基于大语言模型的国际信任民调数据插补*

杨　锋　侯煜欣　庞　珣

【内容提要】 近年来，预训练大语言模型的快速发展催生了区别于传统数据处理路径的“生成性建模”(generative modeling)方法，在社会科学研究中展现出广泛的创新潜力。为探索大语言模型赋能国际关系实证研究和评估其能力和可信度，本文以大语言模型进行民意调查中缺失值插补为切入点，聚焦于其中国际信任度这一高度依赖上下文的主观潜在变量，在这一具有挑战性的缺失数据处理任务上，将大模型的表现与常用且强大的现有方法进行系统性比较。本文借助“中国家庭追踪调查”(China Family Panel Studies，简称 CFPS)这一纵贯性调查数据包含的丰富上下文信息，以及该调查中对国内外不同对象信任度问题所提供的比较契机，考察和对比在不同缺失机制下不同方法的能力表现，以及本土和国际大模型预测能力的优劣，以评估模型训练语料与人类反馈训练中隐含的社会语境对国内外态度预测的影响。实验结果发现，尽管生成性方法在主观民意数据插补中仍面临挑战，但在应对非随机缺失方面表现优异，其性能不逊于现有机器学习方法，且在数据保真方面能力突出。同时，本土模型在理解和模拟本地政治社会语境方面展现的能力胜过国际大模型。鉴于非随机缺失数据在国际关系研究中的常见性和挑战性，本文的发现有助于研究者应用大语言模型来应对这一重要数据问题，并展示了提升大语言模型在国际关系研究中的可信度与规范性的路径。

【关键词】 国际信任度　民调　缺失数据　大语言模型　生成性建模

---

*　本文作者感谢亚洲政治学方法年会 2025 年年会(新西兰惠灵顿)和第三届计算社会科学国际会议(中国香港)参与者给予的宝贵意见和建议，以及北京大学 2024 年度“数字与人文”专项课题的资助。本文通讯作者为庞珣，请将反馈发送至 xpang@pku.edu.cn。

【作者简介】 杨锋,北京大学光华管理学院社会研究中心助理教授。
电子邮箱:feng-yang@pku.edu.cn
侯煜欣,北京大学光华管理学院社会研究中心博士研究生。
电子邮箱:houyx21@stu.pku.edu.cn
庞珣,北京大学国际关系学院教授。
电子邮箱:xpang@pku.edu.cn

一、引言

生成式人工智能正在快速而深刻地重塑社会科学研究,带来广阔的创新空间,但也对现有学术规范和伦理造成冲击。这一影响的一个重要面向是,人工智能为研究者提供了重要工具,形成生成性建模(generative modeling)路径下的"生成性方法",可以前所未有的方式和易得性"创制"数据来为社会科学提供实证补充和增广,但这些数据在可靠性、可解释性和适用性等方面都充满争议,尚待探索。鉴于其巨大的潜在价值和无法回避的规范挑战,研究者正在积极展开基于应用的评估,将大语言模型(LLM)等深度生成性模型产生的文本表征或合成文本运用于社会科学研究中,进行因果纠偏、社会模拟和缺失值填补等任务测试。①

民意调查是实验和运用生成性方法最活跃的领域之一。一方面,民意

① Joon Sung Park et al., "Generative Agents: Interactive Simulacra of Human Behavior," *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, 2023, pp. 1-22; Ruoxi Xu et al., "AI for Social Science and Social Science of AI: A Survey," *Information Processing & Management*, Vol. 61, No. 3, 2024, 103665; Kosuke Imai and Kentaro Nakamura, "Causal Representation Learning with Generative Artificial Intelligence: Application to Texts as Treatments," arXiv Preprint, 2024, https://arxiv.org/abs/2410.00903, 访问时间:2024 年 12 月 30 日; Lincan Li et al., "Political-LLM: Large Language Models in Political Science," arXiv Preprint, 2024, https://arxiv.org/abs/2412.06864, 访问时间:2024 年 12 月 30 日; 庞珣:《人工智能赋能社会科学研究探析——生成式行动者、复杂因果分析与人机科研协同》,载《世界经济与政治》,2024 年第 7 期,第 3—30 页;杰夫·吉尔,等:《国际关系研究的人工智能"方法论"》,载《世界经济与政治》,2025 年第 1 期,第 4—26 页。

调查因成本高昂和低回应率而备受数据缺失问题的困扰。[①] 另一方面，民调包含多方面的多样性问题，能够提供缺失值填补所依赖的丰富的上下文。而大语言模型对上下文本超强的敏感性和对回答进行生成的独特能力，与民调缺失值插补任务相当适配。

普遍而言，对主观态度的推断比对客观事实的猜测更依赖于上下文。由于缺乏直接观测途径，社会科学中对主观态度的实证测量主要来自于民调。大数据兴起后，文本分析——尤其是基于社交媒体语料的分析——是另一个关于主观态度的重要数据来源，但民调仍被用作主观态度测量的基准。然而，在民调中，受访者常因隐私顾虑或认知困难而拒绝回答，因为主观态度来自更深层的认知结构。[②]

现有研究肯定了大语言模型在数据插补方面的巨大潜力，它们通过海量文本数据学习，能有效捕捉民调中主观态度植根的潜在模式。[③] 然而，现有的评估大都以国外民意调查为基准数据，评估国外主流大模型，缺乏对主观态度所处的不同认知框架结构和政治社会背景的考量，也缺乏对大模型的社会文化适应性进行考察，存在系统性比较不足、泛化能力有限的缺陷。[④]

本文采用中国大型调查——“中国家庭追踪调查”(CFPS)数据，应用大模型对其中中国公民信任度问题的回答进行缺失值插补，评估其能力和可靠性，包括：(1)不同缺失机制下大语言模型与传统方法的比较优势；(2)纵贯数据对模型表现的影响；(3)国内外模型差异是否体现“主权 AI”效应。

---

① Lisa Argyle et al.，“Out of One，Many：Using Language Models to Simulate Human Samples，” *Political Analysis*，Vol. 31，No. 3，2023，pp. 337-351；Junsol Kim and Byungkyu Lee，“AI-Augmented Surveys：Leveraging Large Language Models and Surveys for Opinion Prediction，” arXiv Preprint，2024，https://arxiv.org/abs/2305.09620，访问时间：2024 年 10 月 30 日；Junyung Ji，Jiwoo Kim and Younghoon Kim，“Predicting Missing Values in Survey Data Using Prompt Engineering for Addressing Item Non-response，” *Future Internet*，Vol. 16，No. 10，2024：Article 351，DOI：10.3390/fi16100351。

② John Zaller，*The Nature and Origins of Mass Opinion*，Cambridge：Cambridge University Press，1992.

③ Argyle et al.，“Out of One，Many，” pp. 337-351.

④ James Bisbee，Joshua D. Clinton，Cassy Dorff，Brenton Kenkel and Jennifer M. Larson，“Synthetic Replacements for Human Survey Data? The Perils of Large Language Models，” *Political Analysis*，Vol. 32，No. 4，2024，pp. 401-416.

本文采取多种方法针对多种缺失场景和上下文输入进行实验对比。实验设计具有以下核心考量。首先,我们基于唐纳德·鲁宾提出的经典框架,模拟三种典型的数据缺失机制:完全随机缺失(MCAR)、有条件随机缺失(MAR)和非随机缺失(MNAR),[①]对比大语言模型与现有典型插补方法的表现差异,考察这些方法对复杂缺失模式的识别能力和对潜在数据分布的捕捉能力。

其次,针对实际研究中常见的数据可及性问题,我们设计了不同数据可得性场景下的对比实验。考量到在处理一些民调数据时,研究者需要进行跨期"回溯性插补"却缺乏相应训练数据的情况,[②]我们评估了大语言模型在零样本(zero-shot)和小样本(few-shot)学习设置下的表现,探讨其在实际研究场景中的应用潜力。

最后,从"主权 AI"的视角出发,实验对比了国际模型(GPT-4o-mini)和国产模型(Qwen-plus)在中国社会文化语境下的表现差异。本文主要的数据生成与分析工作开展于 2024 年 9—12 月,在此期间,上述两类模型均属当时较为流行的模型。实验验证了国产模型在捕捉中国社会文化特征方面的优势,提示研究者应根据具体需求选择合适模型。

实验结果表明,大语言模型在不同缺失机制下的表现存在差异:在有条件随机缺失(MAR)情境中,传统机器学习方法更具优势;而在非随机缺失(MNAR)情况下,零样本学习的大语言模型展现出接近甚至超越统计方法与机器学习方法的潜力,尤其在训练数据受限、传统方法难以发挥效用时,其优势将更加明显。值得注意的是,国产模型(如 Qwen-plus)在相对更具社会敏感性的指标(如地方官员信任度)上的插补表现优于国外模型,凸显其对本土社会语境的适应能力。

本文的发现有助于研究者应用大语言模型来应对这缺失值插补这一重要数据问题,并展示了在大语言模型国际关系研究中提升可信度与规范性的路径。

---

① Donald B. Rubin, "Inference and Missing Data," *Biometrika*, Vol. 63, No. 3, 1976, pp. 581-592.

② 关于"回溯性插补",参见 Kim and Lee, "AI-Augmented Surveys," arXiv: 2305.09620v3.

二、信任度民意调查：缺失数据问题及处理方法

（一）信任度的测量困难

信任在人类社会中的重要性不言而喻。在国际关系中，信任尤其困难但也更为关键。在无政府状态下，国家间的合作达成和冲突解决缺乏中央权威的安排和约束，也没有具有强制力的第三方执行机构，因此信任缺乏稳固的基础。但同时，对信任的需要又渗透在国际关系互动的方方面面。

信任通常被视为一种关系性（relational）和情境特定（domain-specific）的心理状态，表现为一方（A）相信另一方（B）在特定领域（X）中会以正直和能力行事，并优先考虑 A 的利益。[①] 人与人之间、群体之间以及国家之间的信任，是促成合作与维持秩序的重要机制，尤其在正式制度薄弱或执行成本较高的情境中，其作用更加突出。

正因如此，信任长期以来受到多个学科的广泛关注。在政治学中，研究者重点探讨公众对政府及其官员的信任程度、信任的结构性差异及其政治影响。[②] 国际关系研究则关注一国民众对其他国家的信任、或对国际组织的信任[③]，并以此解释国际合作的基础。此外，公共卫生、社会学、心理学与经济学等领域也分别从社会资本、健康行为、风险感知与制度运行等角度研究

---

① Margaret Levi and Laura Stoker, "Political Trust and Trustworthiness," *Annual Review of Political Science*, Vol. 3, No. 1, 2000, pp. 475-507; Jack Citrin and Laura Stoker, "Political Trust in A Cynical Age," *Annual Review of Political Science*, Vol. 21, No. 1, 2018, pp. 49-70.

② Levi and Stoker, "Political Trust and Trustworthiness," pp. 475-507; Citrin and Stoker, "Political Trust in A Cynical Age," pp. 49-70.

③ Paul R. Brewer et al., "International Trust and Public Opinion about World Affairs," *American Journal of Political Science*, Vol. 48, No. 1, 2004, pp. 93-109; Xiaojun Li, Jianwei Wang and Dingding Chen, "Chinese Citizens' Trust in Japan and South Korea: Findings From A Four-City Survey," *International Studies Quarterly*, Vol. 60, No. 4, 2016, pp. 778-789; Benno Torgler, "Trust in International Organizations: An Empirical Investigation Focusing on the United Nations," *The Review of International Organizations*, Vol. 3, No. 1, 2008, pp. 65-93.

信任问题。①

国际关系理论中,信任曾被认为是国家理性考量的选择。但随着现实中国内政治与国际政治之间跨层次关联变得更为复杂紧密,国际关系研究中的信任概念也越来越嵌入到复杂关系网络和潜在的认知结构中,以及更为直接地植根于微观层面的公民态度。② 与此同时,全球性的"信任危机"被认为不仅是国内政治不稳和极化的重要因素,而且是地缘政治风险上升的推动因素和重要后果。现有研究发现,公民的国内政治信任度以及普遍信任度都与其对"他者"的信任度密切相关,凸显了信任问题的复杂性和在政治社会文化等维度上的联动性,也使得信任度的实证度量更为重要且更具挑战。③

随着国际关系研究对信任度的重视,国际国内的众多大型民意调查中都将关于信任的问题置于重要位置并进行持续调查。然而,信任具有明显的语义多义性和解释模糊性,不同受访者对"你是否信任某些人/某机构"这一提问可能存在截然不同的理解。受访者人可能将信任理解为情感态度(如亲切感),另一些人则侧重行为倾向(如是否愿意与对方合作),还有人将其视为对制度或规则的评价性判断(如认为制度是否公平或有效)。即使是在评价性判断的框架下,不同受访者理解中"能力"与"意图"两个维度的重

① Oliver Schilke, Martin Reimann and Karen S. Cook, "Trust in Social Relations," *Annual Review of Sociology*, Vol. 47, No. 1, 2021, pp. 239-259; Jennifer Richmond et al., "Conceptualizing and Measuring Trust, Mistrust, and Distrust: Implications for Advancing Health Equity and Building Trustworthiness," *Annual Review of Public Health*, Vol. 45, 2024, pp. 465-484.

② Anna Swärd, "Trust, Reciprocity, and Actions: The Development of Trust in Temporary Inter-organizational Relations," *Organization Studies*, Vol. 37, No. 12, 2016, pp. 1841-1860; O. Pesämaa, Torsten Pieper, Rui Vinhas da Silva, W. C. Black, J. F. Hair Jr, "Trust and Reciprocity in Building Inter-Personal and Inter-Organizational Commitment in Small Business Co-Operatives," *Journal of Co-operative Organization and Management*, Vol. 1, No. 3, 2013, pp. 81-92.

③ 参见 Eric M Uslaner, *The Moral Foundations of Trust*. Cambridge University Press, 2002 以及民意调查:Pew Research Center, "Public Trust in Government," 2022, https://www.pewresearch.org/politics/2022/06/06/public-trust-in-government-2/, 访问时间:2024 年 10 月 15 日; Steven W. Webster, "Anger and Declining Trust in Government in the American Electorate," *Political Behavior*, Vol. 40, No. 2, 2018, pp. 933-964.

要性也可能大相径庭。[①] 这类理解差异不仅导致测量等价性(measurement invariance)难以成立,也使得部分受访者因不理解问题含义而回避作答,从而使信任变量呈现出超出可观测变量所能解释的系统性缺失异质性。

此外,部分信任问题具有较强的政治或社会敏感性,容易引发社会期望偏差(social desirability bias),使得受访者出于自我保护或社会考虑而回避作答。例如,在评价地方政府官员时,受访者可能担心表达不满会"驳了面子",引发不必要的注意或影响熟人社会中的人际关系;而在评价对美国人的信任时,也可能因担忧被视为"崇洋媚外"而选择回避。这类机制使得信任变量的缺失不仅频繁发生,而且很可能与其真实取值高度相关,带来后续描述性和解释性分析的偏差。

可见,民调中信任回答的缺失在社会科学研究的数据缺失问题中具有典型的代表性:缺失的原因可能具有个体内部的主观原因,也可能受到社会结构和氛围的影响。这让缺失数据处理成为一个重要的方法论领域——不可轻易忽视的缺失值需要进行数据插补,而数据插补必须遵循科学原理和依赖可靠技术。

(二)缺失值处理的既有方法

如何处理缺失的民意数据?简单地排除缺失数据的观察不仅因减少样本量而增加标准误差,还可能带来统计推论的偏误。缺失数据的严重性和插补处理的复杂性,不仅在于缺失值的多寡,更在于潜在的缺失数据机制。[②] 不同的缺失机制需要的插补难度不一、要求的方法也各异。

数据缺失机制主要包括三种:完全随机缺失(missing completely at random,简称 MCAR)、有条件随机缺失(missing at random,简称 MAR)和非随机缺失(missing not at random,简称 MNAR)。在"完全随机缺失"情形下,缺失的概率与观察到的或未观察到的数据无关,这意味着即使删除缺失值的观察也不会引入偏差,删去具有缺失值的观察量带来的仅是样本

① Li Lianjiang,"Reassessing Trust in the Central Government: Evidence from Five National Surveys," *The China Quarterly*, Vol. 225, 2016, pp. 100-121.

② Rubin, "Inference and Missing Data," pp. 581-592.

量的损失,在缺失值比例不高、或者样本量很大的情况下,可以采取直接删去的方式而不影响描述和推论的结果。然而,这种情况在实践中很少见,尤其在社会科学中,缺失值的产生往往有其系统性原因。

"有条件随机缺失"放松了对随机性假定,是指当数据缺失值与其他观察到的变量取值相关的情况。[①] 比如,受访者是否回答关于信任的问题,取决于其受教育的程度,受教育程度高的受访者可能因对"信任"概念反复思索而举棋不定,最后放弃回答这些问题。但是,受教育程度一般都会包含在民调中,而没有回答信任问题的受访者大都愿意报告自己受到的良好教育。于是,在 MAR 的缺失机制下,我们就可以用教育程度来"预测"信任程度的回答,在缺失值处填补上预测值。基于这样的原理发展出了众多缺失值插补的方法,例如用于多重插补的 Amelia 系列软件包。[②]

然而,有条件随机缺失要求缺失值与任何未被观察到的因素之间无关。但在社会意愿偏差的影响下,受访者可能隐瞒某种特定的答案,导致数据缺失模式与真实的潜在值有紧密关联,而潜在值不可观察。数据"非随机缺失"是研究者在实际研究中常常面临的困境,其关键在于缺失机制无法通过可观察变量来预测。为应对这一问题,研究人员需要借助直接可测变量以外的信息刻画缺失机制,比如对数据的潜在结构进行建模,并利用估计得到的潜在因子辅助缺失值插补。比如矩阵分解和无监督机器学习等技术,通过挖掘观测数据中的潜在模式以推断未观测结构,往往在高维数据的插补任务中有良好表现。[③]

---

① Rubin, "Inference and Missing Data," pp. 581-592.

② 例如 James Honaker and Gary King, "What To Do About Missing Values in Time-series Cross-section Data," *American Journal of Political Science*, Vol. 54, No. 3, 2010, pp. 561-581; James Honaker, Gary King and Matthew Blackwell, "Amelia Ⅱ: A Program for Missing Data," *Journal of Statistical Software*, Vol. 45, No. 7, 2011, pp. 1-47.

③ Nandana Sengupta et al., "Sparse Data Reconstruction, Missing Value and Multiple Imputation Through Matrix Factorization," *Sociological Methodology*, Vol. 53, No. 1, 2023, pp. 72-114; Naijia Liu, "A Latent Factor Approach to Missing Not at Random," Working Paper, 2021, available at https://naijialiu.github.io/pics/LFA_21.pdf,访问时间:2024 年 12 月 30 日.

（三）大语言模型对民调缺失值插补的潜力

现有研究高度关注对大语言模型进行缺失数据插补的潜力，这主要基于大语言模型的两大核心能力：其一是大语言模型强大的自然语言处理能力，尤其是其对语境的理解与推理；其二是其对人类回答模式与行为逻辑的模拟能力。

相较于传统插补方法主要依赖结构化的数值数据，社会调查或民意调查中的问题与回答往往包含丰富的语义信息。这些回答并不仅仅是数字或简单分类，而是嵌入了语境、情感和主观理解。例如，在面对“您在多大程度上信任美国人”这类问题时，受访者可能以“非常不信任”作答，而该回答再由访问员转录成数值。这一转化过程不可避免地压缩了原始语言信息的维度，使得传统基于表格数据的插补方法难以捕捉其中的语义细节，影响预测和插补的效果。①

大语言模型具有克服这一问题的能力，因为它对语言输入构建“内部表征”（internal representation）。简单地说，模型通过将词汇、语境及其语义关系编码为高维向量，使其能够在表层数据缺失的情况下，通过对潜在语义结构的理解进行合理推断。这种表征方式不仅提升了模型对语言内容的感知能力，也使其能够捕捉变量间深层的语义关联和潜在结构，从而实现比传统方法更为精细和一致的数据插补。②

将大语言模型应用于缺失数据问题的现有尝试和测试产生了关于这一能力的实证证据。例如，阿哈特沙姆·哈亚特（Ahatsham Hayat）与穆罕默德·R. 哈桑（Mohammad R. Hasan）利用大语言模型生成缺失数据的描述性文本或提示信息，并将其输入到较小、经过微调的模型中以完成预测任务。③ 王建伟等人的研究表明，大语言模型在处理混合类型数据（即同时包含数值型、分

---

① Junyung Ji, Jiwoo Kim and Younghoon Kim, “Predicting Missing Values in Survey Data Using Prompt Engineering for Addressing Item Non-response,” Article 351.

② Kim and Lee, “AI-Augmented Surveys,” arXiv: 2305.09620v3.

③ Ahatsham Hayat and Mohammad R. Hasan, “A Context-Aware Approach for Enhancing Data Imputation with Pre-trained Language Models,” *Proceedings of the 31st International Conference on Computational Linguistics*, 2025, pp. 5668-5685.

类型与文本型变量的数据)方面具有独特优势,有助于提升插补的灵活性与准确性。[①] 池俊永(Junyung Ji)等人发现,通过设计合理的提示词(prompt engineering),可以筛选出与目标缺失项最相关的受访者与问题,输入给大语言模型,由其利用上下文学习与推理能力进行插补。[②] 这种方式无需繁复的数据预处理或大量训练,显著简化了分析流程,同时提升了插补的效率与准确度。

除了通过文本表征来捕捉上下文以发现潜在数据结构以预测缺失值外,大语言模型可以模拟人类的文本生成,让其在一定程度上可被视为"合成受访者",通过角色扮演的方式对个体受访者没有回答的问题加以补充。这不仅借助了大语言模型对给定文本的上下文理解能力,更是来自于大模型的"内部知识"——通过海量语料训练的模型能够超越给定文本的知识和提示来推测人的知识结构和信念模式,从而推断出特定个体在特定场景下的可能回答。

现有研究对大语言模型作为"合成受访者"对于民调数据的增广进行了系统评估。比如,莉莎·阿格尔(Lisa Argyle)等人运用 2012 年、2016 年和 2020 年美国国家选举研究(ANES)数据,将受访者的关键人口统计特征(如种族、党派归属等)输入 GPT-3,令其给出受访者关于其投票倾向性的回答。结果表明,模型生成的合成数据与实际调查数据之间高度一致,显示出大语言模型能够在给定受访者背景条件下,合理模拟其对具体问题的回答。[③] 尽管部分研究指出,大语言模型可能存在对特定国家或社会群体的偏见[④],但莉莎·阿格尔等人认为,只要输入信息准确匹配目标群体,大语言模型依然能够有效复现其观点分布,展示出对群体差异的高度敏感性。

---

① Jianwei Wang et al., "On LLM-Enhanced Mixed-Type Data Imputation with High-order Message Passing," arXiv Preprint, 2025, https://arxiv.org/abs/2501.02191, 访问时间:2024 年 12 月 15 日。

② Junyung Ji, Jiwoo Kim and Younghoon Kim, "Predicting Missing Values in Survey Data Using Prompt Engineering for Addressing Item Non-response," Article 351.

③ Argyle et al., "Out of One, Many," pp. 337-351.

④ 例如 Fabio Motoki, Valdemar Pinho Neto and Victor Rodrigues, "More Human than Human: Measuring ChatGPT Political Bias," *Public Choice*, Vol. 198, No. 1, 2024, pp. 3-23; Esin Durmus et al., "Towards Measuring the Representation of Subjective Global Opinions in Language Models," arXiv Preprint, 2024, https://arxiv.org/abs/2306.16388, 访问时间:2024 年 11 月 5 日。

此外，金俊硕（Junsol Kim）和李炳圭（Byungkyu Lee）进一步整合多项全美代表性调查数据，尝试突破传统插补方法对变量“表层相似性”的依赖，转而更加充分地利用问卷问题之间的语义相似性作为建模依据。[①] 通过对大语言模型的微调，他们显著提升了对个体态度预测的准确性，并优于如矩阵分解（Matrix Factorization）等前沿机器学习方法。然而，并非所有研究都呈现出一致的乐观结果。詹姆斯·比斯比（James Bisbee）等人指出，ChatGPT 生成的合成数据在还原调查变量之间的真实相关结构（如年龄与政治态度之间的关系）方面存在不足，导致模型无法有效再现社会群体间的意见差异。[②]阿尔诺·帕舒（Arnault Pachot）和蒂埃里·佩蒂（Thierry Petit）对该领域的相关研究进行了系统综述，评估了各类方法（无论是仅基于人口学变量进行插补，还是同时引入意识形态信息等）的有效性与偏误风险，揭示了大语言模型模拟民意的潜力与局限。[③]

总体而言，尽管仍面临诸多挑战，已有研究已清晰展示出大语言模型在生成高质量合成数据以插补社会调查或民意调查数据的潜力。这种潜力主要体现在两个方面：一是在缺乏足够训练数据的情况下，传统插补方法因可用变量有限而难以实施，而大语言模型凭借其零样本学习的能力，能够基于提示生成“可能的回答”；二是在训练数据可得的情形下，大语言模型可作为传统插补方法的补充，与其集成为混合模型，从而进一步提升预测的精度与可靠性。

然而，当前的相关研究缺乏对信任度这样高度依赖上下文和潜在结构的民调问题的数据缺失进行专门性评估。测试和评估的基准数据也以美国民调为主，而缺乏对大语言模型在非西方民意推断上能力的关注。此外，不同国家的大语言模型可能因预训练语料和人类反馈训练使用的语料不同而对本国和外国语境下的民调问题的理解和回答能力不同，但这一点未得到足够重视和系统分析。

---

① Kim and Lee，“AI-Augmented Surveys，” arXiv：2305.09620v3.

② Bisbee et al.，“Synthetic Replacements for Human Survey Data?” pp.401-416.

③ Arnault Pachot and Thierry Petit，“Can Large Language Models Accurately Predict Public Ppinion? A Review，” HAL，2024，hal-04688498，https://hal.science/hal-04688498v1，访问时间：2024 年 12 月 15 日。

三、评估大语言模型的缺失值插补:实验设计

本研究旨在评估大语言模型在模拟和插补个体民意方面的表现。个体民意(比如对某一项具体政策的反馈)通常是社会科学研究重点关注的变量,既可能是分析的自变量也可能是因变量。在一些具体情况下,研究者会对这些个体民意进行加总,从而了解某个具体时点的公众民意,并长期关注和监督公众民意的变化。虽然政治学方法前沿已经开始探索大语言模型在生成合成数据的表现,体现了其在数据插补上的潜力,但目前学界对其具体表现以及与其他模型的比较缺乏了解。本文将这一比较作为重点,将大语言模型插补结果与传统方法(如 OLS 回归分析和机器学习工具)进行系统比较。

(一)基准数据

本研究以中国家庭追踪调查(CFPS)为基准数据。CFPS 调查由北京大学组织实施,是少有的具有全国代表性的高质量纵向社会调查。[①] 在 2010 年,其基线调查成功访问了 14960 户家庭的 33600 名成年人,样本覆盖 25 个省。此后,调查采用每两年一次的追踪模式,持续跟进核心家庭成员及其后代。本文的基准数据取自 2018 年和 2020 年两次调查。[②]

由于其纵向调查设计,CFPS 调查得以持续追踪受访者对重要社会问题的认知变化,以及他们对地方政府官员、医生、外国人(如美国人)等不同群体的态度演变。这些民意测量数据已成为国内及国际信任研究的重要资源。例如,最新研究探讨了中美贸易摩擦等重大国际事件如何影响中国民众对美国人信任

① Yu Xie and Jingwei Hu, "An Introduction to The China Family Panel Studies (CFPS)," *Chinese Sociological Review*, Vol. 47, No. 1, 2014, pp. 3-29.

② CFPS 调查采用面访为主、电访和网络为辅的混合模式,保持了较高应答率。基线调查个体横截面应答率达 84.1%,相邻波次追访成功率(排除已故者)分别为:2012 年 80.6%、2014 年 83.8%、2018 年 80.8%、2020 年 77.0%。2020 年成功率略有下降,主要因新冠疫情导致数据收集转为以电话访谈为主。关于对近期数据质量的讨论,参考:Yu Xie et al., "Declining Chinese attitudes toward the United States amid COVID-19," *Proceedings of the National Academy of Sciences of the United States of America*, Vol. 121, No. 21, 2024, e2322920121, DOI: 10.1073/pnas.2322920121.

程度的动态变化，[①]以及反腐败如何可以提升民众对地方政府官员的信任。[②]

之所以选择 CFPS 数据，是因为 CFPS 数据结构为检验大语言模型在处理主观态度类数据缺失问题中的效能提供了理想基础。一方面，作为全国性大型追踪调查，CFPS 系统采集了个体在多个年份的多维度信息，包括人口学特征、对相关社会问题的感知以及历史回答记录，能为模型识别缺失背后的潜在结构提供丰富的上下文信息。另一方面，其纵向设计使研究者可以利用个体在过往年份中对相同或相关问题的回答来辅助插补当前的缺失数据。例如，在插补 2020 年“对地方官员信任度”的缺失值时，可以参考同一受访者在 2018 年的相关回答，从而提升插补的准确性与个体连贯性。

此外，CFPS 中关于信任问题提供了与国际态度、国内政治及公共服务体系相关的三种信任对象，有助于从多个角度检验大语言模型对社会语境的适应能力。文本重点关注与之相关的三道问题，即对美国人、本地政府官员以及医生的信任度。具体问题如下：

若 0 分代表非常不信任，10 分代表非常信任，请您对以下这几类人的信任程度打分：

（1）对美国人的信任度

（2）对本地政府官员的信任度（“本地政府官员”指当地县/县级市/区政府官员）

（3）对医生的信任度

这三类信任变量在样本中表现出高度的个体差异性。在 2020 年 CFPS 数据（$N=28530$）中，受访者对美国人、本地政府官员、医生的信任标准差分别为 2.41、2.55 和 2.25。[③] 此外，这三类问题的社会和政治敏感性较高，因此缺失值较多，尤其对美国人和政府官员的信任两个问题，缺失比例分别为

---

① Xie et al.，“Declining Chinese attitudes toward the United States amid COVID-19，” e2322920121.

② Siqin Kang and Jiangnan Zhu，“Do People Trust The Government More? Unpacking The Distinct Impacts of Anticorruption Policies on Political Trust，” *Political Research Quarterly*，Vol. 74，No. 2，2021，pp. 434-449.

③ 样本中受访者对美国人、医生、地方官员、父母、邻居及陌生人信任程度（0～10 分量表）的均值分别为：2.01、7.23、5.85、9.33、6.68 和 2.39。

14.56%和 13.59%。考虑到这些信任问题对政治与文化语境的敏感性,比较国产模型(如 Qwen-plus)与国外模型的插补表现能有效地揭示不同模型在处理具有争议性或文化依赖性问题上的能力差异,从而为理解大语言模型在语境敏感型任务中的适用性提供更好的实证检验。

(二)评测指标

基于 CFPS 数据,特别是 2020 年的调查数据,我们模拟了不同的数据缺失机制来人为"掩盖"其中有关信任问题的部分回答,然后应用大语言模型和其他方法对掩盖的"缺失数据"进行插补,通过对比插补数据和实际数据之间的差距评估不同方法的表现。评估和对比的内容主要包括三个方面:对受访者个体层面回答插补的准确性、对样本层面数据分布特征的刻画能力以及插补后对信任度和其他变量之间关系进行捕捉的准确度。①

首先,个体插补准确性衡量的是模型在恢复缺失的个体回答方面的平均预测精度,其衡量指标为均方根误差(Root Mean Squared Error,RMSE),计算公式为 $RMSE=\sqrt{\frac{\Sigma(\hat{y}-y)^2}{N}}$,即插补值($\hat{y}$)与真实值($y$)之差的平方的平均数再开平方,描述的是"点对点"的个体插补准确性。RMSE 数值越小表示预测越接近真实值,模型拟合效果越好。

其次,除了均方根误差这一数值指标,研究者还关注插补数据是否能够有效保留样本的分布特征。通过检视插补值的分布并将其与真实值的分布进行对比,不仅可以判断插补是否准确捕捉了真实值的均值(即中心趋势),更关键的是,可以进一步评估其在分布形态、变异性等方面是否与真实数据保持一致,即所谓的分布相似性(distributional similarity)。如果插补值的分布变异性明显不足,呈现异常的集中趋势,可能表明该插补方法在重建数据分布或模拟缺失机制方面存在系统性偏差,难以还原样本的真实结构。这种对分布形态的比较可作为均方根误差指标的有力补充,有助于更全面

① Maria Thurow et al., "Goodness (of Fit) of Imputation Accuracy: The GoodImpact Analysis," arXiv Preprint, 2021, https://arxiv.org/abs/2101.07532, 访问时间:2024 年 11 月 5 日; Bisbee et al., "Synthetic Replacements for Human Survey Data?" pp. 401-416.

地评估插补结果的合理性与保真度。

最后，变量间的相关性结构是衡量插补质量的另一关键维度。由于插补数据通常用于后续的统计建模和因果推断等下游分析，高质量的插补应尽可能保留原始数据中变量之间的真实关联。例如，我们比较目标变量的插补值（$\hat{y}$）与其他无需插补变量（z）之间的相关系数 $\mathrm{cor}(\hat{y},z)$，以及目标变量真实值（y）与这些变量之间的相关系数 $\mathrm{cor}(y,z)$，以评估插补结果对相关结构的保留程度。若两者差异较大，甚至方向相反，说明插补未能准确反映变量之间的实际关系，可能会削弱插补数据在下游分析中的解释力与可靠性。

（三）模型选择

对于大语言模型的选择，本文考虑 OpenAI 公司的 GPT 系列模型和阿里巴巴的 Qwen（通义千问）系列模型。在本研究主要开展周期内（2024 年 9—12 月），GPT-4 是在当时测试中最常见的模型选择之一[①]，在多种任务中都表现出了良好的零样本（zero-shot）和少样本（few-shot）学习能力，是国外模型的代表。[②] Qwen 系列支持中英双语的模型，既提供开源版本，也有闭源版本，涵盖多种参数规模和模型结构。[③] 其 2.5 版本在预训练阶段使用了高达 18 万亿（18T）个 tokens 的语料，在多个基准测试上表现显著优异，已达到与其他领先大语言模型相当的水平。[④] 我们用 Qwen 来作为与国际大模型相比较的本土模型的代表。

---

① Shervin Minaee et al.，“Large Language Models：A Survey，” arXiv Preprint，2024，https://arxiv.org/abs/2402.06196，访问时间：2024 年 12 月 30 日；Josh Achiam et al.，“GPT-4 Technical Report，” arXiv Preprint，2023，https://arxiv.org/abs/2303.08774，访问时间：2024 年 12 月 30 日；Wayne Xin Zhao et al.，“A Survey of Large Language Models，” arXiv Preprint，2023，https://arxiv.org/abs/2303.18223，访问时间：2024 年 12 月 30 日。

② Patrick Y. Wu et al.，“Large Language Models Can Be Used to Estimate the Latent Positions of Politicians，” arXiv Preprint，2023，https://arxiv.org/abs/2303.12057，访问时间：2024 年 12 月 30 日；Caleb Ziems et al.，“Can Large Language Models Transform Computational Social Science?” *Computational Linguistics*，Vol. 50，No. 1，2024，pp. 237-291.

③ Qwen Team，“Qwen2 Technical Report，” arXiv Preprint，2024，https://arxiv.org/abs/2407.10671，访问时间：2024 年 12 月 30 日。

④ Ibid.

表 1 中方法类别Ⅳ报告了模型版本。OpenAI 提供了多个版本的 GPT-4 系列模型,包括 GPT-4o、GPT-4o-mini、GPT-4 Turbo、GPT-4 以及各种多模态版本。[①]为平衡性能与成本,我们选取了"GPT-4o-mini-2024-07-18"模型(简称"GPT-4o-mini")进行实验。[②]由于 GPT 系列模型为闭源,我们通过 API 调用 GPT-4o-mini,这种方式比本地部署更为便捷和低成本,无需大量计算资源。在通义千问系列中,我们选择了"Qwen-plus-2024-12-20"模型,[③]也通过 API 访问。

表 1 评估模型列表

方法类别	模型
Ⅰ. 传统统计方法	普通最小二乘回归(OLS)
	多重插补(Amelia Ⅱ)
Ⅱ. 机器学习	矩阵分解(Matrix Factorization)
	轻量级梯度提升机模型(LightGBM)
Ⅲ. 双向编码器表示模型(BERT)	Chinese-RoBERTa(微调)
Ⅳ. 生成式大语言模型(LLM)	Qwen-plus(零样本学习)
	Qwen-plus(小样本学习)
	GPT-4o-mini(零样本学习)
	GPT-4o-mini(小样本学习)

资料来源:作者自制。

为了评估大语言模型在缺失数据填补任务中的表现,本文将其结果与其他主流缺失数据插补方法(即表 1 中列出的方法类别Ⅰ、Ⅱ、Ⅲ)进行比较。

在传统统计方法上(方法类别Ⅰ),我们重点考察多重插补。多重插补(Multiple Imputation, MI)是社会科学研究中最常用的插补方法之一,本文进行插补时使用 R 软件包 Amelia Ⅱ。[④]该软件包基于有条件随机数据缺失

① OpenAI, "Models," OpenAI Platform Documentation, 2024, https://platform.openai.com/docs/models,访问时间:2024 年 12 月 30 日。

② OpenAI, "GPT-4o Mini: Advancing Cost-Efficient Intelligence," 2024, https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/,访问时间:2024 年 12 月 30 日。

③ 阿里云:模型广场(模型列表),阿里云官方文档,2024,https://help.aliyun.com/zh/model-studio/getting-started/models,访问时间:2024 年 12 月 30 日。

④ Honaker and King, "What To Do About Missing Values in Time-series Cross-section Data," pp. 561-581.

假设，采用期望最大化（EM）算法与自助法（bootstrapping）结合的方式，通过观测变量推测缺失值，生成多个可能的数据集。这些数据集包含不变的观测值以及变化的插补值，研究者可分别分析各个数据集，并结合结果考虑插补数据的不确定性。在本文的实证分析中，我们采取了一种混合样本策略：从多个插补数据集中各随机抽取一部分观测值，合成为一个新的分析样本。[①]此外，作为对比基准，我们也纳入了基于普通最小二乘回归（OLS）模型预测值的简单插补方法，以体现多重插补相较于更基础方法的潜在优势。

机器学习模型（方法类别Ⅱ）在数据填补任务中也展现出了较大的潜力，尤其在处理复杂关系和稀疏数据时往往优于传统方法。[②] 矩阵分解（Matrix Factorization）技术，如奇异值分解（Singular Value Decomposition）和非负矩阵分解（Non-negative Matrix Factorization），能够在高维稀疏数据中提取潜在结构。[③] 矩阵分解方法通常被视为传统插补方法中的最佳选择，并在近期研究中频繁被用作评估大语言模型相对增益的基准模型。[④] 因此，我们在对比研究中纳入了矩阵分解方法。但是矩阵分解方法在捕捉非线性关系方面存在局限性，而轻量级梯度提升机模型（LightGBM）这种高效的树模型能够更好地捕捉非线性关系，并在数据分布上更具稳健性，因此我们也

---

① 传统多重插补会生成 M 个（比如，5 个）数据集，研究者在每一个插补数据集上分别进行模型估计，获得 M 个参数估计量与标准误，然后使用唐纳德·鲁宾提出的方法（即“鲁宾规则”）对这些结果进行加权合并，以估计总体的点估计和标准误。在本研究中，我们采用了一种替代性的整合方式：在每个插补数据集中随机抽取 N/M 个观察值（其中 N 为原始样本量），拼接为一个新的完整数据集，再将其当作一个数据集简化分析。该方法在一定程度上保持插补结果多样性的同时，显著简化了后续建模与结果整合过程，更便于与机器学习模型和大语言模型的插补结果进行直接比较，因为后两者通常仅生成一个完整的数据集而非多个备选版本。具体可参见 Honaker and King，“What To Do About Missing Values in Time-series Cross-section Data，” p. 564。

② Dimitris Bertsimas, Colin Pawlowski and Ying Daisy Zhuo, “From Predictive Methods to Missing Data Imputation: An Optimization Approach,” *Journal of Machine Learning Research*, Vol. 18, No. 196, 2018, pp. 1-39; Tlamelo Emmanuel et al., “A Survey on Missing Data in Machine Learning,” *Journal of Big Data*, Vol. 8, No. 1, 2021, pp. 1-37; Sengupta et al., “Sparse Data Reconstruction, Missing Value and Multiple Imputation Through Matrix Factorization,” pp. 72-114.

③ Sengupta et al., “Sparse Data Reconstruction, Missing Value and Multiple Imputation Through Matrix Factorization,” pp. 72-114.

④ Kim and Lee, “AI-Augmented Surveys,” arXiv: 2305.09620v3.

考虑将大语言模型的表现与之进行比对。①

因为大语言模型的缺失值插补方法本质上是“生成性方法”,因此我们也将其表现与非生成性预训练模型相比较,评估生成性在其中的作用。对于非生成性语言模型,我们考虑双向编码器表示模型(Bidirectional Encoder Representations from Transformers,简称 BERT),②这是最早且最具影响力的预训练语言模型之一,通过预训练有效地捕捉了通用知识(表 1 中方法类别Ⅲ)。虽然属于广义上的大语言模型,但 BERT 不具有生成性,而是用于输出分类、提取、打分等语言任务。另外,BERT 系列模型在规模上也远小于大语言模型,虽有上亿参数但仍远小于强大的生成式大语言模型。BERT 相对较小的参数规模使其在模型微调过程中能够更好地捕捉缺失值填补任务中的隐含数据模式。因此,我们微调了 BERT 系列模型中的 Chinese-RoBERTa-wwm-ext-large(以下简称 Chinese-RoBERTa)模型,用于对比大语言模型的表现。该模型在中文数据集上进一步预训练,以更好地适应中文特定任务。③

可见,文本选择用以和大模型比较的“传统”方法并非“过时”方法,相反,它们大都是目前常用或非常先进的缺失值插补方法,尤其是我们还增强了其中一些方法来进行比对。大语言模型的缺失值插补生成性方法的表现如果能与这些方法不相上下,即可以认为其表现已经相当出色。

(四)数据准备与提示词模板

为系统评估不同模型在缺失数据插补中的表现,本研究采用模拟实验(simulation experiment)的设计框架,通过操控缺失机制、输入结构与模型

---

① 一个潜在的顾虑在于,树模型在高维或多变量数据中可能面临性能下降的问题,且容易发生过拟合。为此,我们进一步引入了一种集成学习方法——堆叠(Stacking)模型,通过集成多个梯度提升决策树(GBDT)模型,期望提升插补的准确性。然而,在后续分析中我们发现,其性能与 LightGBM 模型相差无几,故不再单独呈现和讨论其结果。

② Jacob Devlin et al.,“BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” arXiv Preprint, 2019, https://arxiv.org/abs/1810.04805,访问时间:2024 年 11 月 5 日。

③ Yiming Cui et al.,“Pre-training with Whole Word Masking for Chinese BERT,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, Vol. 29, 2021, pp. 3504-3514.

类型，构建具备实验性质的评估环境。具体而言，我们基于 2020 年 CFPS 实测数据，模拟 MCAR、MAR 和 MNAR 三种不同的数据缺失机制，设计统一的“掩盖”方案，用以构造测试集，并保留真实值用于性能评估，从而在可控条件下对比各类插补方法的表现。

我们还特别关注调查数据结构（即横截面与纵贯设计）所对应的信息环境对插补效果的影响，因而设计了两种输入数据配置。此外，我们还比较了中外大语言模型在不同情境下的插补表现，并通过改变缺失机制的复杂程度和输入与训练数据的可得程度，评估在不同挑战性上各类方法的相对表现。

（1）输入设置

本研究旨在预测受访者在 2020 年 CFPS 调查中对美国人、医生以及地方政府官员的信任程度。我们设定两类信息环境，以评估大语言模型在不同输入条件下的插补能力：（A）仅使用基本人口统计学变量；（B）在此基础上，进一步加入受访者在 2018 年对相关社会问题的感知及其信任评分。这两种设定分别对应横截面数据结构与纵贯数据结构的应用情境。

需要强调的是，B 模块中新增的信息不仅作为大语言模型执行预测性插补时的提示输入，也被纳入传统统计方法的回归模型、用于机器学习模型的训练过程，并作为 BERT 模型微调阶段的关键输入变量。因此，该信息环境的拓展在所有方法路径中均具有重要影响。

上述三类变量分别是：

1. 人口统计学变量：年龄、性别、民族、城市居住情况、户口类型、职业、党派身份、教育水平、家庭收入和所在省份。①

2. 过去的相关社会问题感知：涉及信任水平的社会感知问题，包括对地方政府表现的评价，以及对医疗问题、社会保障问题和地方政府腐败程度的

① 我们使用 CFPS 的个人标识符将 2018 年的输入变量与 2020 年的目标变量匹配，仅保留在所有必需变量上数据完整的个体，并排除缺失值样本。对于需要数值型输入的模型（如 OLS、多重插补、机器学习模型），性别、民族、城市居住情况、户口类型和党派身份被转换为二元变量；教育水平被转换为受教育年限；职业映射到国际社会经济指数（ISEI）；家庭收入（总收入及人均收入）采用对数变换以减少偏态影响；省份则作为类别变量处理。此预处理步骤确保了数据一致性和模型兼容性。

看法。所有数据均来自 2018 年 CFPS 调查。

3. 过去的信任变量:个人在 2018 年对美国人、医生和地方政府官员的信任程度。

对于需要文本输入的模型(如 BERT、GPT 和 Qwen 模型),我们将结构化数据转化为中文文本描述,以便模型理解个体信息。表 2 是提供给大语言模型的一个来自于信息模块 B 的提示词示例。

表 2 大语言模型提示词示例

系统提示词(System Prompt):

个人的信任水平与其基本人口统计学特征、特定问题上的态度及其社会背景有复杂的关系。以下是一个中国个体在 2018 年的概况,包括基本人口统计学信息(如年龄和性别)以及其对特定问题的看法。请基于这些信息,推测该个体在 2020 年的信任评分。

信任评分包括三个维度:对美国人的信任(trust_usa)、对医生的信任(trust_doctor)以及对地方政府官员的信任(trust_govern)。每个信任评分范围为 0 到 10,其中 0 表示"完全不信任",10 表示"完全信任"。

请按照以下格式回复:trust_usa: 0 trust_govern: 0 trust_doctor: 0

请直接提供评分,不要附加额外信息。

用户提示词(User Prompt):

以下是一个中国个体的基本信息:

该个体为{23}岁{女性},{汉族},居住在{安徽省},拥有{农业户口}。目前生活在{农村地区},职业为{农业生产人员},并且{是中共党员}。2018 年,其家庭总收入为{1 万元},人均收入为{3000 元}。最高受教育程度为{初中毕业}。

在 2018 年,她对县级政府工作的评价为{在一定程度上有所作为}。她还评估了中国的一些问题严重程度(0 表示问题不严重,10 表示非常严重)。她对医疗问题的评分为{6},对社会保障问题的评分为{5},对政府腐败问题的评分为{7}。

2018 年,她对美国人的信任评分为{1},对地方政府官员的信任评分为{7},对医生的信任评分为{6}。

请根据以上信息,结合当时的社会背景,预测她在 2020 年的信任评分,并按照指定格式回复。

资料来源:作者自制。

(2) 数据准备

我们将整个 CFPS 在 2020 年的调查分成三部分:60%作为训练集,20%作为验证集,20%作为测试集。训练集用于拟合 OLS 模型、训练机器学习模型以及对 BERT 进行微调。在使用 Amelia Ⅱ进行多重插补时,训练集与测试集被合并为一个数据集进行处理。在每个实验设定中,我们从训练集中随机选取 10 个少样本学习(few-shot)示例,并在该设定内保持一致。验证集仅用于机器学习模型和 BERT 的超参数调优。

我们参考既有文献,针对三种典型的数据缺失机制设计了相应的数据"掩盖"方案,以构造测试集并用于评估各类模型的插补性能,具体设计详见表 3。[①] 鉴于本研究包含三个待插补的信任类目标变量,在每种缺失机制下,我们统一设定一个共同的测试集,即对这三个变量采用相同的一组缺失个体进行"掩盖",所有模型均在此基础上开展插补。换言之,我们并非为每个变量分别生成测试集,而是在多变量插补情境中对三个目标变量实施一致的缺失处理,以便更系统地评估模型的整体表现。采用这一策略的主要原因在于,在 CFPS 2020 年度调查数据($N=28530$)中,三项信任变量的缺失高度重合:14.97%的受访者至少缺失其中一项,而同时缺失三项的比例为 13.07%。

表 3 不同缺失机制下的数据"掩盖"设计

数据缺失机制	描述	数据"掩盖"方式
完全随机缺失(MCAR)	缺失完全随机,与任何变量无关	随机从完整数据集中抽取 20%样本作为测试集,每个受访者被选中的概率相同。
有条件随机缺失(MAR)	缺失依赖于其他可观测变量	使用常见且缺失较少的人口学变量拟合逻辑回归模型,以预测三个信任评分在 2020 年皆缺失的概率。根据模型估计的缺失概率,选取概率最高的 20%个体作为测试集。上述人口学变量在插补阶段均可获取。

① Sengupta et al., "Sparse Data Reconstruction, Missing Value and Multiple Imputation Through Matrix Factorization," pp. 72-114; Kim and Lee, "AI-Augmented Surveys," arXiv: 2305.09620v3.

续表

数据缺失机制	描述	数据“掩盖”方式
非随机缺失(MNAR)	缺失依赖于不可观测变量	构造如下权重函数:weight=[trust_usa+(10−trust_govern)+(10−trust_doctor)]^2,根据该权重从样本中随机抽取 20%个体作为测试集。信任美国人较高或对地方政府官员、医生信任较低的个体更易出现缺失。该设计模拟现实中较为常见的非随机缺失机制,即变量缺失值与其真实值直接相关,难以被观测变量充分解释。

资料来源:作者自制。

图 1 展示了在非随机缺失(MNAR)机制下抽样后的训练集与测试集的分布情况。其中,测试集即为通过抹去部分观测值所构建的“缺失数据”,用于后续插补评估。

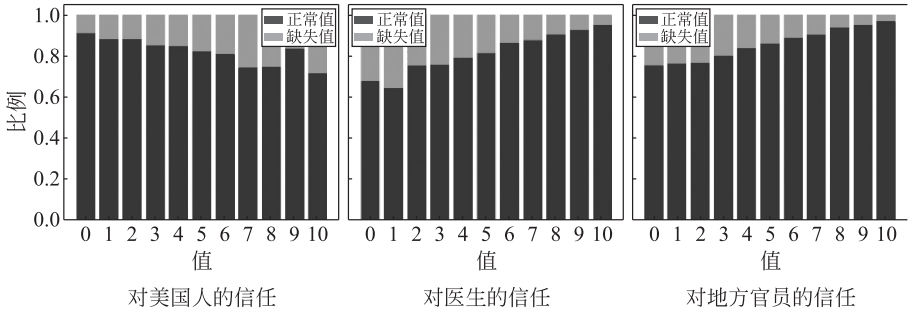

图 1 在非随机缺失机制下模拟生成的缺失数据(测试集)分布情况

资料来源:作者自制。

四、实验结果报告与分析

本文采用不同的方法,对根据不同缺失机制进行数据“掩盖”所得的缺失值进行插补,并根据前文所述的维度和指标对大语言模型的生成性方法和其他方法的表现进行对比评估。实验结果显示,由于各种方法在前提假定上都是至少用来处理条件随机缺失等复杂情况的,因此在完全随机缺失

的插补任务上表现和条件随机缺失的情况非常相似，但在非随机缺失和条件随机缺失两种情形下的表现差异较大。我们将聚焦后两种情况进行报告和分析。实验还发现，小样本学习（few-shot）并未显著提升大语言模型的表现，而零样本学习（zero-shot）设置下的缺失值插补更能体现大语言模型的能力。因此，我们对大语言模型表现的讨论主要聚焦在零样本学习设置下的各项指标（除特别说明外，后文“大语言模型”均指此情形）。

（一）有条件随机缺失（MAR）

表4报告了“有条件随机缺失”情形下，各种方法的数据插补表现。如上所述，在模块A中，模型的输入（以及微调与学习过程）仅包括个体的基本人口学变量，如性别、年龄等；而在模块B中，除了这些基本信息外，模型还接收个体在前一期对相关社会问题的感知以及对三类对象的信任水平作为额外输入。模块B所对应的，是纵贯数据插补中常见的情形，即模型可利用受访者在前一期对类似主观问题的回答；而模块A更贴近截面或重复截面数据的插补任务，在此类情形下往往无法获得个体既往的回答，只能依赖其基本人口学特征。

表4 有条件随机缺失条件下各模型的数据插补表现（RMSE）

方法	模型	信任对象		
		美国人	医生	地方官员
	模块A：输入人口学变量			
传统统计方法	OLS	3.584	3.303	3.627
	MI（Amelia Ⅱ）	3.011	3.129	3.483
机器学习	Matrix Factorization	2.782	2.772	3.029
	LightGBM	2.626	2.528	2.742
双向编码器表示模型（BERT）	Chinese-RoBERTa（微调）	2.626	2.529	2.728
大语言模型	Qwen-plus（零样本学习）	2.798	2.758	2.953
	GPT-4o-mini（零样本学习）	2.602	3.180	3.641

续表

方法	模型	信任对象		
		美国人	医生	地方官员
	模块 B:输入人口学变量+上一期信任及社会感知			
传统统计方法	OLS	3.288	3.074	3.271
	MI(Amelia Ⅱ)	2.926	2.885	3.147
机器学习	Matrix Factorization	2.574	2.439	2.604
	LightGBM	2.534	2.365	2.491
双向编码器表示模型(BERT)	Chinese-RoBERTa(微调)	2.549	2.417	2.490
大语言模型	Qwen-plus(零样本学习)	2.833	2.766	2.853
	GPT-4o-mini(零样本学习)	2.838	2.884	3.239

注:表中所展示数值为均方根误差,取值越大表示模型拟合误差越大,插补效果越差。
资料来源:作者自制。

表 4 显示,就个体插补的准确性而言,机器学习优于大语言模型。例如,在模块 B 中对"美国人信任"的插补,Qwen-plus 的均方根误差为 2.833,显著高于 LightGBM 的 2.534 和矩阵分解的 2.574。进一步比较模块 A 与模块 B 的结果可见,随着输入与训练信息的增加,所有机器学习模型的插补性能均实现提升,而大语言模型的表现则较为分化:在两种大语言模型与三个插补目标变量构成的 6 个组合中,只有 3 个情境显示出性能改进,分别是 Qwen-plus 对地方官员信任的插补、GPT 对医生信任的插补,以及 GPT 对地方官员信任的插补;而在其余 3 个情境中,大语言模型在引入上一期信息后反而表现下降,均方根误差增大。

这一结果出现的一个可能的解释是,引入纵贯信息后,大语言模型更倾向于依赖历史数据,从而忽视个体信任水平在不同时间点的异质性变化。这一倾向可从插补值与真实值的比较中得到体现(受限于篇幅,具体结果略去)。简而言之,与机器学习方法相比,大语言模型对 2020 年信任值的插补更集中于与 2018 年真实值相差−2 至 0 的区间,表明其预测主要是在 2018 年受访者回答的基础上进行整体性负向调整,而未能充分捕捉个体间信任水平变动的多样性与差异性。相比之下,机器学习模型生成的插补值则呈现出更大的变异性,对前期信任水平的依赖程度更低,更能体现个体信任变化的非均质性。

尽管在个体层面的插补准确性上不如主流方法,大语言模型在恢复数

据的整体分布特征，特别是在重建样本变异性（variation）方面展现出显著优势。图 2 对比了 LightGBM 与 Qwen-plus 在插补样本上的分布情况，并将其与真实测试数据进行了对照分析。真实数据中，受访者对“美国人信任”的评分多集中于低分段（如 0 或 1），而对“医生”与“地方官员”的信任值则整体偏高，呈现出较强的变量间分布差异。

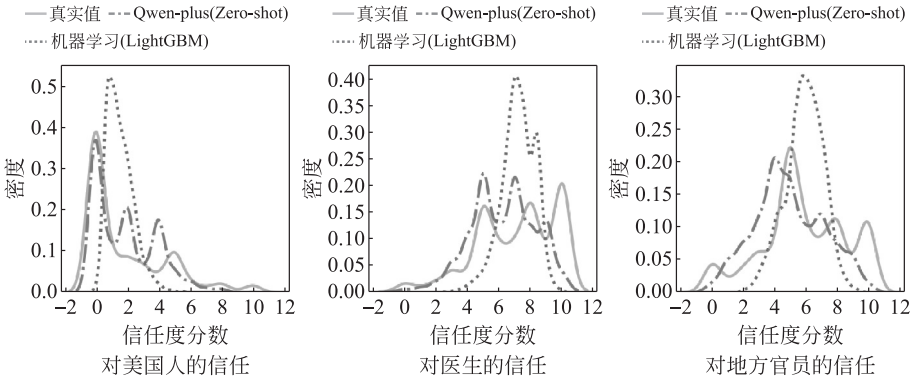

图 2 有条件随机缺失情形下插补数据及真实数据的样本分布

注：数据插补基于条件随机缺失场景，插补模型的输入包括人口学变量以及上一期的社会感知与信任指标（对应表 4 模块 B 的设置）。

资料来源：作者自制。

相比之下，LightGBM 所生成的插补值分布明显更加集中，呈现出对信任水平变异性的低估，其标准差显著低于真实样本。进一步计算显示，在测试数据中，“美国人信任”的真实标准差为 2.59，LightGBM 的插补值标准差仅为 0.85，而 Qwen-plus 的标准差为 1.96，更接近真实数据。这一结果表明，大语言模型在保持数据分布的变异性方面具有更强的能力。而机器学习插补方法倾向于低估样本变异性，具体表现为对个体信任值变动幅度的压缩。这带来两个潜在后果：第一，在下游的描述性分析中，可能导致对社会共识程度的误判；第二，在依赖这些插补数据开展的相关性分析中，插补数据的过度集中可能引发分析结果的系统性偏误。尽管已有研究指出大语言模型在预测与插补中可能低估个体差异，[①]我们的结果表明，其他插补方

① 例如 Bisbee et al.，“Synthetic Replacements for Human Survey Data?” pp. 401-416.

法在这个问题上很可能更严重,大语言模型实际上可能对这一问题有所改善。

比较 GPT 和 Qwen 这两种大语言模型的表现可见,本土大模型 Qwen-plus 的表现比国际大模型 GPT-4o-mini 更优,尤其在插补受访者对地方政府官员信任时,Qwen-plus 的插补结果更能贴近中国民众的真实回答。例如,在表 4 模块 B 的信息环境下,Qwen-plus 在插补"地方官员信任度"时的均方根误差为 2.85,明显优于 GPT-4o-mini 的 3.24。进一步分析表明,这一差异主要源于 GPT 模型倾向于低估民众对地方政府官员的信任。这一发现表明大语言模型的表现很有可能受其对本土文化和上下文背景理解的影响。

表 4 的实验结果中还有一些值得一提的发现。比如,绝大多数模型在加入上一期信任和感知变量(模块 B)后表现有所提升,验证了纵贯数据的信息增益价值。又如,传统统计方法整体表现不如机器学习方法,插补误差相对较高。例如,在模块 B 中对"美国人信任"变量的插补中,表现最好的传统统计方法为 Amelia Ⅱ(RMSE=2.926),而机器学习方法 LightGBM 的均方根误差更低,仅为 2.534。此外,经过微调的双向编码器表示模型(BERT)接近 LightGBM(RMSE=2.549),显示其作为深度语言模型在插补任务中的潜力。

(二) 非随机缺失(MNAR)

表 5 报告了在数据非随机缺失情形下各模型的插补表现。我们发现表 5 中各模型的均方根误差普遍高于表 4,表明在非随机缺失情境下,缺失数据的插补任务更具挑战性。此外,表 5 中各种方法的均方根误差并没有都随着输入信息的增加(即从模块 A 到模块 B)而降低,但大多数模型表现出误差逐步下降的趋势,这与表 4 中观察到的模式一致,说明即便在非随机缺失条件下,更多的输入与训练信息依然具有提升插补效果的价值。

就大语言模型的表现而言,表 5 显示了零样本学习大语言模型的相对优势,尤其体现在本土大语言模型 Qwen-plus 的表现上。在信息模块 B 中,Qwen-plus 针对美国人、医生和地方官员信任的插补,均方根误差分别为 2.731、2.405 和 2.389。这些结果不仅明显优于传统统计方法(如 Amelia

Ⅱ),在均方根误差这一指标上也与机器学习模型 LightGBM 以及 BERT 模型相当甚至略优。国际大模型 GPT-4o-mini 的表现虽然不如本土大模型,但也显示出了与最优模型相差不多的能力。

表 5 非随机缺失条件下各模型的数据插补表现(RMSE)

方法	模型	信任对象		
		美国人	医生	地方官员
模块 A:输入人口学变量				
传统统计方法	OLS	3.804	3.488	3.822
	MI(Amelia Ⅱ)	3.147	3.130	3.344
机器学习	Matrix Factorization	3.064	2.532	2.546
	LightGBM	2.882	2.752	2.782
双向编码器表示模型(BERT)	Chinese-RoBERTa(微调)	2.924	2.896	2.883
大语言模型	Qwen-plus(零样本学习)	2.673	2.657	3.479
	GPT-4o-mini(零样本学习)	2.784	2.495	2.626
模块 B:输入人口学变量+上一期信任及社会感知				
传统统计方法	OLS	3.459	3.145	3.281
	MI(Amelia Ⅱ)	2.965	2.889	3.083
机器学习	Matrix Factorization	2.818	2.376	2.295
	LightGBM	2.660	2.471	2.426
双向编码器表示模型(BERT)	Chinese-RoBERTa(微调)	2.764	2.633	2.524
大语言模型	Qwen-plus(零样本学习)	2.731	2.405	2.389
	GPT-4o-mini(零样本学习)	2.714	2.545	2.728

注:表中所展示数值为均方根误差,取值越大表示模型拟合误差越大,插补效果越差。
资料来源:作者自制。

表 5 模块 B 显示,在非随机缺失情境下,国产大语言模型(Qwen-plus)在数据插补中的均方根误差仅在“对美国人信任”这一指标上略低于国外模型(GPT-4o-mini),而在“对地方官员信任”的插补上则明显优于后者。为进一步揭示国产模型的相对优势,图 3 展示了两种大语言模型所生成的插补数据的分布,并将其与真实数据的分布进行对比。

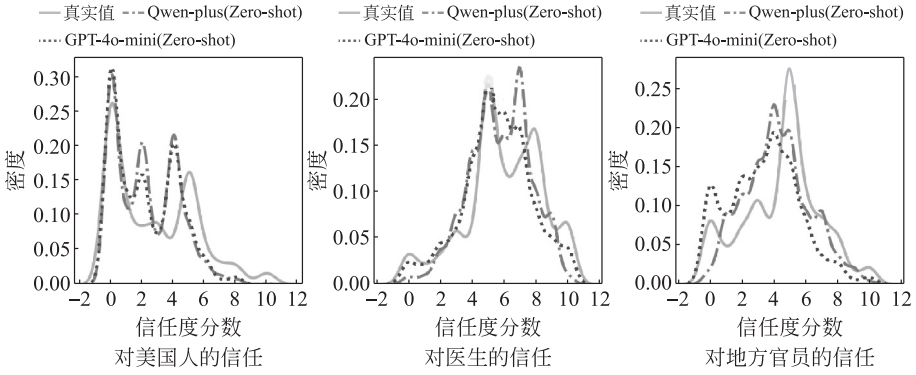

图 3 对比国内外大语言模型数据插补的表现

注:数据插补基于非随机缺失场景,插补模型的输入包括人口学变量以及上一期的社会感知与信任指标(对应表 5 模块 B 的设置)。

资料来源:作者自制。

结果表明,GPT 模型在插补中普遍低估了受访者对地方政府官员的信任,表现为对"完全不信任"(值为 0)的插补频率偏高;相比之下,Qwen-plus 显著弱化了这一低估对地方官员信任水平的倾向,有效降低了插补值为 0 的比例(见最右侧小图)。这一结果表明,国产大语言模型在理解我国民众对地方政府官员的态度方面表现更为贴近实际回答,插补更为准确,凸显了其在把握本土社会情境方面的优势。

对比表 4 和 5 可以看出,大语言模型在非随机缺失情境下相对表现的上升,并不是由于其绝对能力的提高,而是由于机器学习模型在该情境下的表现有所下降。但这种相对提升也具有重要意义:当缺失机制更偏向非随机时,零样本学习的大语言模型可能成为更具优势的替代方案。另外,值得注意的是,与机器学习模型需依赖训练数据、BERT 需依赖微调不同,零样本学习的大语言模型可在无需额外训练的前提下直接生成插补结果,这在某些场景中具有独特优势。例如,研究者若希望对过去某些年份中未被调查的问题进行"回溯性插补",其对应的训练数据可能根本不存在,此时大语言模型的零样本能力便尤为关键。

图 4 进一步显示,在非随机缺失机制下,大语言模型所生成的插补数据在样本变异性上更接近真实数据,而机器学习模型则倾向于低估信任程度

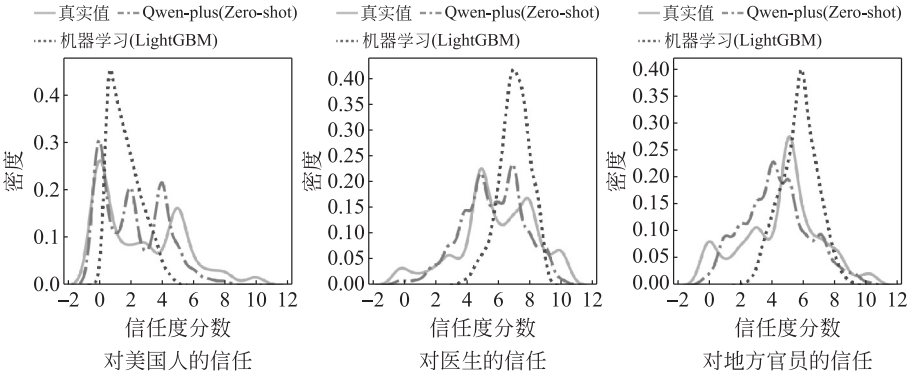

图 4 非随机缺失情形下插补数据及真实数据的样本分布

注：数据插补基于非随机缺失场景，插补模型的输入包括人口学变量以及上一期的社会感知与信任指标（对应表 5 模块 B 的设置）。

资料来源：作者自制。

的个体差异。这一发现与我们在条件随机缺失情境中（见图 2）所观察到的模式一致。

在下游相关性分析方面，大语言模型生成的插补数据在表现上也不逊于机器学习模型 LightGBM。图 5（见下页）显示，基于插补数据进行的相关性分析中，Qwen-plus 能够较好地再现真实数据中年龄和教育年限与信任变量之间的正向或负向关系。其中一个特例是：在真实数据中，年龄与对医生的信任呈负相关，而基于 Qwen-plus 生成的插补数据却显示出微弱的正相关趋势；不过，LightGBM 同样未能准确还原这一真实的负相关关系。

五、结语

本文旨在评估大语言模型对民意调查或社会调查中缺失数据进行插补的表现。基于“中国家庭追踪调查”（CFPS）数据，本文模拟了多种主观信任数据的缺失机制，并在不同情境下将大语言模型与传统插补方法（如多重插补、机器学习等）进行了系统比较。实验结果表明，大语言模型在缺失值插补领域展现出良好的应用前景。即使在零样本学习的条件下，只要输入信息充分（例如受访者在上一期的相关回答），其插补性能不仅优于传统统计

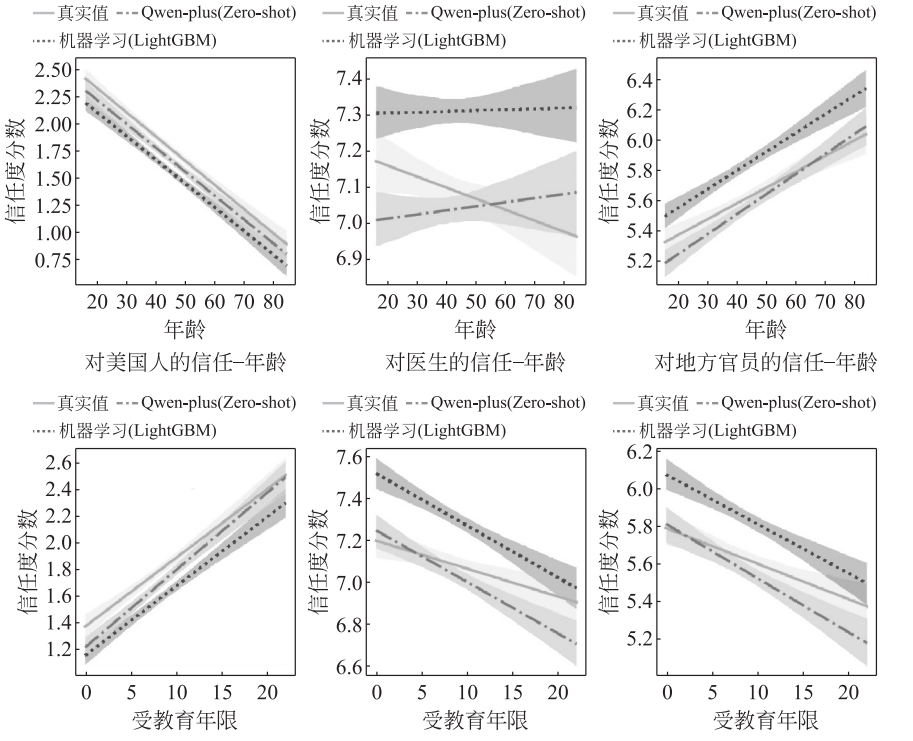

图 5 非随机缺失场景下基于信任插补数据的相关性分析

注:相关性分析基于未缺失数据和插补的缺失数据。数据插补基于非随机缺失场景,插补模型的输入包括人口学变量以及上一期的社会感知与信任指标(对应表 5 模块 B 的设置)。

资料来源:作者自制。

方法,且在多个关键指标上并不逊色于主流机器学习模型。尤其在数据非随机缺失这一既常见又较难处理的情境中,大语言模型表现尤为突出。此外,考虑到零样本学习的大语言模型无需额外的数据处理和训练,其优势在无法进行模型训练或缺乏历史样本的实际应用场景中愈加明显。

本文还评估了国内外大语言模型在中国社会文化语境下的适应性表现,发现 Qwen 在模拟受访者对地方政府官员的信任时展现出明显优势。这一结果不仅揭示了模型训练背景与其理解本地社会语境能力之间的联系,也体现了“主权 AI”在处理本国社会科学任务中的潜在价值。该方向值得在未来研究中进一步拓展和深化。

此外，本文还比较了截面数据与纵贯数据输入对模型插补表现的影响。尽管机器学习方法在引入上一期信息后普遍实现了性能提升，大语言模型却未展现出类似的增益，甚至在某些情形下出现了准确性下降的情况。大语言模型在处理纵贯信息时可能存在过度依赖历史数据的问题，从而忽略了个体信任水平的动态变化及其与社会环境变迁之间的复杂互动。相比之下，机器学习方法由于依赖目标期数据进行训练，在捕捉当期变异性方面表现出更强的适应性。

这一发现提示我们，在使用大语言模型进行数据插补时，需格外警惕输入信息使用所带来的悖论：尽管历史信息的引入总体上有助于提升预测准确性，但若依赖过度，可能反而削弱模型的模拟能力。未来研究应进一步探讨如何在利用既有信息与保持预测灵活性之间实现更优平衡，特别是在融合截面与纵贯数据结构特征的框架下，发展更具适应性与泛化能力的插补策略。