基于大语言模型的数据生成与验证*

——以中国健康与养老追踪调查(CLHLS)为例

侯亚杰　庄亚儿**

提要:传统统计调查面临成本高昂、样本损耗及时效滞后等系统性难题,而大语言模型(LLMs)驱动的生成式人工智能(GAI)为革新数据获取范式提供了新途径。本文以中国健康与养老追踪调查(CLHLS)为实证场景,构建基于 LLMs 的老年健康数据生成框架,借助知识增强技术注入先验规则,对2021 年追踪样本的自评健康和日常生活活动能力(ADL)进行高保真模拟。研究发现,知识增强有效突破了通用大模型的局限,校正了模型偏差,较为准确地复现了健康指标与健康行为、人口学因素的关联模式。然而,技术落地仍面临三重挑战。基于此,本文提出在认识上构建"人机互馈"、方法上建立"人机共审"、生态上实现"人机共生"的"人类—AI"协同的社会研究新范式。

关键词:大语言模型　数据生成　验证　CLHLS

一、引　言

统计调查作为社会科学研究的重要基础,长期支撑着推断统计、因果分析等核心方法的实践,推动了全国性大型调查项目和社会科学量化研究的蓬勃发展。然而,传统社会调查模式正面临严峻挑战:随访成本高、样本损耗

*　本文系国家社会科学基金重点项目"人口发展综合调查通用平台建设与应用研究"(项目批准号:23ARK005)的阶段性研究成果。

** 侯亚杰,中国人口与发展研究中心。庄亚儿,中国人口学会、中国人口与发展研究中心。

大、敏感性议题应答率低、数据时效性差等问题日益突出,已成为制约研究深度与广度的普遍障碍。与此同时,生成式人工智能(GAI)技术,特别是大语言模型(large language models,LLMs)凭借其强大的深度语义理解和多模态数据生成能力,在扩展数据来源、优化调查流程、提升数据可信度等方面展现出独特潜力,正重塑统计调查的范式。

LLMs 通过学习海量历史数据中社会人口学特征与健康指标等复杂的关联模式,展现出强大的复现真实数据规律的能力。这使其既能作为研究者可控的高效"智能调查员",向各类对象获取信息,也能直接生成高质量数据供研究使用。这两种模式从根本上革新了定性与定量数据的获取逻辑,为有效模拟特定群体的状态演变、弥补数据断层、降低调查成本并提高时效性开辟了新途径。

在此背景下,本文依托中国健康与养老追踪调查(CLHLS)2014—2018 年纵向数据集,将基于 LLMs 的数据生成方法引入老年健康研究领域,旨在构建基于 LLMs 的数据生成框架,对 CLHLS 追踪样本 2021 年的关键健康指标[如自评健康、日常生活活动能力(ADL)]进行高保真度生成。通过将生成数据与 2021 年 CLHLS 实际调查数据进行严格的分布一致性检验、变量关联分析及预测效度验证,本文将科学评估 LLMs 生成数据在老年健康研究中的可靠性与应用价值,为推动统计调查向智能化、高效化、可持续化转型提供方法和实践参考。

二、文献综述

近年来,GAI 在多个领域展现出强大的数据生成与处理能力。在社会科学研究中,GAI 逐渐成为推动数据收集、调查设计、数据填补以及结果分析等环节创新的重要工具。通过生成数据,GAI 不仅能弥补调查数据的缺失,还能模拟传统调查对象,填补因样本难以触及或数据稀缺造成的空缺。在此情境下,GAI 为研究人员提供了前所未有的解决方案,显著提高了数据收集的效率和覆盖面,为传统调查方法带来了新视角与新思路(Aher et al., 2023; Bail, 2024)。

(一)GAI 在数据获取中的应用

基于大语言模型的 GAI 取得突破性发展,正深刻变革着社会科学研究的数据获取范式。早期探索主要集中在 GAI 在数据缺失填补和虚拟受访者模拟等基础领域。霍顿(Horton,2023)通过测试 GPT-3 模型在四项理性选择实验中的表现,发现其模拟的"硅基人"(silicon subjects)行为模式与经典"经济人"假设高度相似,为 AI 模拟人类决策行为提供了初步证据。进一步研究表明,GAI 能有效填补调查数据中的缺失项,特别是在受访者未提供完整回答时,还能基于已有信息推测可能的答案,显著提升数据的完整性(Bail,2024)。这些初步成果为 GAI 在社会调查中的应用奠定了理论基础,其潜力被普遍认为在于助力触及传统调查难以覆盖的群体、进行大规模调查前的预研、降低实验成本以及填补关键数据缺失(Kozlowski et al., 2024)。

随着 ChatGPT、DeepSeek 等先进 LLMs 的迭代发展,GAI 在社会调查数据获取中的应用迅速拓展,展现出替代或补充传统数据收集手段的显著能力。相关研究不仅证实了其在提升调查效率、降低调查成本方面的优势,还深入探索了其对复杂社会现象的模拟潜力(边燕杰、缪晓雷,2025;陈忱,2025)。阿吉尔等人(Argyle et al., 2023)利用 GPT 模型模拟美国社会群体的政治观点与投票行为,发现生成数据与实际观测数据高度一致。这一发现有力地支持了 LLMs 能够精准模拟特定群体行为模式的观点,尤其在传统概率抽样难以覆盖或成本过高的情况下,为研究提供了有效的替代性数据来源。同样,阿赫尔等人(Aher et al., 2023)通过 GAI 模拟群体行为和社会互动模式,探索不同社会结构中的行为规律,为社会学研究开辟了新路径,拓宽了社会调查的视角。

除了直接生成虚拟数据替代或补充传统数据获取方式,LLMs 在优化传统社会调查设计环节的作用也日益突出。研究者利用 LLMs 高效生成、筛选和优化调查问题,并能根据预设的受访者特征动态调整问题的表述方式和内容,从而提升问卷设计的精准度和适应性。例如,迪利恩等人(Dillion et al., 2023)利用 LLMs 生成伦理道德场景问题,成功减少了传统问卷设计中可能存在的偏见,提高了数据获取的质量与可靠性。这种 AI 辅助的设计过程,使研究者能够在更短时间内制定出更具代表性和科学性的调查

工具。

过去两年间,LLMs(尤其是 GPT 系列模型)在社会科学研究中的应用呈爆发式增长,其核心争议点在于它模仿人类受试者的能力边界。特别值得关注的是,阿吉尔等人(Argyle et al., 2023)提出了“算法保真度”(algorithmic fidelity)和“硅样本”(silicon samples)的概念。他们认为,经过适当调试的 LLMs 能够以与人类细分群体反应模式高度一致的方式生成输出,既体现特定群体的典型倾向,也包含合理的偏离,进而实现对特定社会群体行为的模仿。这为理解和评估 LLMs 生成数据在社会科学研究中的效度提供了关键理论框架,标志着该领域研究取得重要进展。

(二)GAI 在数据获取中的争议与挑战

尽管 GAI 在社会调查中应用潜力巨大,但在实际应用中仍面临一系列不可忽视的挑战。首先,生成数据存在系统性偏差与失真风险。研究普遍表明,LLMs 生成的“硅样本”往往带有其训练数据固有的社会偏见和结构性不平等。当训练语料库缺乏多样性或对某些群体覆盖不足时,模型生成的数据可能无法准确反映这些群体的独特行为和观点,导致输出失真(周穆之等,2025;Bender et al., 2021;Mehrabi et al., 2019;Binns & Kirkham,2018)。这种偏差在模拟复杂社会现象或特定群体行为时更为明显。例如,多项政治倾向预测研究发现,未经严格校准的模型输出常对特定国家、社会群体或意识形态表现出系统性偏向,与真实人群的反应差异显著(Bisbee et al., 2024;Santurkar et al., 2023;Heyde et al., 2023;Motoki et al., 2024;Hartmann et al., 2023;Rozado,2024;Rutinowski et al., 2024)。阿吉尔等人(Argyle et al., 2023)的研究进一步量化了数据覆盖不足的危害,指出当训练数据对目标人群覆盖率低于 85% 时,模型输出偏差会呈指数级增长,导致对未充分覆盖群体产生系统性误判。这种偏差在迭代应用过程中可能形成“认知回声室效应”(cognitive echo chamber effect),强化既有模式而非真实呈现社会结构的复杂性。更令人担忧的是,主流文化叙事在训练数据中占比过高会显著压缩少数群体文化特征的语义表征,跨语言场景下的隐喻捕捉能力也会大幅降低,形成一种存在内在缺陷的“数字社会晶体”(digital social crystal),通过算法递归

强化认知扭曲(Bail,2024)。对于生成数据而言,这种偏差问题更为复杂。生成模型不仅会复制原始数据中的偏见,还可能因算法设计倾向于强化主流分布特征,导致少数群体样本在生成数据中进一步边缘化或失真,形成“双重选择性失真”(Stadler et al., 2022;Santurkar et al., 2023)。

伦理与隐私困境是另一重大挑战。社会调查常涉及健康、行为等高度敏感的个人信息。利用 LLMs 生成此类数据,或在训练过程中处理敏感数据,若保护不当,极易引发严重的隐私泄露风险,触碰伦理红线并可能违反相关法律法规。因此,如何在利用 GAI 提升数据获取效率与质量的同时,确保严格的隐私保护与伦理合规性,成为平衡技术创新与社会责任的关键难题(Lipton,2018)。生成数据虽被宣传为隐私保护的理想解决方案,但其实际效果存在显著争议。普通生成数据生成模型可能因记忆效应泄露原始数据信息,尤其是异常值(如罕见病例、高收入群体)因统计独特性更容易被识别,导致成员推理攻击风险(Jordon et al., 2022;Stadler et al., 2022;Mehrabi et al., 2019)。

最后,LLMs 固有的“黑箱”特性(black-box nature)及其导致的可解释性缺失,可能会严重削弱研究者对生成数据的信任和应用信心。模型内部决策过程高度复杂且不透明,这使得研究人员难以理解和验证生成数据的具体逻辑与依据(Lipton,2018)。在社会调查这类对数据来源和可靠性要求极高的领域,这种可解释性的匮乏极大地限制了生成数据的广泛应用。开发“可解释人工智能”(explainable AI,XAI)技术以提高模型透明度,被视为应对这一挑战的重要方向(Gilpin et al., 2022;郭茂灿等,2025)。

面对上述挑战,学界提出了相应的反思与解决办法。价值对齐(value alignment)被视为缓解偏见的关键,它强调在模型训练和生成前应扩展样本,覆盖多元群体,确保“更多人被代表”,而非仅反映主流叙事(Sakib & Das, 2024)。同时,知识增强(knowledge augmentation)通过向模型注入特定领域的专业知识,旨在减少事实性错误与推理偏差,并在一定程度上破解“黑箱”困境。强化去偏见算法研究、提升训练数据多样性、进行更精细化的社会调查设计,也被认为是提升生成数据公平性与科学性的必要手段(Dastin, 2023)。

三、数据生成

为测试大语言模型生成数据的能力,本文以中国健康与养老追踪调查(CLHLS)为例,聚焦评估模型对老年人群体客观健康状态(日常生活活动能力评分,ADL)及自我健康评价的理解能力。一般流程是:利用社会人口统计学信息,引导大语言模型代入具有特定特征的个体视角进行“思考”,再向其提出调查问题,所得到的回答结果即为生成数据。同时,为明确知识增强与模型本身的作用边界,本文通过多组对比实验(无知识增强、仅知识增强、知识增强 + 大语言模型)分解两者的贡献,并重点对增强知识的有效性进行独立验证。

(一)数据来源与说明

本文选择中国健康与养老追踪调查(CLHLS)数据作为数据生成实例,主要是基于其独特的数据特征与研究适配性。

从时间维度看,CLHLS 自 1998 年至 2021 年持续了 20 多年,积累了 13 余万人次的样本,构建了罕见的老年群体纵向健康数据库。这种长期追踪特性,为评估模型捕捉老年人核心健康指标的时间关联、生成符合真实演变规律的数据提供了关键基准。从数据维度看,CLHLS 包含 200 余项调查问题,涵盖社会人口学特征、健康行为、医疗资源利用等多维度信息,构建了高维度的老年健康画像。这种复杂性为检验大语言模型的多变量关联建模能力提供了天然场景。从研究对象看,CLHLS 聚焦高龄老人群体,该群体社会经济活动相对稳定,主要健康指标受外部波动影响较小,日常生活活动能力、自评健康等核心指标的变化呈现强规律性(如随年龄增长失能风险递增)。这种稳定性为判断生成数据的合理性提供了明确参照,便于检验模型是否准确捕捉老年健康的内在规律。此外,高龄老人数据相对稀缺,生成高质量数据的需求迫切,CLHLS 的代表性使其成为验证生成数据可用性的典型案例。

为简化研究流程,本文暂不考虑纵向数据的删失问题,仅采用两期追踪

数据:2014—2018 年数据用于模型训练,以此捕捉老年群体健康指标的动态变化;在 2018—2021 年数据中,2018 年数据为生成 2021 年数据提供背景信息,2021 年数据则用于与生成数据进行对比验证。

(二)模型选择与知识增强

本文选用国产大模型 DeepSeek-v3 进行数据生成,主要基于两方面优势。一方面是语境适配性。该模型在中文语义理解和中国老年健康特征建模方面更具优势,能够精准捕捉 CLHLS 数据中“日常生活自理能力”“自评健康等级”等本土表述,减少跨文化语义偏差。另一方面是垂直领域适配潜力,DeepSeek-v3 支持灵活的知识注入机制,便于针对老年健康细分领域进行定制化增强。本文的知识增强体系包含三部分,具体构建如下。

1. 结构化领域知识图谱

基于 2014—2018 年 CLHLS 追踪数据,运用时序分析方法提取核心健康指标的动态规律。核心变量包括“ADL 评分”“自评健康”“慢性病数量”“锻炼频率”等 12 类,系统地捕捉 2014—2018 年 ADL 衰退、慢性病数量增加、自评健康下降等变量的动态变化趋势,将此作为基础知识图谱,为预测 2018 年老年人群在 2021 年的健康状况提供支撑。

2. 健康指标预测

为提高大语言模型生成数据的有效性,本文构建了可解释的回归模型,并将其作为核心增强知识之一。

该模型的因变量为 2018 年 ADL 评分和自评健康评分。为简化模型构建过程,本文将这两个因变量均处理为连续变量,采用多元线性回归方法构建预测模型。模型自变量包括 2018 年人口统计学特征(包含年龄、城乡分类等四项指标)和健康状态及行为指标(包含 2018 年 ADL 评分、慢性病数量等五项指标)。具体回归结果如表 1 所示。

基于上述回归结果,本文将回归系数转化为可直接调用的规则化表述,例如:“年龄每增长 1 岁,ADL 评分平均增加 0.035 分。”“饮食多样性评分每增加 1 分,自评健康得分平均降低 0.015 分。”

表 1　　2018 年老年人健康指标预测模型

变量	2018 年 ADL 评分	2018 年自评健康评分
2014 年 ADL 评分	0.413***	
女性	0.131	−0.100
年龄	0.035***	−0.002
镇	−0.046	−0.028
乡	−0.062	0.076
饮食多样性评分	−0.006	−0.015***
抽烟	−0.116	−0.118
喝酒	−0.089	−0.063
体育锻炼	−0.134*	−0.035
家庭经济条件评分	−0.032	0.278***
在婚	0.018 (0.075)	0.046 (0.066)
曾患病	0.251** (0.106)	0.446*** (0.094)
卧床不起	2.939*** (0.499)	1.679*** (0.434)
体重(kg)	0.0005 (0.003)	−0.0003 (0.003)
身高(cm)	0.006 (0.005)	−0.010** (0.004)
2014 年自评健康评分		0.224*** (0.033)
常数项	−3.414*** (0.988)	3.176*** (0.872)
观测	873	873

3. 知识增强与大语言模型的协同机制

本文中,知识增强与 DeepSeek-v3 大语言模型的协同通过“规则嵌入提示词 + 样本微调”两步来实现。首先,依托前文构建的结构化领域知识图谱提炼出的老年健康指标动态规律(如 2014—2018 年 ADL 衰退趋势),以及可解释回归模型转化的规则化表述(如年龄与 ADL 评分的关联规则),共同约束模型输出边界(如将 ADL 评分波动控制在真实值 ± 1 分以内),确保生成结果符合老年健康演变的客观规律;其次,选取 2014—2018 年真实数据进行

模型微调,进一步增强知识与模型的适配性。最终,模型在生成数据时,既能依托大语言模型的优势保障输出的逻辑合理性,又能通过知识增强确保结果与老年健康指标规律一致,实现通用文本生成能力与领域专业知识的深度融合。

(三)提示词与数据生成

本文同时选取了多个与老年人 ADL 评分和自评健康相关的协变量,具体涵盖社会人口学特征、生理特征、生活习惯、健康状况历史信息。剔除含有缺失值的样本后,最终获得 2014—2018 年有效样本 873 个,2018—2021 年有效样本 3203 个。具体提示词如下:

个体特征:

这是一位[2018 年年龄]岁的中国[性别][民族]老人,居住在[省份][城乡]地区。婚姻状态为[婚姻状态],家庭经济地位在本地属于[经济地位];生理特征:身高[身高(cm)],体重[体重(kg)];生活习惯:目前[吸烟状况:吸烟/不吸烟]、[饮酒状况:饮酒/不饮酒],体育锻炼频率为[锻炼频率:从未/偶尔/定期],饮食多样性评分为[饮食多样性评分(0—30 分)];健康状况:过去两年内[患病情况:未患重病/患过重病/卧床不起]。2018 年调查时日常生活活动能力(ADL)评分为[ADL 评分(0—6 分)],健康自评结果为[健康自评:非常好/较好/一般/较差/很差]。

老年健康变化规律:

年龄每增长 1 岁,ADL 评分平均增加 0.035 分;

饮食多样性评分每增加 1 分,自评健康得分平均降低 0.015 分;

过去两年曾患重病,自评健康评分等级平均下降 1 级。

……

请结合上述个体特征和规律,回答:

2021 年这位老人的自评健康为_____;

2021 年这位老人的 ADL 评分为_____;

2021 年这位老人的婚姻状态为_____;

2021 年这位老人的社会经济评分为_____;

……

最终生成的 2021 年老年人主要健康指标和变量情况如表 2 所示。

表 2　　生成样本与人类样本主要变量基本特征

变量		人类样本			DeepSeek-v3（知识增强）		DeepSeek-v3（未知识增强）	
		N	Mean	SD	Mean	SD	Mean	SD
ADL 评分		3203	0. 4249	1. 1856	0. 4302	1. 1621	0. 6497	1. 3500
自评健康评分		3203	2. 4880	0. 9453	2. 4833	0. 9406	2. 4917	0. 9402
居住地	城市	3203	0. 2485	0. 4322	0. 2548	0. 4357	0. 2401	0. 4254
	镇	3203	0. 4471	0. 4973	0. 3598	0. 4799	0. 3757	0. 4831
	乡	3203	0. 3044	0. 4602	0. 3865	0. 4873	0. 4106	0. 4919
饮食多样性评分		3203	19. 0166	7. 5061	18. 6295	6. 8915	18. 2110	6. 8123
抽烟		3203	0. 1673	0. 3733	0. 1537	0. 3606	0. 1789	0. 3835
喝酒		3203	0. 1520	0. 3591	0. 1689	0. 3743	0. 1699	0. 3757
体育锻炼		3203	0. 4639	0. 4988	0. 4783	0. 4996	0. 4023	0. 4912
家庭经济条件评分		3203	2. 9532	0. 5702	2. 7835	0. 6318	2. 8901	0. 6288
在婚		3203	0. 5083	0. 5000	0. 5013	0. 5000	0. 5588	0. 4965
体重(kg)		3203	58. 8831	17. 8457	57. 3269	12. 8936	56. 1235	12. 4568
身高(cm)		3203	157. 4160	9. 8953	155. 9873	9. 9216	156. 6890	9. 7457
两年内患病情况	未患病	3203	0. 8935	0. 3085	0. 8794	0. 3258	0. 9123	0. 2835
	曾患病	3203	0. 0965	0. 2953	0. 1186	0. 3236	0. 1110	0. 3123
	卧床不起	3203	0. 0100	0. 0995	0. 0062	0. 0786	0. 0021	0. 0457

四、生成数据的验证

鉴于 LLM 的核心功能为预测，本文将模型的预测结果视为该特征人群对调查问题的回答，并将其与真实调查答案（人类受访者对同一问题的回答）进行对比分析。

(一)验证方法

本文与其他研究不同,并非直接应用通用大模型生成数据,而是在生成数据前增设知识增强环节。为检验知识增强后 LLM 生成数据的准确性,研究通过统计检验对比人类样本与两类生成样本,以此判断知识增强的有效性。在验证维度上,本文借鉴前人的分析策略(凌宛莹、崔思瞻,2025):通过统计检验评估生成样本与人类样本在主要变量上均值差异的统计显著性;运用列联表矩阵和 Gamma 相关系数衡量模拟生成的个体与真实个体在 ADL 评分和自评健康的重合度,揭示具体应答的相似程度;通过回归分析,对比人类样本与生成样本中各协变量与老年健康指标的回归系数及置信区间,反映不同样本间变量关系的异同及系统性偏差。通过上述分析,综合评判 LLM 生成的老年健康样本表现及其与人类样本的差异。

(二)生成样本与人类样本的一致性

表 3 报告了未进行知识增强和进行知识增强后的两类生成样本在主要变量上与人类真实样本的差异。从核心健康指标来看,知识增强样本对人类真实样本的模拟精度显著高于未知识增强样本。ADL 评分作为衡量日常活动能力的核心指标,知识增强样本与人类样本的差异极小,无统计学显著性,这表明其在日常活动能力相关特征的生成上与真实情况高度契合;而未知识增强样本与人类样本的 ADL 评分差异显著,差异值达 -0.2248,显示其对基础生活能力的模拟与真实水平偏差较大。在自评健康评分方面,两类生成样本与人类样本的差异均不显著,整体都接近真实样本的分布特征。

在其他变量方面,知识增强样本的整体表现优于未知识增强样本,与人类样本的偏离趋势更为缓和。在居住地分布方面,两类生成样本在“镇”和“乡”的占比上均与人类样本存在显著差异。在饮食多样性评分方面,知识增强样本与人类样本的差异程度及显著性均低于未知识增强样本,说明其对饮食相关特征的模拟更为准确。在生活习惯方面,在抽烟、喝酒、参加体育锻炼这三个变量上,未知识增强样本与人类样本存在显著差异,而知识增强样本则无明显偏离,体现了知识增强在行为特征生成上的优势。在家庭经济条件评分方面,两类样本都与人类样本存在显著差异。在人口学特征方面,在“在婚”状

态的模拟上,知识增强样本对婚姻状态相关特征的捕捉更为精准。在身体指标方面,在体重和身高的模拟中,知识增强样本与人类样本的差异均小于未知识增强样本,更贴近真实身体特征分布。在患病情况方面,未知识增强样本在“未患病”和“卧床不起”等健康状态上与人类样本差异显著,知识增强样本的偏离则主要体现在“曾患病”指标上,整体对健康状态相关特征的模拟更为可靠。

总体而言,知识增强样本在核心健康指标和多数协变量上都更接近人类真实样本,整体上证实了知识增强对提升生成数据真实性的作用。

表 3　　生成样本与人类样本在主要变量上的差异

变量		人类样本 VS 知识增强样本			人类样本 VS 未知识增强样本		
		差异	t 值	显著性	差异	t 值	显著性
ADL 评分		-0.0051	-0.1741		-0.2248	-7.0692	***
自评健康评分		0.0047	0.2000		-0.0037	-0.1570	
居住地	城市	-0.0063	-0.5800		0.0084	0.7800	
	镇	0.0873	7.1500	***	0.0714	5.8200	***
	乡	-0.0821	-6.9600	***	-0.1062	-8.9600	***
饮食多样性评分		0.3871	2.1500	*	0.8056	4.4600	***
抽烟		0.0136	1.4800		-0.0116	-1.2300	
喝酒		-0.0169	-1.8400		-0.0179	-1.9500	
体育锻炼		-0.0144	-1.1600		0.0616	4.9700	***
家庭经济条件评分		0.1697	11.3100	***	0.0631	4.2300	***
在婚		0.0070	0.5600		-0.0505	-4.0700	***
体重(kg)		1.5562	4.0000	***	2.7596	7.2000	***
身高(cm)		1.4287	5.7700	***	0.7270	2.9700	**
两年内患病情况	未患病	0.0141	1.7800		-0.0188	-2.5400	*
	曾患病	-0.0221	-2.8700	**	-0.0145	-1.9100	
	卧床不起	0.0038	1.6900		0.0079	4.0900	***

注:$^{*}p<0.05$,$^{**}p<0.01$,$^{***}p<0.001$。

(三)生成样本的准确性

除了验证生成样本与人类样本在主要指标上的一致性之外,还需要检验个体层面的数据生成能否准确模拟人类样本。图 1、图 2 分别展示了知识增强前后的生成样本在 ADL 评分和自评健康评分上的预测准确度。

从整体关联来看,知识增强样本、未知识增强样本与人类样本在较为客观的 ADL 评分上的 Gamma 系数分别为 0.97 和 0.90,在自评健康得分上的 Gamma 系数分别为 0.95 和 0.93。这表明两类生成样本对老年人主客观健康指标的预测,都符合真实样本的整体分布逻辑。

进一步分析生成样本对老年人 ADL 得分的预测表现(图 1),两类生成样本的准确预测概率变化趋势基本一致,但差异显著。整体而言,未知识增强样本的 ADL 评分预测准确率为 82.1%,而知识增强样本则达到了 95.3%。具体而言,知识增强样本的预测准确率整体维持在较高水平,虽有波动但均保持在 60%以上;尤其在对无失能(ADL=0)和重度失能(ADL=6)情况的预测上,准确率均超过 90%。未知识增强样本准确率变动趋势与知识增强样本一致,但除了在无失能场景下准确率较高外,其他失能阶段的预测准确性明显偏低,轻中度失能(ADL=2—3)的准确率仅为 20%—30%。两类样本的准确率差异在各 ADL 得分段均有体现,并且随着失能程度的加深而扩大——在中重度失能(ADL=4—6)阶段,知识增强样本准确率保持在 80%以上,未知识增强样本则不足 50%,差异尤为明显。在较为主观的自评健康得分方面(图 2),两类生成样本与人类样本的分布趋势较为一致,各评分占比相差不大,但准确率存在明显差异:未知识增强样本的预测准确率为 82.2%,知识增强样本为 91.4%。特别是在对“很差”(5 分)自评健康的预测上,知识增强样本准确率达到 57%,未知识增强样本仅为 38%。

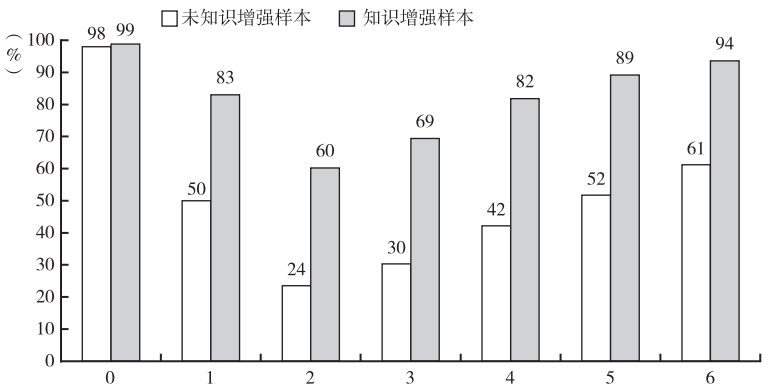

图 1 生成样本在 ADL 评分上的准确性

注:$Gamma_{人类样本vs未知识增强样本}=0.9005$;$Gamma_{人类样本vs知识增强样本}=0.9653$。

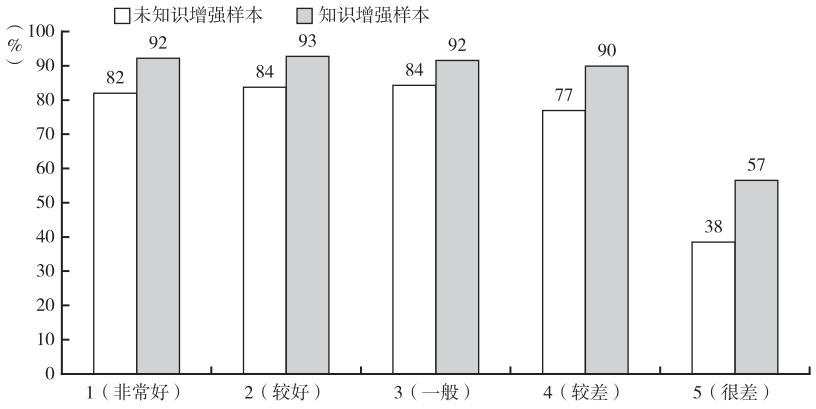

图 2　生成样本在自评健康评分上的准确性

注：$Gamma_{人类样本vs未知识增强样本}=0.9330$；$Gamma_{人类样本vs知识增强样本}=0.9534$。

（四）生成数据在变量关系上的稳定性

图 3 展示了通过回归分析对比人类真实样本、知识增强生成样本和未知识增强生成样本在 ADL 和自评健康方面的变量关联模式。在 ADL 评分的影响因素分析中（图 3a），知识增强生成样本呈现出与人类真实样本高度一致的关联特征，大部分变量的效应值完全覆盖了真实样本的置信区间。未知识增强样本虽然在方向和强度上未呈现显著差异，但在刻画患病状况、健康行为习惯（抽烟、锻炼）以及居住地（城乡类别）与 ADL 评分的关系时，与真实样本存在较大偏差。此外，值得注意的是，无论是知识增强生成样本还是未知识增强样本，模型中各变量的置信区间相较于真实样本都相对较小。在自评健康模型中（图 3b），上述差异仍然存在，且未呈现较大波动。

五、总结与讨论

（一）主要研究发现与启示

本文基于中国老年健康影响因素跟踪调查（CLHLS）开展实证分析，系统探讨了 LLM 生成老年健康数据的可行性与应用边界。研究发现，经知识增强的 GAI 技术能够有效模拟老年健康数据样本，在主要变量分布一致性、个体

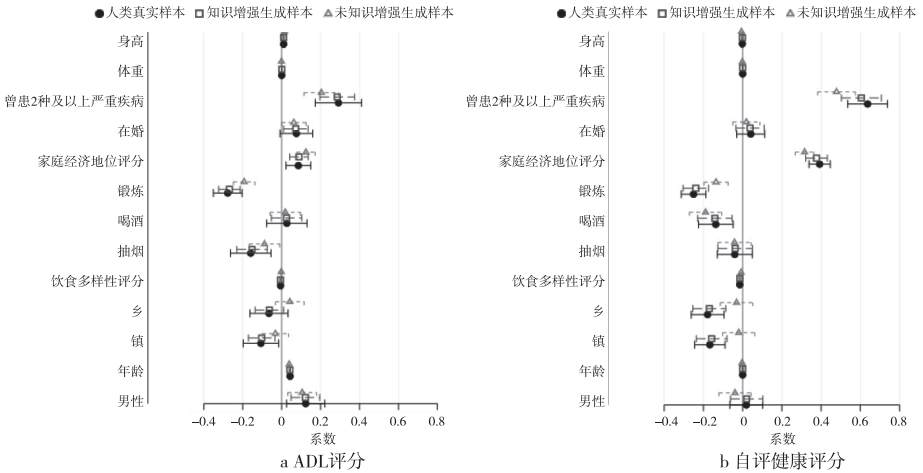

图 3　生成样本与人类真实样本在变量关系上的差异

数据微观准确性及变量关系稳定性等方面,均展现出与真实人类样本的统计学相似性。特别值得关注的是,知识增强技术的引入显著提升了生成数据的质量。这一提升在日常生活能力(ADL)等客观健康指标的模拟中尤为明显,它不仅有效修正了基础模型对失能程度预测的系统性偏差,还准确复现了健康指标与健康行为、人口学特征之间的关联模式。这一发现具有重要的方法论启示:通用 LLM(如 ChatGPT、DeepSeek 等)在处理老年健康等专业领域任务时,必须通过结构化嵌入医学、社会学等先验知识,才能突破模型固有的"认知浅层化"局限,使生成结果更接近真实情况。

同时,必须强调的是,本文基于大语言模型的数据生成技术本质上仍处于"模式复现"层面,尚未实现真正的"机制重构"。即便经过知识增强优化,该技术在落地应用中仍面临三重核心挑战,这些挑战直接关系到生成数据的科学效度与应用价值。

首先,模型固有的"幻觉"风险与知识更新滞后性带来挑战。在模拟健康状态关联时,模型可能虚构不存在的影响机制,或因训练数据未涵盖最新状况而导致预测偏差。虽然检索增强生成(RAG)技术可动态注入新信息,但模型缺乏深层因果推理能力,仍可能产生表面合理却违背科学逻辑的输出。因此,我们需及时更新时效性参数,锁定老年健康领域的核心知识权重,避免关

键科学规律被噪声数据稀释。

其次,健康数据本身的多维度复杂性是第二重挑战。老年健康状态是生理机能、心理适应与社会环境交互作用的结果,具有显著的系统性特征。以ADL评分为例,它不仅反映个体的躯体功能状况,还受家庭照护资源、社区养老设施等环境因素的深刻影响;而自评健康则隐含代际文化价值观的差异。现有模型在整合社会人口特征、健康行为习惯与医疗史等多源信息时,往往难以捕捉这些跨维度的动态关联,导致特征割裂与归因失真。更需警惕的是,知识增强过程可能意外放大训练数据中的隐性偏见。若训练数据过度强调某些因素对健康的影响,模型可能机械强化这些因素的贡献度,而忽视其他保护因素的作用。解决这一问题需要构建多源多模态数据融合架构,通过定义关键变量间的交互规则,引导模型识别健康决定因素的非线性耦合机制,而非简单拟合表面相关性。

最后,生成数据的代表性与偏见控制问题构成第三重挑战。如果模型简单复制CLHLS样本的特征分布(如高龄老人占比过高),不仅会弱化其他亚群体的健康特殊性,更可能导致算法对隐性偏见的递归强化。例如,训练数据对“80岁以上群体失能风险”的过度强调,可能扭曲年龄与ADL受限之间的真实关联模式。针对这一问题,需要建立超越传统统计指标的多层级评估体系,引入“群体覆盖率”“偏见敏感指数”等创新维度。在技术实现上,可采用动态样本加权方法,对代表性不足的群体赋予更高生成权重,在保持数据整体逻辑一致性的前提下,增强生成结果的群体异质性表征能力。

(二)未来方向:“人类—AI”协同的社会研究新范式

本文蕴含的深层次启示是,GAI在社会科学研究中,并非简单的工具化应用,而是必将引发人机互构的研究范式革命。这种革命性转型将通过认识论、研究方法和研究生态的多重创新,重塑社会认知的生产逻辑。

在认识论层面,我们必须摒弃将GAI视为被动工具的固有观念,构建“人机互馈”的知识生产框架。研究者通过注入领域知识来塑造模型的语义理解结构,而模型则凭借跨模态融合能力,揭示微观个体行为与宏观社会机制之间的隐性关联,反哺人类的理论盲区。这种人类锚定概念内核、AI验证现实

表征的双向互动,将推动知识生产从经验假设转向证据驱动,最终实现从个体生命历程到社会结构的多层次机理辩证统一。

在研究方法方面,重构体现为“人机共审”的辩证评估体系。传统的机械验证仅依赖统计显著性等单一标准,而新范式则是覆盖数据生成全周期的综合评估体系。在生成端,嵌入人口结构、经济社会发展等动态监测指标和偏见审查规则,通过对抗训练约束算法输出;在验证端,引入实践主体的质性评价,结合强化学习持续优化。这种框架将技术理性(模型置信区间)与社会理性(田野经验)相融合,形成量化—质性双轨校验机制,使生成数据既符合逻辑严谨性,又承载社会实践意义,实现从技术中立到价值负载的转变。

研究生态的变革在于打破人类样本与生成数据的二元对立,构建“人机共生”的协同体系。一方面,以严格质控的人类样本(如 CLHLS 队列)作为“事实锚点”,保障研究的实证基础;另一方面,充分发挥生成数据的认知拓展潜能——在广度上,生成罕见群体样本,破解实证研究中的长尾困境;在深度上,模拟追踪中断期的社会事实演变,重建纵向研究的连续性逻辑;在动态性上,构建不同情境的虚拟对照实验,实现社会过程的可控推演。这种协同增效机制标志着研究范式的根本转型,二者在辩证统一中加深了对社会复杂性的理解。

当大语言模型从数据合成工具升级为“社会机制模拟器”时,其释放的认知潜能与潜藏的系统性风险构成辩证统一。这就要求我们构建“人类—AI”协同的社会研究新范式。人类研究者依靠价值理性与理论洞见把控认知方向,人工智能凭借超大规模计算与模式识别拓展认知范围。只有这样,才能避免陷入技术乌托邦与怀疑论的两极误区,使智能革命真正提升对社会复杂性的解构能力,为应对老龄化等重大挑战提供兼具科学精确性与人文关怀的决策支持体系。

参考文献:

边燕杰、缪晓雷,2025,《大数据视野下的社会科学实证研究》,《智能社会研究》第 1 期。

陈忱,2025,《生成式人工智能赋能随机实验——机遇与挑战》,《智能社会研究》第 2 期。

郭茂灿等,2025,《智能社会学:生成式人工智能驱动下的社会变迁与社会学范式重构》,《智能社会研究》第 2 期。

凌宛莹、崔思瞻,2025,《复现的限度——“硅基样本”的可能与可为》,《智能社会研究》第 2 期。

周穆之等,2025,《大语言模型智能体能模仿问卷调查受访者吗?——来自中国人口数据的基准比较》,《智能社会研究》第 2 期。

Aher, G. et al. 2023, “Using Large Language Models to Simulate Multiple Humans and Replicate Human Subject Studies.” arXiv Preprint. doi: 10.48550/arXiv.2208.10264.

Argyle, L. et al. 2023, “Out of One, Many: Using Language Models to Simulate Human Samples.” *Political Analysis* 31(3).

Bail, C. 2024, “Can Generative AI Improve Social Science?” *PNAS* 121(21).

Bender, E. et al. 2021, “On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?” Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency.

Binns, R. & R. Kirkham 2021, “How Could Equality and Data Protection Law Shape AI Fairness for People with Disabilities?” Preprint.

Bisbee, J. et al. 2024, “Synthetic Replacements for Human Survey Data? The Perils of Large Language Models.” *Political Analysis* 32(4).

Dillon, D. et al. 2023, “Can AI Language Models Replace Human Participants?” *Trends in Cognitive Sciences* 27(7).

Gilpin, L. et al. 2022, “‘Explanation’ is Not a Technical Term: The Problem of Ambiguity in XAI.” Proceedings of the 5th IEEE International Conference on Data Science and Advanced Analytics.

Hartmann, J. et al. 2023, “The Political Ideology of Conversational AI: Converging Evidence on ChatGPT's Pro-Environmental, Left-libertarian Orientation.” Preprint.

Heyde, L. et al. 2023, “Vox Populi, Vox AI? Using Language Models to Estimate German Public Opinion.” SocArXiv Preprint. doi: 10.31235/osf.io/8je9g.

Horton, J. 2023, “Large Language Models as Simulated Economic Agents: What Can We Learn from Homo Silicus?” NBER Working Paper 31122.

Jordon, J. et al. 2022, “Synthetic Data: What, Why and How?” arXiv Preprint. doi: 10.48550/arXiv.2205.03257.

Kozlowski, A. et al. 2024, “In Silico Sociology: Forecasting Covid-19 Polarization with Large Language Models.” arXiv Preprint. doi: 10.48550/arXiv.2407.11190.

Mehrabi, N. et al. 2019, “A Survey on Bias and Fairness in Machine Learning.” arXiv Preprint. doi: 10.48550/arXiv.1912.04889.

Motoki, F. et al. 2024, “More Human Than Human: Measuring ChatGPT Political Bias.” *Public Choice* 198(1).

Rozado, D. 2024, “The Political Preferences of LLMs.” *PLOS ONE* 19(7).

Rutinowski, J. et al. 2024, “The Self-Perception and Political Biases of ChatGPT.” *Human Behavior and*

Emerging Technologies 2024(7115633).

Sakib, S. & A. Das 2024, "Challenging Fairness: A Comprehensive Exploration of Bias in LLM-Based Recommendations." 2024 IEEE International Conference on Big Data.

Santurkar, S. et al. 2023, "Whose Opinions Do Language Models Reflect?" *Proceedings of Machine Learning Research* 202.

Stadler, T. et al. 2022, "Synthetic Data-Anonymisation Groundhog Day." 31st USENIX Security Symposium.

责任编辑:王誉梓、赵海峰