YANG Y,DUAN H,LIU J,et al. LLM-Measure:generating valid,consistent,and reproducible text-based measures for social science research. (2024-09-19)[2025-07-21]. https://doi. org/10. 48550/arXiv. 2409. 12722.

[22] Andreessen Horowitz. Top100 Gen AI Consumer Apps. (2025-03-06)[2025-07-03]. https://a16z. com/100-gen-ai-apps-4/.

[23] Open Science Collaboration. Estimating the reproducibility of psychological science. Science, 2015, 349 (6251): aac4716.

[24] ANDERSON S F,MAXWELL S E. There's more than one way to conduct a replication study:beyond statistical significance. Psychological methods,2016,21(1):1-12.

[25] VAN DER LINDEN D,TE NIJENHUIS J,BAKKER A B. The general factor of personality:a meta-analysis of big five intercorrelations and a criterion-related validity study. Journal of research in personality,2010,44(3):315-327.
[26] PEREZ E,RINGER S,LUKOSIUTE K,et al. Discovering language model behaviors with model-written evaluations. (2022-12-19)[2025-09-16]. https://doi.org/10.48550/arXiv.2212.09251.
[27] XIE Y,XIE Y. Variance reduction in output from generative AI. (2025-03-02)[2025-07-21]. https://doi.org/10.48550/arXiv.2503.01033.
[28] CHEN J,WANG X,XU R,et al. From persona to personalization:a survey on role-playing language agents. (2024-10-09)[2025-07-21]. https://doi.org/10.48550/arXiv.2404.18231.
[29] BECK T,SCHUFF H,LAUSCHER A,et al. Sensitivity,performance,robustness:Deconstructing the effect of sociodemographic prompting. (2024-02-08)[2025-07-21]. https://doi.org/10.48550/arXiv.2309.07034.
[30] CHENG M,DURMUS E,JURAFSKY D. Marked personas:using natural language prompts to measure stereotypes in language models. (2023-05-29)[2025-07-21]. https://doi.org/10.48550/arXiv.2305.18189.
[31] WANG Y,ZHAO J,ONES D Z,et al. Evaluating the ability of large language models to emulate personality. Scientific reports,2025,15(1):1-9. [2025-07-21]. https://doi.org/10.1038/s41598-024-84109-5.

Which Model Best Simulates Human Behavior? The Impact of Model Type and Personality Setting on the Accuracy of Large Language Models in Replicating Communication Studies Experiments

Zeng Xiuqin, *Chen Kelu*(Xiamen University)

Abstract:This study selected four mainstream large language models—ChatGPT-4o, DeepSeek-R1, Doubao-1.5, and Kimi-K1.5—as experimental subjects, employing a 2 (with vs. without Big Five personality settings) × 4 (model type) factorial design to construct virtual participants to replicate a communication experiment. Findings show clear performance differences:DeepSeek-R1 best approximated human averages and behavioral variability, while ChatGPT-4o displayed larger variance-fitting deviations but produced highly accurate and stable replications of main and indirect effects. Without personality prompts, models more closely matched human descriptive statistics; with personality prompts, causal-effect replication improved, and the only complete reproduction of two main effects occurred in this condition. Mediating effects remained difficult to replicate, though personality prompting helped reduce directional and magnitude deviations. Additional analyzes with ChatGPT-4o indicate that participant type(human vs. no-personality vs. partial-personality vs. full-personality groups) significantly moderates certain main effects and mediation pathways, with personality prompts suppressing extreme responses. This study provides cross-model comparison under personality-prompt conditions, demonstrating the value of Big Five prompting for enhancing simulation fidelity while underscoring LLMs' limitations in modeling complex psychological mechanisms.

Key words:large language model; Big Five personality traits; replication study; silicon-based subjects

■收稿日期:2025-07-21
■作者单位:曾秀芹,厦门大学新闻传播学院;福建厦门　361005
陈珂璐(通讯作者),厦门大学新闻传播学院
■责任编辑:肖劲草