整体事实还是有偏样本

——基于大语言模型生成数据的测量

龚为纲　黄思源

摘　要　大语言模型（LLM）在社会科学中的应用日益广泛，其生成数据是否能反映真实社会图景仍存争议。以中国综合社会调查（CGSS2021）为基准，构建多模型对比实验框架，系统评估不同 LLM 生成“硅基样本”的拟合度与偏差特征，可发现，主流模型可较好复现宏观变量间的统计关系，但存在表征偏差，易强化主流话语、忽略边缘声音。通过引入思维链（Chain-of-Thought）分析，发现模型在生成评分理由时呈现出标准化的因果推理结构，反映其潜在的社会观念建构路径。此外，提示策略与微调机制可能无形中影响模型对公共议题的认知方式。LLM 在社会测量中既存在潜能也有局限，建议未来应提升数据多样性、模型可解释性，并推动社会科学领域的专用大模型发展。

关键词　大语言模型（LLM）　整体事实　数据偏误　对齐机制　CoT

作者龚为纲，武汉大学社会学院/人工智能学院教授（湖北武汉 430072）；黄思源，武汉大学社会学院博士研究生（湖北武汉 430072）。

中图分类号 C91　　**文献标识码** A　　**文章编号** 0439-8041(2025)06-0123-14

一、引言

近年来，以大语言模型（Large Language Models，LLMs）为核心技术的生成式人工智能（Generative AI）正深刻重塑社会科学的研究范式。以自然语言处理（Natural Language Processing，NLP）为基础，LLM 能够高效分析和生成文本，使其成为计算社会科学（Computational Social Science，CSS）中的重要工具。[①] 借助这一技术，研究者以前所未有的规模和深度探究社会现象，包括舆论演化、社会网络互动、情绪传播、意识形态分化等诸多议题。[②] 然而，LLM 在社会科学中的应用也引发了关于其可靠性的讨论，包括对社会认知、数据偏误和研究方法论的反思。[③] 这一问题不仅涉及 LLM 的训练数据和算法对齐（Alignment）过程，也影响其作为社会测量工具的适用性。

在此背景下，本文试图探讨以下问题：首先，LLM 生成的数据是否能够真实反映“整体事实”，抑或其仅是训练语料和对齐机制塑造下的带有偏误的样本？这一问题涉及 LLM 的训练数据来源及其对社会观念的再现能力，尤其是在不同文化、语言和社会群体间的表现差异。其次，在社会科学研究中，LLM 是否

① 张咏雪：《从自动化技术到生成式人工智能——技术对劳动者影响的技能异质性研究》，《社会学研究》2024 年第 4 期。

② Marino E. B., Benitez-Baleato J. M., Ribeiro A. S., “The Polarization Loop: How Emotions Drive Propagation of Disinformation in Online Media—The Case of Conspiracy Theories and Extreme Right Movements in Southern Europe,” *Social Sciences*, 2024, p. 603.

③ Khudabukhsh A. R., “Deceptively simple: An outsider’s perspective on natural language processing,” *AI Magazine*, 45(4), 2024, pp. 569-582.

可以作为测量人类观念的有效工具？LLM在训练过程中累积了海量的社会认知经验，能否充分且准确地再现宏观社会事实与群体趋势？其与传统社会调查或官方统计的结果相比，吻合度如何？如果LLM的生成内容能够准确模拟真实社会调查数据，其或许能在传统社会调查方法（如问卷、访谈）受限的情况下提供有价值的补充。然而，若LLM生成的观念和态度模式与真实人类数据存在显著偏差，则其在社会科学研究中的适用性需要进一步审慎评估。本文希望厘清LLM在社会科学研究和知识生产中的潜力、局限，分析其可能引发的社会认知影响，为未来相关研究提供理论和方法上的参考。

二、文献综述

长期以来社会科学研究主要依赖问卷调查、访谈、实验等传统方法，以获取与人类社会行为和观念相关的数据。然而，这些方法存在诸多局限，例如数据收集成本高、样本代表性不足、受访者反应偏差等问题。计算社会科学的兴起部分弥补了这些不足，尤其是LLM的应用使得大规模文本分析成为可能。[①] 近年来，研究者开始利用LLM模拟社会个体，以测试其是否能够复现真实社会调查的数据分布。[②] 部分研究发现，LLM能够在一定程度上模拟社会群体的行为模式，例如在社会心理学、经济学实验和政治学研究中，其表现出较高的预测性。[③] 这一能力使得LLM被视为一种新的社会研究工具，能够帮助研究者克服传统调查方法的不足。

“整体事实”是一种从整体视野出发所揭示的社会过程中的动态因果结构与演变过程，其强调了对社会现象的宏观审视及在时间与空间维度上的系统性理解。[④] 因此，相较于割裂的、碎片化的或情感化的“局部事实”，“整体事实”具有更高的综合性与跨尺度性。可以说，“局部事实”是在特定的观念视角或目标导向下，被选择性抽取、夸大或在传播过程中情绪化处理的事实片段，它们不足以反映社会现象的全貌，反而可能消解对“整体事实”的认知和理解。随着数字化转型的深入，社交媒体及互联网平台塑造了一个高度碎片化的信息传播格局，在这一环境中，基于即时情绪关注的“局部事实”被大幅度放大，由此进一步削弱了从整体视角呈现社会事实的可能性。数字时代中的主体，由于身处于被局部“事实”裹挟和情感驱动的传播空间，逐渐丧失了从整体视角反观社会过程的能力。[⑤]

LLM生成内容能否有效表证“整体事实”，在学界仍存在争议。[⑥] LLM的训练语料往往来自互联网文本，虽然涵盖了广泛的信息，但其数据来源的不均衡性可能导致某些社会群体的观点被过度放大或削弱。此外模型训练过程中的对齐机制也会对模型输出产生影响。例如，在对职业、性别和收入之间关系的分析中，不同LLM的词嵌入结果表现出显著差异，表明模型的训练数据和校正策略可能影响其对社会观念的模拟能力。[⑦] LLM的输出往往会体现开发者设定的价值偏向，从而影响其对社会现实的客观呈现。

LLM不仅仅作为社会事实的测量工具，其生成机制本身亦可能深度参与社会认知的建构过程。随着人类与AI的互动增多，人们的观点可能逐渐受到LLM输出的影响，形成对现实世界的某种“对齐”。[⑧] 在社

---

① 龚为纲：《大语言模型助力计算社会科学迭代》，《中国社会科学报》2024年3月8日；Suh S. C., *Artificial Intelligence for Design and Process Science*, Cham: Springer Nature Switzerland, 2025.

② 梁玉成：《基于生成式大语言模型的“测试社会学”》，《探索与争鸣》2024年第11期。

③ 吕鹏：《大型社会模拟器：社会知识生产的大科学装置》，《探索与争鸣》2024年第11期；Ferraro A., Galli A., La Gatta V., et al., “Agent-Based Modelling Meets Generative AI in Social Network Simulations,” in *Social Networks Analysis and Mining*, Aiello L. M., Chakraborty T., Gaito S. (eds.), Cham: Springer Nature Switzerland, 2025, pp. 155-170.

④ 梁玉成、马昱堃：《对青年的计算文本“远读”——数字时代基于降维的整体认识论》，《青年探索》2022年第3期。

⑤ 邱泽奇：《数字化与社会学的时代之变》，《中国社会科学评价》2024年第2期。

⑥ Gross R., “Stochastic contingency machines feeding on meaning: on the computational determination of social reality in machine learning,” *AI & SOCIETY*, 2024.

⑦ Marino E. B., Benitez-Baleato J. M., Ribeiro A. S., “The Polarization Loop: How Emotions Drive Propagation of Disinformation in Online Media—The Case of Conspiracy Theories and Extreme Right Movements in Southern Europe,” *Social Sciences*, 13(11), 2024, p. 603.

⑧ Garcia B., Qian C., Palminteri S., “The Moral Turing Test: Evaluating Human-LLM Alignment in Moral Decision-Making,” arXiv, 2024.

交媒体环境中，AI 生成的文本可能加剧观点极化，使不同群体在信息环境中进一步割裂。[①] 此外，LLM 的预测性与其数据来源密切相关，不同的训练语料和模型微调方法可能导致其在不同文化或语言环境下表现出不同的社会偏见。[②]

LLM 可以在社会调查和经济实验中表现出类似人类的行为。有学者使用 LLM 模拟选民行为，并预测美国大选结果，发现模型生成的结果与人类调查数据具有一定的一致性。[③] 此外，也有学者利用 LLM 模拟城市居民的日常活动，显示出 LLM 能够在一定程度上再现现实中的社会动态[④]，甚至可以用于取代传统 ABM 的模拟方法[⑤]。

LLM 生成的数据仍可能存在系统性偏误。有学者对多个 LLM 在博弈实验中的表现进行分析，发现模型在再现人类行为时存在显著的偏差，尤其是涉及复杂的策略性推理时，其生成的分布与人类数据存在较大差异，或是相同角色设定下，LLM 在不同情境下可能给出相互矛盾的回答[⑥]，这可能影响其在社会科学研究中的可靠性；通过提供人类行为样本作为输入，LLM 的生成结果则可以更接近真实的调查数据。[⑦] 然而，即使采用优化技术，目前的 LLM 仍然无法完全复制人类行为，其适用范围仍需进一步探索。

三、LLM 的数据特性

随着数字传播环境的演化，尤其是在社交媒体和搜索引擎等平台的算法呈现机制下，社会认知的碎片化现象愈加明显。[⑧] 在这一背景下，LLM 所生成的数据能否反映整体事实，抑或仅是对训练数据中的偏见和局部事实的放大，成为一个关键的社会科学研究议题。

（一）LLM 的训练数据：整体事实的潜力与局限

LLM 常被视作一种能够“洞察”宏观社会现实、呈现“整体事实”潜力的创新工具。然而，数据本身并非中立载体，任何知识的生产都受到社会、文化与历史等因素的共同形塑。对于 LLM 而言，其训练数据更是这一复杂过程的核心：一方面，规模庞大、来源多元的文本语料为模型提供了高度广阔的知识基础，使其在理论上具备从宏观层面“再现”社会事实的潜力；另一方面，数据本身的代表性不足、分布不均衡以及模型后期对齐机制的干预，也会在结构与内容上限制其再现能力。

基于此，本文从数据来源与多样性、数据分布的偏差以及对齐机制的干预三个方面，探讨 LLM 训练数据在社会科学研究中所面临的机遇与局限，并进一步说明在利用 LLM 进行社会现象分析时，研究者应如何在技术潜力与社会结构交互性之间实现平衡与发展。

1. 数据来源与多样性。

理论上，LLM 的训练数据往往跨越不同领域与社会群体，包括互联网开放文本（新闻、博客、社交媒体）、学术文献、政策文件与各类书籍等。[⑨] 这种“大规模+多样化”的数据来源使模型能够同时捕捉主流与边缘、专业与大众等多重视角。对于社会科学研究而言，学术文献与政府报告提供了相对系统且经过一定审查或同行评议的知识框架，而社交媒体与新闻报道则能反映更具时效性与草根性的社会情绪和议题偏好。当 LLM 在训练过程中“整合”这些多元信息时，便有机会在理论层面上构建一种多学科、多视角交

---

① 李春南、王山：《生成式人工智能时代的语言安全：系统性风险与治理路径》，《国际安全研究》2024 年第 5 期。

② 吴冠军：《大语言模型的技术政治学评析》，《中国社会科学评价》2023 年第 4 期。

③ Jiang S., Wei L., Zhang C., “Donald Trumps in the Virtual Polls: Simulating and Predicting Public Opinions in Surveys Using Large Language Models,” arXiv, 2025.

④ Yan Y., Zeng Q., Zheng Z., et al., “OpenCity: A Scalable Platform to Simulate Urban Activities with Massive LLM Agents,” arXiv, 2024.

⑤ Ju D., Williams A., Karrer B., et al., “Sense and Sensitivity: Evaluating the simulation of social dynamics via Large Language Models,” arXiv, 2024.

⑥ Huang Y., Yuan Z., Zhou Y., et al., “Social Science Meets LLMs: How Reliable Are Large Language Models in Social Simulations?” arXiv, 2024.

⑦ Gao Y., Lee D., Burtch G., et al., “Take Caution in Using LLMs as Human Surrogates: Scylla Ex Machina,” arXiv, 2025.

⑧ 吕鹏、龚顺、梅笑等：《智能社会的崛起和人工智能的社会影响》，《智能社会研究》2022 年第 1 期。

⑨ Ju Y., Ma H., “Training Data for Large Language Model,” arXiv, 2024.

叉的社会认知图景。例如，在社会舆情分析与政策评估的研究中，LLM可以同时参考学术研究中的宏观理论模型与社交媒体中的微观个体表达，从而在一定程度上校正单一视角所带来的局限，形成对政策影响和公众情绪的综合性判断。在跨文化研究领域，利用多语言、多地域的训练数据，LLM还能够对不同社会背景下的文化要素进行比较分析，提升研究者对跨文化沟通与冲突的理解深度。

从计算社会科学的角度来看，对海量文本数据的处理和分析常被认为能为社会科学提供一种获取“整体事实”的研究路径。通过挖掘文本中的语言、话语与情感模式，研究者期待获得以往依赖小样本或质性研究难以捕捉到的宏观结构。在这一过程中，LLM可以协助研究者快速提炼多样化的信息内容，如不同社会群体对公共议题的态度差异、跨国比较视角下的价值观冲突等，展现出对“整体事实”的潜在映射能力。有学者提出LLM“算法保真性”的观点，认为模型内部关于思想、态度和社会文化背景的复杂关系模式，能够具有一定程度上准确表证特定人群态度分布的能力①，通过简单实验证明LLM可以准确模拟人类群体的思想模式，且在投票预测、党派态度等任务上与真实调查数据高度一致。我们可以将这一理论以数理社会学方式进行呈现：

设社会群体空间为测度空间(Ω,F,μ)，其中Ω为全体社会成员，F为可测群体结构，μ为人口分布测度。给定身份特征向量$I=(a_1,\cdots,a_m)\in A\subset R^m$（$a_i$表征年龄、性别、教育程度等人口学参数），则LLM的条件生成过程可表述为：

$$P(y\mid x,I)=\int_Z P(z\mid I)\ P(y\mid x,z)\ \mathrm{d}z \tag{3.1}$$

其中潜在变量$z\in Z$表示训练语料中隐含的亚群体响应模式（subgroup response pattern）。根据大数定律，当训练数据量满足$|D|\geqslant N(\varepsilon,\delta)$时，存在算法保真性：对于任意真实子群体$G\subset U$($U$为总体)，存在模型参数$\theta$使得：

$$D_{|KL|(P_{|G|}P_{|\theta|})}\leqslant\epsilon \tag{3.2}$$

这表明通过恰当的身份条件设置，模型能够从高维响应分布中析取出与真实群体高度契合的概率子空间。

LLM的训练数据构成其认知世界的“符号资本库”，本质上是对人类社会实践的数字化摹写。这种通过参数化表征建立的社会认知范式，既蕴含再现整体事实的符号潜能，又受制于数据生产的社会拓扑学限制与技术治理的意识形态规训。或者说，“整体事实”的建构仍然是理论化与实践操作之间的动态过程。即使LLM能够在文本层面汇聚多源知识，其对社会现实的呈现依旧有赖于数据本身的覆盖广度和结构。如果训练数据从一开始就缺乏对某些群体或地域的完整记录，那么所谓的“整体性”也难免出现偏颇。

2. 数据分布的偏差。

LLM通过大规模文本数据训练，以模拟人类语言和社会观念，实现对社会知识的结构化建模。然而，尽管LLM在某些情况和特定条件下能够捕捉、模拟和生成近似于“整体事实”，其生成内容的再现效能受到训练数据分布特征的显著制约，导致如图1所示的数据偏差（Data Bias）或数据分布偏差（Data Distribution Bias），这些偏差可能对研究结果的信度和可解释性构成挑战，甚至引发了方法论上的质疑。②

数据分布的偏差主要体现在：（1）语言与文化的代表性问题，（2）社会群体在数据中的不均衡性，（3）议题选择的系统性偏差。这些因素共同影响LLM在生成内容时的真实性，使其可能偏离整体事实，而倾向于特定文化、群体或主流话语框架。

由于训练语料以英语为主且文化分布不均，LLM易形成西方话语主导的知识框架，难以真实反映多元

① Argyle L. P., Busby E. C., Fulda N., et al., “Out of One, Many: Using Language Models to Simulate Human Samples,” *Political Analysis*, 31(3), 2023, pp. 337-351.

② Grossmann I., Feinberg M., Parker D. C., et al., “AI and the transformation of social science research,” *Science*, 380(6650), 2023, pp. 1108-1109.

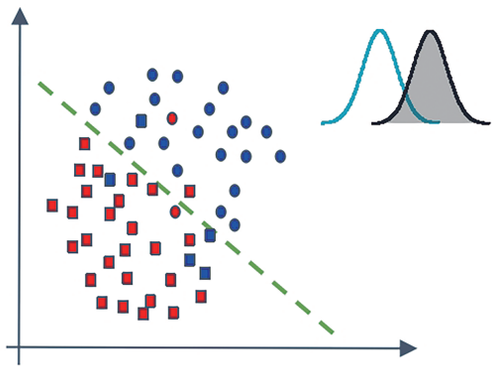

图 1　数据分布偏差的二维表示

社会现实。同时，高活跃度群体在语料中占优，边缘声音被系统性忽视，导致模型模拟出的社会观念呈现出代表性失衡与议题偏倚，加剧了全球知识表达的不平衡。

知识的生产、分配与流通总是与权力结构交织在一起[①]，LLM 的训练数据也不例外：偏见的产生不仅源于语言或群体代表性的数量不均衡，更是受制于社会中不同主体的话语权力与符号权力。当学术资源、媒体话语与在线平台的算法机制都倾向于某些主流观点时，LLM 从中“学习”到的社会事实也就自然“拷贝”了此种不对称的权力分布。对社会科学研究而言，这种现象意味着，若不对数据源的结构性偏差进行充分审视与修正，LLM 所生成的结论可能进一步加强既有的社会不平等和刻板印象，而非起到“揭示真相”的作用。即使计算社会科学的方法可以在一定程度上帮助我们识别和调整 LLM 在数据分布上的偏差，提高其对整体事实的再现能力，然而，LLM 仍然受到数据来源、社会群体代表性和议题覆盖范围的影响，研究者在使用 LLM 进行社会科学研究时，不得不考虑这些数据分布偏差可能给研究过程及结论带来的影响。

（二）LLM 的社会认知塑造与对齐效应

除了作为社会观念的测量工具，LLM 自身也可能成为重构社会认知的力量。这一现象可能加剧信息环境中的认知趋同，使个体在面对复杂社会问题时更容易接受 LLM 所提供的框架，而忽视其可能存在的偏误。

1. 作为“认知中介”。

社会认知理论认为，个体的认知模式在很大程度上受到环境输入的影响，在信息环境日益算法化的背景下，人们接触的信息越来越多地由人工智能生成，在语境建构和信息组织中，LLM 日益扮演“认知中介”（Cognitive Mediator）的角色，即其生成的文本不仅影响人们对外部世界的理解，还可能重塑他们的态度、信念和价值观。[②] 不同的语言结构和表达方式会影响个体对信息的解读，而 LLM 由于其训练数据的特点，在信息的组织和输出方式上可能与传统媒介有所不同，其文本生成往往呈现出高度流畅、逻辑自洽的特点，这使得用户更倾向于将其内容视为权威性知识，而忽略其潜在的偏差与局限性。尤其在涉及复杂社会议题时，LLM 的输出并非单纯地再现客观事实，而是在已有文本语料的基础上进行知识重构，而这一过程可能在无形中影响着用户的认知模式和价值判断。例如，在城乡发展问题上，LLM 可能会倾向于强调如通过土地流转、市场经济带动农村产业升级等市场机制在资源配置中的作用，而较少提及政府如何通过财政转移支付、乡村基础设施建设等方式推动农村发展。这可能影响人们对城乡关系的认知，使市场化手段的作用被放大，而国家宏观调控的贡献被弱化。这一倾向并非源于 LLM 对现实的独立分析，而是由于训

① 米歇尔·福柯：《惩罚的社会》，陈雪杰译，上海：上海人民出版社，2018 年。

② Ziems C., Held W., Shaikh O., et al., “Can Large Language Models Transform Computational Social Science?” *Computational Linguistics*, 50 (1), 2024, pp. 237–291.

练数据本身的统计分布使然。这种语言生成机制意味着 LLM 在传播知识的同时，也在塑造知识的边界，并可能导致社会认知的单一化。

2. 算法驱动的社会认知演变。

LLM 作为社会互动的一部分，其信息筛选和呈现方式在深度参与社会认知结构的动态重塑过程。在数字传播环境中，信息的流通模式已逐步由传统媒体主导的线性传播模式转向由算法驱动的个性化推荐系统，而 LLM 进一步加速了这一进程。由于 LLM 具备强大的语境适应能力，其输出能够根据用户的提示词进行调整，从而在一定程度上“迎合”用户的预设立场。这种迎合机制与社交媒体的“回音室效应”（Echo Chamber Effect）相似，即个体在信息选择过程中倾向于接受符合其既有认知的内容，而忽视或排斥与自身观点相左的信息。长此以往，LLM 不仅不会提供更加多元的社会认知视角，反而可能进一步固化社会认知结构中的偏见，加剧群体间的信息隔阂与认知极化。这一现象在政治传播、文化讨论以及科学议题的公共论争中尤为显著。例如，在全球气候变化的议题上，LLM 在不同的语境中可能提供截然不同的表述方式：在倾向环保主义的语境下，其回答可能更强调“气候危机”的紧迫性，而在商业导向的环境下，则更可能采用“能源转型”这类语义中性的表述方式。这种语言框架的调整不仅是对既有数据的复现，同时也影响着公众对现实议题的认知方式。

3. 社会知识体系的重构。

值得注意的是，LLM 的“对齐效应”不仅体现在个体层面，也可能在更大范围内影响主流社会知识体系的生成机制与合法性结构。随着人工智能在学术研究、政府决策、新闻传播等领域的广泛应用，LLM 被越来越多地用于信息整合和政策分析。然而，其训练数据主要集中于主流出版物、新闻报道和学术论文，这些来源本身就存在出版偏倚和意识形态筛选等问题，导致 LLM 难以避免继承其潜在偏见。以中国家庭结构为例，LLM 在分析时往往依据全球婚育趋势进行推测，较少纳入儒家文化、家庭养老压力、“躺平”现象等本土因素。如果训练数据缺乏社交媒体和地方文化文本，其对婚姻观念或新生代就业认知的分析更可能倾向全球化下的个体主义叙事，而忽视中国社会内部关于家庭责任、代际支持、单位体制等复杂变量。这种认知偏差若未被警觉，可能误导社会科学研究者对中国现实的理解，甚至在政策制定中造成判断偏差，对知识生产带来潜在风险。

这一现象对社会科学研究的可信性提出了新的挑战。传统的社会科学研究强调多视角分析和实证验证，而 LLM 生成的文本往往缺乏可追溯的数据来源和明确的证据链，这可能导致其在科学研究中的适用性受到质疑。尤其在涉及公共政策和社会治理的议题时，政策制定者若过度依赖 LLM 进行分析，可能会导致决策过程过度技术化（Technocratic Policy Making），即政策的制定更多地基于算法模型的推演，而非来自社会现实的多方讨论。一些政府单位已经开始使用 LLM 进行社会舆论分析，并据此调整政策宣传策略，但如果 LLM 生成的舆论数据存在偏差或被特定话语体系主导，那么这种“数据驱动”的决策方式不仅不会提高治理效率，反而可能削弱决策的多元性。因此，在 LLM 参与社会认知塑造的过程中，如何确保信息生成的透明度、减少算法偏倚、提高多样性和公平性，已成为当代社会科学研究亟待解决的重要问题。

总体而言，LLM 在社会认知塑造中的作用日益凸显，其影响不仅体现在个体认知的微观层面，也深刻作用于社会知识体系的宏观结构。然而，这种影响并非绝对中立，而是受到数据来源、模型优化、用户互动模式等多重因素的塑造和调节。计算社会科学的研究表明，LLM 的“对齐效应”在一定程度上强化了社会观念的同质化趋势，并可能影响公共知识体系的开放性和多元性。未来的研究需要进一步探索如何优化 LLM 的训练方法，使其在知识生产和社会认知塑造过程中更具透明性、公正性和可解释性。只有在确保数据多样性、模型公正性和用户意识提升的前提下，LLM 才能真正成为社会科学研究的有效工具，而非无形中固化社会偏见和成为认知失衡的助推器。

四、LLM 的社会模拟：基于硅基样本的测量

随着 LLM 在社会科学领域的应用日益深入，其不仅被视为社会测量工具，也逐渐成为社会科学研究

的潜在测量对象[①]，模拟真实社会调查的结果。传统的社会调查方法，如问卷调查、访谈研究和实验设计，依赖于人类受访者的数据采集，而 LLM 生成的“硅基样本”是否能够复现真实社会调查数据，成为计算社会科学研究的新兴议题。在理论层面，若模型具备足够的算法保真度，则其具备通过条件生成机制模拟特定人群态度分布的能力，从而再现真实调查数据的统计特征［见公式（3.2）］。然而，LLM 在实际应用中是否能够可靠地复现代际观念、社会价值观和群体行为模式，仍然缺乏系统的实证检验。以下通过一个社会经济地位与社会公平感关系的简单案例，比较 LLM 生成的数据与中国综合社会调查（CGSS）等真实社会调查数据之间的相关性，以评估 LLM 在社会测量中的一致性和有效性。

（一）研究方法与实验设计

本文以中国综合社会调查（CGSS2021）数据为基准对照组，通过 LLM 生成模拟实验组数据，系统性评估 LLM 在社会测量中的保真性。研究聚焦社会经济地位（SES）与社会公平感（Perceived Social Fairness）之间的结构性关联机制，通过统计分析与数据分布对比，探讨 LLM 生成数据对真实社会态度的复现能力及其作为社会科学研究工具的可行性。

1. 数据来源与研究变量。

真实调查数据来源于 CGSS2021 全国性社会调查，该数据集包含社会经济地位、社会公平感知等核心变量的标准化测量结果。经过数据清洗剔除无效和空白样本后，最终保留 5249 份有效受访者数据作为对照组。实验组数据由 LLM 模拟生成，通过结构化提示词引导模型以虚拟受访者身份完成问卷回答，旨在捕捉不同 LLM 对社会态度测量任务的响应特征。研究重点观测变量包括两个 5 分量表指标：社会经济地位（1=上层至 5=下层）反映受访者主观阶层定位，社会公平感（1=完全不公平至 5=完全公平）表征个体对社会资源分配体系的整体评价。同时引入性别、年龄、教育程度、婚姻状况等八项人口学变量作为生成硅基样本的参数。

2. 样本描述。

本文采用的人口学变量包括性别（gender）、年龄（age，截至 2021 年）、教育程度（education）、婚姻状况（marriage）、2020 年的总收入（income）、工作经历（work_experience）、当前工作状态（work）、职业类型（occupation）、社会公平感（fairness）和自评社会经济地位（ses）。其中，社会经济地位变量为受访者对自身社会层级的主观评价，采用 5 分量表测量（1=上层，2=中上层，3=中层，4=中下层，5=下层）；社会公平感变量则衡量受访者对当今社会公平状况的感知，采用 5 分量表测量（1=完全不公平，2=比较不公平，3=说不上公平但也不能说不公平，4=比较公平，5=完全公平）。

调查结果显示，约 57.08% 的受访者认为社会“比较公平”，21.66% 选择“中立”选项，极端评价较少（“完全公平”6.89%，“完全不公平”1.20%）。社会经济地位方面，自评为中层及以上者占 46.85%，中下层及以下占 53.15%，显示受访者对社会地位认知的广泛差异。

3. LLM 与实验设计。

实验设计采用多维度对比框架，设立五个差异化实验组以解析 LLM 性能的影响因素。实验组 A 与 B 分别采用最新版本的 DeepSeek-R1 和 Llama3.3 模型[②]，严格遵循 CGSS2021 原问卷格式生成数据，旨在考察不同模型架构对测量结果的影响。实验组 C 在 DeepSeek-R1 模型基础上改用两次对话进行问卷模拟，一方面用于检验不同生成方式输入对 LLM 响应模式的潜在偏差。实验组 D 与 E 将 5 分量表扩展为 0—100 分连续量表，分别通过 DeepSeek-R1 和 Llama3.3 生成数据后按五分位区间离散化，以此探究量表精细化对数据分布特征的改进作用。模型生成温度参数（Temperature）均设置为 0.6，均使用零样本（Few-shot）方式。

---

① 梁玉成：《基于生成式大语言模型的“测试社会学”》，《探索与争鸣》2024 年第 11 期。

② DeepSeek-R1 使用 BF16 精度的 DeepSeek-R1 本地模型，参数量为 671B，其开发者未披露训练数据的截至日期（Knowledge cutoff），第三方测试的截至时间为 2024 年 7 月。Llama 3.3 使用 BF16 精度的 Llama-3.3-70B 指令微调的本地模型，参数量为 70B，数据截至时间为 2023 年 12 月。

表1 实验组与对照组的设置及比较维度

组别	使用模型	Prompt	量表类型	核心对比维度
实验组A	DeepSeek-R1	单次生成	5分量表（1—5级）	模型基准性能（基于原生Prompt）
实验组B	Llama3.3	单次生成	5分量表（1—5级）	模型架构和训练数据差异对比
实验组C	DeepSeek-R1	多次对话	5分量表（1—5级）	Prompt对话方式影响
实验组D	DeepSeek-R1	单次生成	100分量表（0—100分转换）	测量精度改进
实验组E	Llama3.3	单次生成	100分量表（0—100分转换）	模型架构与量表交互作用
对照组	N/A	N/A	5分量表（1—5级）	N/A

使用的系统提示（System Prompt）为：“你将扮演一位生活在2021年的中国人，你正在接受一份访谈，请根据你的认知，如实回答提出的问题。你的性别为｛gender｝，年龄为｛age｝岁，婚姻状况为｛marriage｝，工作情况是｛work_format｝，个人2020年的总收入为｛income｝元，您自认为的经济地位属于｛ses｝。作为受访者，您只需要告诉我你的答案，不要进行任何解释和说明。”① 其中，各变量在实际生成时替换为真实的样本。

按原问卷5分量表和扩展后的100分量表所使用的用户提示（Prompt）分别为：“我的问题是：总的来说，您认为当今的社会公不公平？请选择：‘完全不公平’‘比较不公平’‘说不上公平但也不能说不公平’‘比较公平’‘完全公平’。”和“……请从0—100中选择一个分数来表示您认为社会公平与否的程度，0为完全不公平，100为完全公平”。

对于实验组C，在使用系统提示作为第一次对话提示，待模型生成内容后，使用用户提示作为第二次对话提示。

4. 研究假设。

由此可以提出三条假设：

H1（一致性假设）：在相同的人口统计学变量匹配条件下，LLM生成的数据分布与CGSS真实调查数据之间具有较高的一致性。

H2（关联性假设）：LLM生成的数据中变量间的关联模式与实际社会调查数据相比可能存在系统性的偏差，具体表现为变量间关联强度的系统性高估或低估，以及关联模式的简化趋势（如过度单一化或线性化社会经济地位对社会公平感的影响）。

H3（模型影响假设）：不同LLM之间的测量能力可能存在差异，例如DeepSeek-R1和Llama3.3下可能表现出不同的社会态度分布。

H1和H2分别探讨了LLM生成数据与真实社会调查数据之间在不同层面上的差异。前者关注的是数据整体分布的再现程度，后者则更深入地探讨变量之间的关联模式差异，强调社会经济地位对社会公平感影响的强度和结构在LLM生成数据中可能发生系统性偏离。H3则强调了不同LLM模型架构及训练数据对测量结果的可能差异影响。

（二）研究结果与发现

结果如图2所示。由于触发模型审查机制，实验组C和实验组D部分样本有缺失，但仍在允许误差范围内。

研究结果表明，不论采用原问卷的5分量表还是修改后的100分量表，不论使用哪种基座模型，不论采用单次生成还是多次对话，LLM生成的样本在相关性方向和样本基本分布情况都与真实社会调查数据在变量分布与方向性上呈现出较高程度的一致性。这表明LLM具备一定的社会模拟能力，在生成硅基样本（Silicon Sample）时，能够保持基本的人类社会认知模式。

① 根据DeepSeek的相关说明文档，应避免设置DeepSeek-R1模型的系统提示。对于该模型，本文将系统提示与用户提示合并。

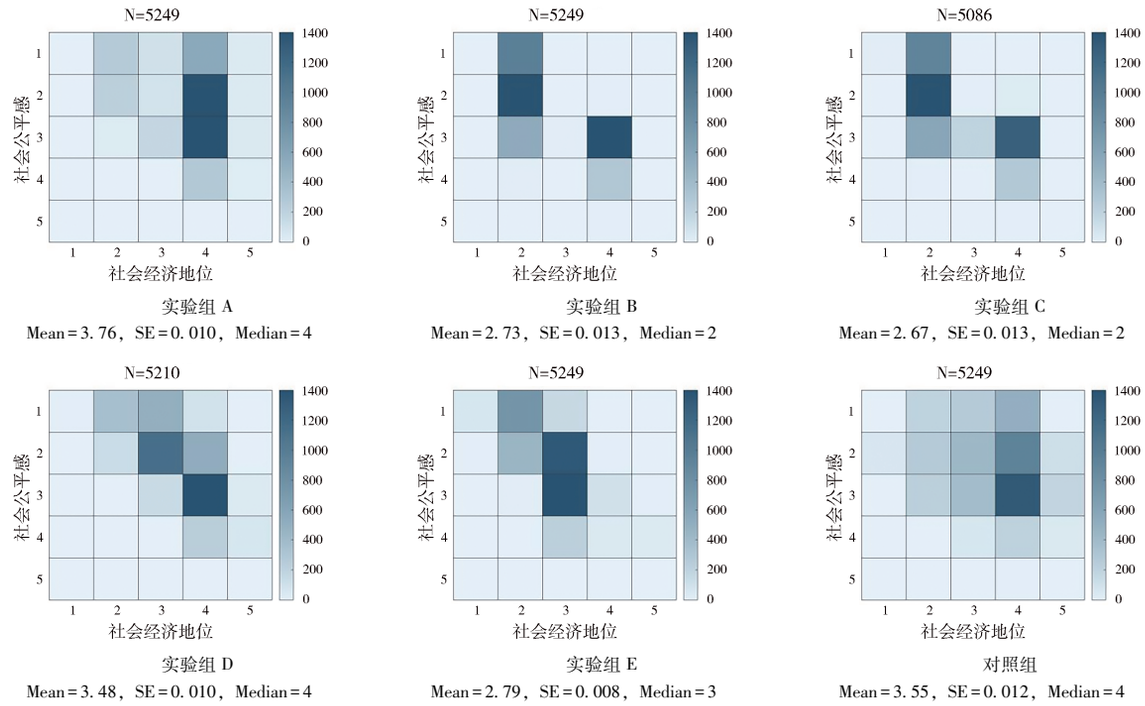

图 2　实验组与对照组数据测量结果

图中颜色深浅表示社会公平感评分（1=完全不公平至5=完全公平）在各组中的样本量，颜色越深表示该评分选项的样本量越大。各子图下方列出的统计量（Mean=均值，SE=标准误，Median=中位数）均为对应实验组或对照组社会公平感评分变量的描述性统计结果

通过图 3 的概率密度曲线可以观察到，使用 DeepSeek-R1 生成的实验组 A 和实验组 D 与对照组最为接近，实验组 C 和实验组 E 次之，可以表明 DeepSeek-R1 在社会调查数据模拟上较为稳定，尤其是当使用单次生成（One-shot Generation）时，其生成的数据分布最接近真实调查数据，而多轮交互的 Prompt 设计方式对 LLM 生成数据有一定影响，在这之中可能会引入新的偏差。

而 Llama3.3 使用 5 分量表的实验组 B，“说不上公平但也不能说不公平”（社会公平感=3）的样本数远少于在“比较不公平”（社会公平感=2）上的分布，与对照组分布相差较大。Llama3.3 在 5 级量表上的数据生成存在明显偏差，未能准确复现真实社会调查中的中间选项分布。转换为 100 分量表后（实验组 E），数据分布比实验组 B 稍有改善，但仍然显示出偏差，表明 Llama3.3 在量表尺度转换后，仍存在一定的架构偏差。

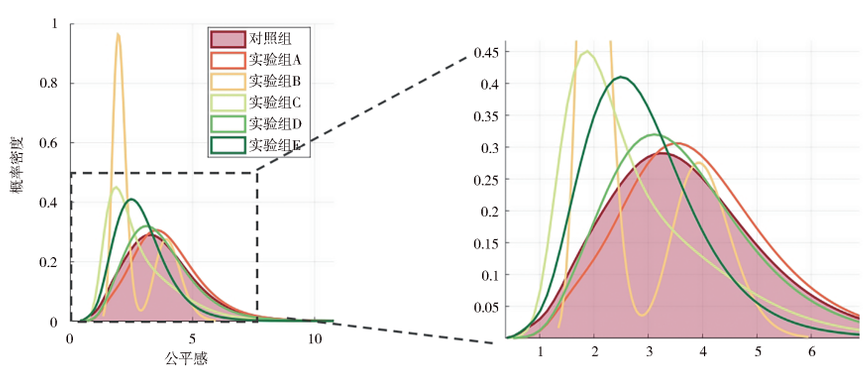

图 3　实验组与对照组概率密度比较

通过对比 LLM 生成的硅基样本与 CGSS2021 真实调查数据，在社会经济地位与社会公平感两个变量上的分布情况，验证了一致性假设（H1），即在相同的人口统计学变量匹配条件下，LLM 生成的数据能够较

为准确地再现真实社会调查的统计分布。这一实验结果表明，LLM 在模拟社会观念、重构社会事实的过程中，具备一定的算法保真性（Algorithmic Fidelity），并具有生成“整体事实”的潜力。

具体而言，单次生成条件下，DeepSeek-R1 在原始 5 分量表（实验组 A）与 100 分量表转换（实验组 D）中的表现，均与 CGSS 数据在关键变量分布和联合结构上高度一致，说明在输入充分的前提下，模型可较好模拟群体观念结构与变量间关系。

此外，相关性分析显示，LLM 生成的数据在社会经济地位与社会公平感间的相关系数与真实数据接近，表明其并非简单复制训练语料，而是在学习社会模式基础上生成新样本，体现出类似于自然语言任务中的泛化能力，即在复杂社会变量组合下生成符合统计规律的模拟数据。

实验结果亦支持模型影响假设（H3），即不同 LLM 在社会态度测量上的能力存在差异。以 DeepSeek-R1 与 Llama3.3 为例，在相同实验条件下，二者生成的数据分布明显不同。这种差异不仅与强化学习反馈（RLHF）机制相关，也源于训练数据构成的差异：Llama3.3 主要基于英语国家的公开数据，在指令微调时中文语料覆盖有限，而 DeepSeek-R1 中文语料权重更高，因而在模拟中文社会调查时表现更优。受训练语料分布影响，Llama3.3 在五级量表模拟中显示出对立场明确型选项的成生偏倚，呈现特定模式偏差。这一现象并不意味着 Llama3.3 具有极化效应，即模型不会倾向于生成极端态度，而是可能由于开发者在模型训练和强化学习调整过程中有意规避极端回答，使得其在较强约束下更倾向于生成社会上普遍可接受的回答，而非完全极端的观点。[①] 这种调整策略可能导致中间态度选项的概率下降，使得 Llama3.3 在社会调查模拟时低估了现实社会中模糊立场的存在。

此外，测量尺度的调整对 LLM 生成数据的影响也较为复杂。采用 100 分量表的实验组 D 和 E 相较于 5 分量表，在数据分布的连续性和精细度上有所改善，但在 Llama3.3 生成的实验组 E 中，这一调整未能完全消除偏差。这进一步表明，Llama3.3 在处理社会态度测量时的系统性误差较 DeepSeek-R1 更为显著，即便在尺度转换后，数据仍存在一定程度的测量不稳定性。这种现象可能与 Llama3.3 主要基于欧美社会文本训练，而欧美社会在社会议题上通常呈现出更强的共识性叙事（Consensus-driven Narrative）相关，使得 Llama3.3 在生成社会调查数据时，更倾向于符合社会规范的回答模式，而对模糊态度的再现能力相对较弱。

实验结果支持了关联性假设（H2），即由 LLM 生成的社会经济地位与社会公平感之间的关联模式，与实际社会调查数据相比可能存在系统性的偏差，具体表现为变量间关联强度的系统性高估或低估，以及关联模式的简化趋势（如过度单一化或线性化社会经济地位对社会公平感的影响）。如图 4 所示，实验组 B 至实验组 E 在“社会经济地位—社会公平感”变量上的 Spearman 相关系数均显著高于对照组（真实调查数据），显示出 LLM 生成数据中变量间的相差强度普遍高于真实数据，呈现出身为理想化的线性结构。这一现象表明，尽管 LLM 在整体变量分布层面可能较为接近真实数据（H1 支持），但在变量间的结构性关系建构方面却表现出一种“合理预期强化”的偏向倾向。也就是说，LLM 在生成社会态度数据时，倾向于输出符合其训练语料中主流叙事的关联路径，例如高社会经济地位者感受更强烈的社会公平感，从而弱化了现实社会中大

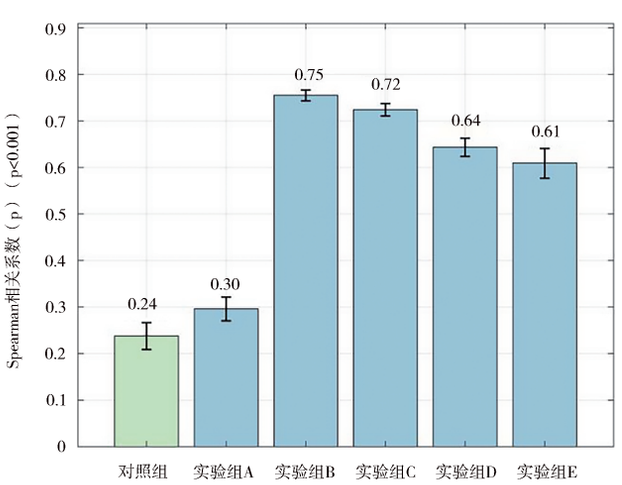

图4 各实验组与对照组中社会经济地位与社会公平感的 Spearman 相关系数比较

① Lin B., *Reinforcement Learning Methods in Speech and Language Technology*, Cham: Springer Nature Switzerland, 2025.

量存在的非典型个体模式（如高收入但感知社会不公者、低收入但认同社会公平者）。这种关联模式的单一化，实质上反映了 LLM 在学习社会模式时对主流统计结构的固化倾向，进一步印证了其生成数据中的关联性偏差特征。

LLM 生成中的系统性偏差，主要源于其训练数据的主流价值导向和生成机制。训练语料通常强调社会的稳定趋势，较少涵盖对立、边缘或极端观点，甚至可能在预处理阶段被有意过滤或降权处理。因此，在测量社会公平感时，LLM 更倾向于输出符合主流叙事的“合理”回答，而忽视现实中广泛存在的态度差异。此外，强化学习对齐（RLHF）技术进一步弱化了对社会矛盾和极端表达的呈现，使模型生成结果更趋向“温和”或“理性”立场，与真实调查中多样化、分歧显著的态度分布不尽一致。

另一方面，LLM 在生成评分时可能依据统计学习原则，将社会经济地位高的群体赋予更高公平感评分、低地位群体赋予更低评分。这种趋势可能符合统计学意义上的总体规律，但难以涵盖高收入却感知不公、低收入却认同公平等“非典型”个体。由于缺乏对个体心理、社会互动与历史经验的深度理解，模型容易高估变量间的线性关系，形成有偏样本。

（三）生成社会观的逻辑：从思维链到认知结构的跃迁

尽管本文验证了 LLM 在社会调查数据模拟中的一致性和关联模式的系统性偏差，并揭示了不同模型在测量能力上的差异，但 LLM 生成社会态度的具体考量机制仍然值得进一步探讨。相较于传统社会调查依赖的受访者主观报告，LLM 生成的硅基样本具有一个显著的优势，即其生成过程可追溯到模型内部的“思考”过程。得益于 DeepSeek-R1 具备长思维链（Chain of Thought，CoT）推理能力，我们可以通过分析模型生成社会公平感评分时的思考路径，提取出其考量的关键因素，并尝试理解 LLM 在模拟人类社会认知时的内在逻辑结构。

1. 语义生成的路径分析。

我们在实验组 A、实验组 C 和实验组 D 中，进一步设计了 CoT 输出追踪实验，即在生成社会公平感评分的过程中，截取模型结果“<think></think>”标签中的思考过程，作为评分的原因解释，并基于文本分析方法提取 LLM 进行判断时所依据的核心变量。随后，我们采用主题聚类和文本频次分析技术，以分析模型在生成社会态度时主要参考的因素。

经过清洗后，我们对 13681 个有效思考过程文本进行分析。通过主题聚类，我们发现其主要考虑的变量集中于经济、社会保障、政府政策、就业、教育资源、医疗资源六大类别（主要的考量因素主题如表 2 所示）。这表明 LLM 在生成社会公平感数据时，主要依赖结构性变量，而较少涉及个体主观经验或社会心理因素。相比于 CGSS2021 真实调查数据，这种模式反映出 LLM 生成的样本虽能再现社会变量之间的统计关系，但其测量结果可能偏向于“理性化、主流化”模式，忽视了个体层面的社会心理复杂性。

表 2　模型考量因素主题（部分）

考量因素主题	频次	考量因素主题	频次	考量因素主题	频次
城乡不平衡	8332	就业机会不均	659	中层经济地位	213
收入差距	3760	医疗资源分配不均	535	收入分配	210
教育资源分配不均	1821	政府努力	526	性别歧视	205
资源分配不均	1515	扶贫政策	517	政策改善	194
收入较低	1507	认可国家发展	501	个人经历	175
社会保障	1374	收入不稳定	488	教育和医疗资源分配不均	169
经济地位中下层	1009	政策支持	478	个人经济状况	157
收入分配不均	1006	市场竞争压力	351	失业状态	155
经济压力	980	医疗保障	327	国家发展	151
地区发展不平衡	909	对城乡不平衡感到不公	307	基础设施改善	150
自认为中层	785	社会进步	284	教育机会	149
收入分配不均	1006	市场竞争压力	351	失业状态	155
经济压力	980	医疗保障	327	国家发展	151

研究发现，LLM 在衡量社会公平感时主要依赖于经济状况、社会保障、政府政策、就业机会等结构性

变量，而较少涉及个体经验、社会心理和文化认同等因素。在经济因素方面，LLM 生成的数据更加强调收入差距、城乡发展不均等要素，表现出较高的社会经济地位与社会公平感的相关性，但相比真实社会调查数据，其在推理过程中更倾向于强化收入水平对公平感的决定性作用，而忽略了社会认知的复杂性。与此同时，LLM 在决策过程中高度关注社会保障体系，如医疗资源、养老体系、政府扶贫政策，这可能与其训练数据较多来源于政策文件和官方文本相关，使其在测量社会公平感时，较少反映个体基于社会互动和人生经历形成的公平认知。例如，DeepSeek-R1 由于其更高比例的中文政策类文本训练，使得其在社会公平感测量上表现出更强的政策导向性，而 Llama3. 3 由于更多基于全球互联网数据，其测量结果可能更受西方自由市场经济观念的影响。此外，模型在解释社会公平感时倾向于主流政策叙事，例如“政府在改善社会公平”或“社会经济发展带来公平感提升”，而对于社会心理、群体认同等因素的影响较少涉及，这可能导致其在测量社会公平感时放大了政策干预的作用，而低估了个体经验的多样性。在就业因素方面，LLM 生成的数据更加强调劳动市场竞争、工作稳定性等问题，但较少涉及职场歧视、社会关系网络、文化认同等更具主观性的公平认知模式。

整体而言，LLM 在社会公平感测量中展现出较高的统计一致性，但其生成的数据仍然受到训练数据的约束，在一定程度上强化了社会主流认知，而未能完整再现社会个体的多样性经验。因此，对于本文所述的实验而言，如果进一步提供与社会公平感高度相关的变量，如社会流动性和政策受益情况，LLM 可能会生成更加精确的硅基样本，减少测量偏差。但如果提供的是无关变量（如饮食习惯、兴趣爱好等），LLM 可能仍会试图在这些变量与社会公平感之间构建“合理”关联，进而可能引入虚假相关性或模式化偏误。这样的实验值得进一步探索，以测试 LLM 生成的社会调查数据的因果稳健性和变量敏感性，更进一步探讨 LLM 在多大程度上能够表征“整体事实”。本文认为，LLM 在社会科学测量中的适用性仍需进一步优化，例如通过扩展训练数据的异质性、引入社会心理变量、优化提示词设计等方式，提高其对社会测量的全面性。

2. 从评分理由到社会观：LLM 如何“思考”公平。

传统上，CoT 被视为辅助性分析工具，用于解释模型的输出路径或提升其在多步推理任务中的表现。在社会科学语境中，CoT 亦可被视为观察语言模型生成逻辑与社会认知建构的一种窗口。本文尝试将 CoT 理解为语言模型“构造社会观念”的机制路径，即：模型不仅在执行任务，更在通过语言生成的过程中，主动建构其对“社会如何运作”的拟态理解。

LLM 在模拟社会公平感评分时，所生成的理由语句通常呈现出因果链式的结构逻辑。例如，诸如“收入差距”“资源不均”“教育机会缺失”等词项频繁出现于评分理由之中，构成了模型对“社会不公平感”来源的基本解释框架。这些语句通常沿着“结构性原因→机会受限→心理反应→价值判断”的推理路径展开，表现出一种高度结构化的社会认知模式。这一模式在一定程度上与主流社会科学对结构性不平等的因果推理逻辑相契合，显示出 LLM 在训练语料中积累的社会知识表征能力。但同时也需指出，这种结构化路径易于简化个体经验的复杂性，忽视了现实社会中异质性的认知表达。边缘视角、反常态观点、以及多重因果路径往往在模型生成中被有意无意地弱化，从而导致“社会经济地位—社会公平感”之间关系在生成结果中被系统性放大。这正是本文所强调的“关联模式偏差”在认知层面的延展体现。

进一步而言，LLM 生成的“社会观念”，并非直接来源于对现实社会认知结构的模拟，而更可能是一种源自训练语料的主流叙事模式的统计重构。在这一过程中，模型学习并再生产了大量文本中所隐含的“社会常识”与制度逻辑，从而在生成中呈现出一种“合理世界”的预设图景。这种生成逻辑构成了硅基叙事空间中的社会观念投影机制，其对社会公平的“思考”更多体现为对主流价值链的提纯与强化，而非对个体经验的真实还原。

然而，需要强调的是，这一机制虽具结构性偏向，却也反映出 LLM 在宏观层面具备模拟主流社会趋势与因果结构的能力。CoT 中的因果链条，虽存在简化与规训，但也揭示了模型在处理复杂社会问题时能够调用一定的结构性知识框架进行语言构建，这为社会科学研究提供了一种有条件使用的工具性价值。

因此，我们可以认为，CoT 不仅是模型输出的解释框架，也是一种反映其社会观念生成逻辑的“认知

镜像”。它提示我们：LLM 在模拟社会态度时所表现出的“观念合理性”，是一种兼具算法统计性与语料构成性的复合产物。这一过程既展现了模型的“保真性潜力”，也揭示了其在结构性叙事中的偏差风险。

需要说明的是，以上关于 LLM 思维链与社会观念建构机制的讨论，作为对 LLM 生成数据为“整体事实还是偏误样本”这一核心问题的理论延展，意在探索语言模型生成结果背后可能存在的观念塑造路径。尽管该部分分析揭示了 LLM 在社会认知建构中可能具备的结构性偏向，但当前探讨仍属初步尝试，更多体现为对测量机制背后逻辑的一种可能性想象，而非对实证结果的直接补充。未来研究可在更系统的语料分析基础上，进一步验证 LLM 在观念生成层面的建构倾向，以深化我们对其在“复现整体事实”与“再生产结构性偏误”之间张力的理解。

五、讨论与展望：整体事实还是偏误样本

本文通过系统考察 LLM 在社会科学研究中的数据特性与认知塑造效应，揭示了其在社会事实表征中的双重属性：一方面，LLM 依托海量训练数据的统计建模与算法保真性（algorithmic fidelity），展现出模拟宏观社会趋势的潜力，在特定条件下能够反映“整体事实”的动态结构；另一方面，其生成数据的可靠性与有效性受到训练数据覆盖度、对齐机制的技术干预及提示工程的语境约束等因素的影响，导致模型态度生成易受社会主流叙事的隐性塑造，甚至可能强化既有偏见的结构性再生产。

研究表明，LLM 的“整体事实”再现能力本质上是技术理性与社会权力共同作用的结果。在结构性变量层面，LLM 生成的数据在社会经济地位与社会公平感的关系上表现出较高的统计一致性，能够较准确地模拟社会整体模式。值得注意的是，当训练数据具备跨群体、跨文化、跨层级的均衡性时，模型可以通过参数化表征构建社会变量的关联网络，并较好地复现真实调查数据的统计分布特征。

尽管 LLM 具备较强的统计建模能力，可在一定程度上模拟社会结构中的规律性模式，但在数据分布、认知偏差与社会叙事建构等方面仍存在显著局限。例如，模型在生成变量间协变结构时，易受训练语料中主流叙事影响，表现出“关联模式偏差”——即单变量分布虽接近真实数据，但变量间的相关性强度与结构常被系统性放大或简化。在社会经济地位与社会公平感的关系建构中，真实调查中存在大量非典型样本（如高 SES 感知不公者），而 LLM 则更趋向于生成符合“合理性预期”的主流路径，从而压缩了异质性。

其一，LLM 对个体经验的模拟缺乏社会互动与历史语境的嵌入性，其“硅基样本”本质上是统计模式的外推，易将训练数据中的聚合趋势误认为现实中的因果机制，进一步强化结构性变量间的既有关系，忽略真实社会中个体认知的多样性。在涉及群体认同、代际流动、政策信任等更复杂议题时，LLM 更易高估主流共识，低估社会冲突与分歧。其二，LLM 在个体层面的社会认知塑造能力有限，生成数据更像是对主流偏见的整合反馈，而非社会事实的全景再现。训练数据多来源于主流媒体、政策文件与学术文献，这些内容往往经由既有话语体系筛选，使模型在测量社会公平感时倾向高估政策调控作用，低估个体经验、群体抗争与历史情境等因素的影响，边缘叙事也常被系统性遮蔽。

因此，LLM 所生成的数据更接近对主流社会观点的系统性归纳，而非社会调查的真实复现。即便扩大训练数据规模，LLM 也难以彻底消除选择性偏差与算法推理的局限。其生成结果可在特定条件下作为社会调查的有益补充，用于拓展分析维度、识别潜在模式，但远不能取代传统社会调查方法。

本研究选用 DeepSeek-R1 作为实验工具，以检验其生成数据对“整体事实”的再现能力。结果显示，DeepSeek-R1 相较于通用 LLM 更具中国地方性知识优势，适合本土社会科学研究。作为近年来中国人工智能的重要突破，DeepSeek 不仅在模型性能上取得显著提升，也降低了 LLM 在社会科学领域的使用门槛，提升了其研究可及性，尤其适用于社会调查受限、政策评估复杂、社会态度难以量化等情境下的补充性测量。

我们同时呼吁推动面向社会科学的专业化 LLM 发展，不仅着眼于提升分析效率，更应精准对齐社会权力结构、解构数据话语体系，并以重构知识秩序为目标，更有效服务于社会科学研究。

DeepSeek-R1 具备的“思考”能力，使得社会科学研究者可以通过 CoT 更深入地洞见模型生成结果背后的逻辑推理过程，从而揭示模型在决策中的关键考量因素。通过这一能力，研究者不仅可以获得模型生成数据的最终输出，还可以追踪模型在生成数据时的推理路径，解析其依据哪些变量、如何进行关联推

理，以及是否在不经意间放大了特定社会偏见。DeepSeek-R1的这一特性不仅增强了LLM在社会科学研究中的透明度，也为研究者提供了一种新的“可解释性AI”研究方法。相较于传统的社会调查方法，DeepSeek-R1的CoT机制允许研究者在不直接干预模型运算逻辑的前提下，解析LLM在模拟社会态度时的认知模式。这意味着，研究者可以通过LLM生成的硅基样本，结合其思考链分析，构建出更具理论解释力的社会科学模型。例如，在社会分层、政策评价、公众舆论测量等研究中，研究者可以通过LLM生成大量虚拟调查样本，并基于其思考链追踪变量之间的关系，进而分析LLM在不同情境下对社会态度的模拟机制；再如本文对“社会经济地位—社会公平感”关系的考量中，可以进一步建立如图5的“社会公平感认知—原因标签—硅基样本特征”的三元关系组，将思考链文本做形式化建模，构建“社会认知图谱”，为后续通过网络分析进行机制解释提供新的思路。这种方法不仅拓展了LLM在社会科学测量中的适用性，也为未来的计算社会科学研究提供了更具智能化的分析框架。

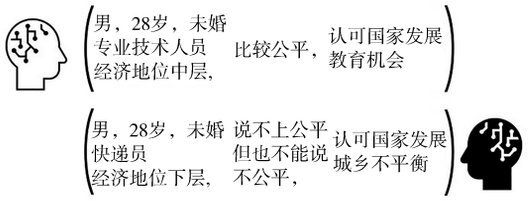

图5 “社会公平感认知—原因标签—硅基样本特征”三元关系组

我们认为，社会科学的未来，不应只是对数据的计算，而更应该是对人类社会脉络的深刻理解。LLM的出现让我们看到了技术如何辅助社会科学，让研究者能够以前所未有的规模模拟社会现象、探究人类观念的演变、努力向整体事实靠近。然而，无论模型多么先进，社会科学研究的本质始终无法被算法完全捕捉。经验质感的厚度、个体叙事的细腻、田野调查中对社会真实的感知，仍然是任何人工智能都无法取代的洞察来源。

〔本文系国家社会科学基金项目“基于大数据的社会情绪风险与网络集群事件治理研究”（22BSH024）、武汉大学社会科学数智创新研究团队项目“大国竞争背景下的战略情报分析”的阶段性成果〕

（责任编辑：朱 颖）

Approximating Facts or Reproducing Bias

—— Evaluating LLMs in Social Research

GONG Weigang, HUANG Siyuan

Abstract: The application of large language models (LLMs) in the social sciences is expanding rapidly, yet whether the data they generate accurately reflect real-world social phenomena remains contested. Using the 2021 Chinese General Social Survey (CGSS) as a benchmark, this study develops a multi-model comparative framework to systematically assess the representativeness and biases of “silicon-based samples” produced by various LLMs. The results show that mainstream models can reproduce statistical relationships among macro-level variables, but they exhibit representational biases—tending to reinforce dominant discourses while marginalizing alternative perspectives. Through the incorporation of Chain-of-Thought (CoT) analysis, we find that the models generate standardized causal reasoning structures when explaining their responses, revealing implicit pathways of social cognition embedded in their outputs. Furthermore, prompt design and fine-tuning mechanisms may inadvertently shape users’ perceptions of public issues. This paper highlights both the potential and limitations of LLMs in social measurement and recommends enhancing data diversity, improving model interpretability, and developing domain-specific models tailored for the social sciences.

Key words: large language models (LLMs), algorithmic fidelity, data bias, alignment mechanism, Chain-of-Thought (CoT)