大语言模型能否成为可靠的硅基受访者？——基于 CFPS 回答分布的微调训练实验

林泽腾[1] 叶振业[2] 王仁和[2*]

（1. 香港科技大学，广东 广州，511453；

2.*通讯作者，华南师范大学 政治与公共管理学院，广东 广州，510006）

摘 要：随着生成式人工智能在社会科学研究中的应用不断扩展，大语言模型能否作为“硅基受访者”参与问卷预实验、调查设计和公众意见研究，成为计算社会科学与社会调查方法领域的重要问题。本文以中国家庭追踪调查（CFPS）2010—2022 年的个人面板数据为基础，围绕生活满意度和主观健康两个主观变量，让大语言模型在学会变量分布特征的基础上，以“硅基受访者”的身份参加后续年份的问卷问答。方法上，我们采用基于 Kullback-Leibler 散度约束的低秩适应有监督微调（FT-KL）训练大模型，并与问答式、第一人称自传式和第三人称角色描摹式三类零样本提示词工程进行效果对比。实验结果显示，相较三类零样本提示词方法，FT-KL 训练取得统计显著的改善，经过训练后的大模型可以很好地充当“硅基受访者”回答未来年份的问卷。稳健性检验进一步表明，这一结论不受主观变量设置、大模型参数规模、提示词设置的影响。研究表明，在真实的调查数据分布监督下，大模型可以成为可靠的硅基受访者。

关键词：大语言模型；硅基受访者；中国家庭追踪调查；计算社会科学

Can Large Language Models Become Reliable Silicon Respondents? A Fine-tuning Experiment Based on CFPS Response Distributions

Lin Zeteng[1] Ye Zhenye[2] Wang Renhe[2*]

(1. Hong Kong University of Science and Technology (Guangzhou), Guangzhou 511453, Guangdong, China;
2. School of Politics and Public Administration, South China Normal University, Guangzhou 510006, Guangdong, China)

Abstract: With the expanding use of generative artificial intelligence in social science research, whether large language models can serve as reliable “silicon respondents” has become an important methodological question for computational social science and survey research. Based on individual-level panel data from the China Family Panel Studies (CFPS) from 2010 to 2022, this study examines whether large language models can predict future response distributions for two subjective variables: life satisfaction and self-rated health. Methodologically, we fine-tune large language models using supervised fine-tuning with Low-Rank Adaptation under a Kullback-Leibler divergence constraint (FT-KL), and compare this approach with three zero-shot prompting strategies: question-answer

prompting, first-person biographical prompting, and third-person portrayal prompting. The results show that FT-KL consistently outperforms all prompting-based baselines across different model sizes, task settings, and subjective variables. Models calibrated with real survey response distributions generate predictions that are closer to actual CFPS distributions in terms of JSD, KL, L1, and EMD. These findings suggest that large language models do not naturally become reliable survey respondents through demographic role prompting alone. However, when constrained by real survey distributions, they can generate statistically more realistic silicon samples within clearly defined task boundaries. This study provides empirical evidence for using calibrated large language models as an auxiliary tool for questionnaire pretesting, policy pilot evaluation, and evidence-informed decision-making.

Keywords: Large Language Models; Silicon Respondents; China Family Panel Studies; Response Distribution; Computational Social Science

一、引言

问卷调查是社会科学理解公众态度、社会认知和群体差异的基础方法之一。与访谈、田野调查和小规模实验相比，标准化问卷能够在较大样本范围内收集结构化、可比较的数据，并通过连续追踪或重复调查记录社会态度与社会行为的变化。特别是中国家庭追踪调查（China Family Panel Studies，CFPS）这类长期面板调查，不仅能够描述不同群体在价值观、社会信任、公平感、幸福感、健康感知和社会经济地位判断等方面的差异，也能够帮助研究者观察同一受访者或同一类群体在不同时点上的变化轨迹。因此，标准化问卷长期以来都是社会科学开展经验研究、政策评估和公众意见分析的重要数据基础。

然而，传统问卷调查也存在明显限制。高质量调查通常需要严格的抽样设计、问卷开发、访问执行、质量控制和数据清理，成本较高、周期较长。在正式调查开展之前，研究者往往难以低成本地预估不同群体对特定题项的回答分布，也难以提前判断题目设计是否会造成理解偏差、选项集中或群体差异被遮蔽。随着样本接触难度上升以及电话调查回应率下降，传统问卷调查的执行成本和组织难度进一步增加。正是在这一背景下，学者开始讨论大语言模型作为“合成人类样本”“人工智能替身”甚至“硅基受访者”的可能性，期望其在问卷预实验、题项设计、公众意见趋势估计和社会模拟中发挥辅助作用。

围绕大模型能否辅助社会科学研究，国内相关讨论已经从一般性工具应用转向方法论反思。相关研究指出，生成式大语言模型不仅能够辅助写作、编码和资料整理，也可能为社会科学问题的预演、模拟和测试提供新的方法空间[1-3]。但已有研究也发现，大模型生成的“硅基样本”虽然形式完整，却容易存在代表性不足、异质性不足和结构性偏差，表现为强化主流话语、忽略边缘声音，并将复杂社会关系简化为更线性、更类型化的“合理预期”[4-5]。因此，大模型能否用于问卷调查研究，关键不在于其是否能够生成看似合理的个体回答，而在于其生成结果是否具有统计真实性。换言之，评价大模型生成社

会科学数据的有效性，不能只看个体层面的预测准确率，更应关注其生成的群体回答分布是否接近真实调查中的人口分布[6-8]。

由此可见，现有研究的核心挑战并不是让大模型“回答问卷”，而是如何使其回答分布更接近真实人群的统计分布。基于这一问题意识，本文将大模型在问卷选项对应词元上的概率分布作为核心约束，训练模型预测未来波次中的回答分布。具体而言，本文关注的不是模型最终输出了哪一个选项，而是各个选项的概率分布能否接近真实调查中人类受访者的回答比例。一方面，这一实验设计更符合社会调查研究关注群体分布而非单个答案的基本逻辑；另一方面，它也使大语言模型的问卷回答能够通过可量化的分布指标进行评价，而不是停留在“回答像不像人”的主观判断层面。

由于 CFPS 具有真实面板追踪结构，能够连续记录同一受访者在不同时点上的主观变量变化，因此为检验大语言模型能否学习受访者群体的主观回答规律提供了合适的数据基础。CFPS 不仅能够呈现某一类人在某一年如何回答，也能够观察同一批人在后续年份中的态度变化。本文以生活满意度和主观健康状况为预测目标，将上一波或历史波次的主观变量，以及性别、教育、城乡、地区、年龄等客观变量作为训练数据，检验经过训练后的大语言模型能否预测未来波次中相应群体的主观变量回答分布。

据此，本文主要回答两个研究问题：第一，对于问卷调查中具有连续追踪结构的主观变量，经过基于 Kullback–Leibler 散度约束的低秩适应监督微调后（下文简称 FT-KL），大语言模型能否准确预测未来波次的主观变量回答分布？第二，与社会科学研究中常见的零样本提示词工程方法相比，FT-KL 方法是否具有更高的预测准确性和统计真实性？通过回答上述问题，本文旨在为大语言模型在问卷调查研究中的应用提供一种可校准、可评估且边界明确的计算范式。

二、文献综述

（一）大语言模型在社会科学中的应用与影响

在大语言模型问世之前，部分社会科学研究已经出现由传统经验研究向大数据驱动和社会计算转型的趋势。互联网平台、社交媒体、移动通信和数字治理系统等所积累的大规模数字痕迹数据，为研究者观察社会行为、社会关系和社会结构提供了新的资料基础[9]。此类研究以社会科学问题意识为核心，将大数据、人工智能算法和社会科学理论结合起来，实现对社会现象和社会规律的观察与分析[10]。然而，早期计算社会科学虽然拓展了社会科学的研究对象和数据来源，但也存在较高的技术门槛。大语言模型出现后，其以研究辅助工具的形式进入文献综述、理论观点提出、研究设计、数据收集、数据分析和论文写作等多个环节，并对传统社会科学研究方式产生直接影响[11]。在计算政治学等具体分支中，大语言模型进一步拓展了方法体系和议题边界，推动了分支学科的发展[12]。

国内关于大模型赋能社会科学的讨论，已经从一般性工具应用逐渐转向方法论和研究范式层面的反思。相关研究指出，数字社会研究需要新的范式、方法与思路，生成式大语言模型不仅能够辅助写作、

编码和资料整理，也可能为社会学问题的预演、模拟和测试提供新的方法空间[1-3]。在“人工智能驱动的科学研究”这一第五研究范式背景下，以 ChatGPT 为代表的生成式人工智能正在推动计算社会科学从“大数据驱动”向“大模型驱动”转型：研究过程由“数据驱动”向“知识驱动”跃升，研究主体由单一研究者转向人机交互系统，研究目标也从社会解释进一步扩展到社会预测[13-15]。

在这一范式演进中，既有研究将大语言模型赋能社会科学的方向概括为生成式行动者、复杂因果分析、人机科研协同、硅基样本和社会预测等方面，强调大语言模型可以通过模拟人类语言表达、认知判断和互动过程，为社会科学提供新的观察对象、干预方式和知识生产路径[16-17]。由此产生的关键问题是：大模型生成的社会科学数据或问卷回答结果，能否在经验研究中具有一定可信度？正是在这一问题意识下，硅基样本开始从一般的大模型应用中分化出来，成为社会科学方法论和社会调查研究共同关注的新议题。

（二）硅基样本与硅基受访者

硅基样本，是指研究者通过提示词设置，让生成式人工智能平台“扮演”具有特定社会人口学特征的人类对象，并要求其回答问卷、参与实验或表达观点，由此形成的合成样本或人工智能受访者[18-19]。早期关于大语言模型社会模拟能力的研究，主要关注模型能否在特定人物背景或情境设置下生成与人类行为相近的反应。相关研究通过“图灵实验”检验语言模型是否能够复现经典人类被试研究，说明大语言模型在特定条件下可以模拟多个人类参与者的行为反应[20]。在此基础上，研究者通过真实人类受访者的人口学背景信息构造样本生成回答，并比较模型生成回答与美国调查数据之间的相似性。由此，大语言模型从一般文本生成工具演进为社会科学模拟对象[18]。从应用领域看，硅基样本已经逐渐进入社会学、政治学、法学、教育学和公共治理等研究场景，用于社会模拟、政策预评估和公共议题讨论等研究[21-23]。

从社会科学研究的循证视角看，硅基样本的意义并不只是技术层面的“模拟人类回答”，也在于其能否为政策分析和公共决策提供新的辅助证据来源。循证决策理论强调，高质量政策制定必须建立在“最佳可得证据”基础之上[24]，而社会调查正是政府获取公众意见、评估政策效果的重要证据来源[25]。然而，在现实决策过程中，证据不完备、决策情境复杂和调查资源有限等问题，导致风险社会中长期存在“循证困境”[26]。因此，如果大语言模型能够生成具有统计真实性的硅基样本，就有望在问卷预测试、政策试点评估和公众态度模拟等环节，为公共决策提供一种低成本、可快速迭代的辅助证据生成路径[27-28]。

但是，硅基样本的快速发展也伴随着明显争议。生成式人工智能可以帮助研究者扩展实验设计、理论探索和调查预演，但并不能取消对数据来源、偏误和外部效度的审查[29]。模型生成数据也不是现实经验观察本身，而是由训练语料、模型结构、价值对齐和提示词共同塑造的人工产物[30]。已有研究表明，大模型虽然能够生成形式上完整的受访者样本，但生成样本在代表性和异质性方面仍存在明显不足。周穆之等[4]发现，大模型生成的问卷回答可能过于“整齐”、过于“典型”，难以充分呈现真实社会中

不同群体之间以及群体内部的差异。龚为纲和黄思源[5]以 CGSS2021 为基准，系统评估不同大模型生成“硅基样本”的拟合度和偏差特征，发现大模型容易强化主流话语、忽略边缘声音，并将真实社会中复杂、非线性的关系简化为更线性、更类型化的“合理预期”。这说明，未经校准的大语言模型即使能够生成看似合理的问卷回答，也可能在代表性、异质性和结构性偏差方面存在系统问题。

进一步而言，硅基样本研究的核心问题不应停留在“模型生成的个体回答是否可信”，而应推进到“模型生成的总体是否具有统计真实性”。已有研究开始从单变量分布、双变量关联、多变量预测、生命历程序列及其关联结构等方面，评估大语言模型生成社会科学数据的统计真实性[8]。关于大语言模型预测社会科学实验结果的研究也表明，模型可以作为实验预判工具，但其预测能力必须通过真实实验结果持续检验[31]。总体来看，当前大语言模型尚难精准复现公众态度，也不能直接替代人类应答[32]，但其作为硅基样本、调查预演和社会模拟工具的潜力，已经构成社会科学方法创新的重要方向。

（三）大语言模型在社会调查中的应用

社会调查是检验硅基样本有效性的重要场景。与一般文本生成任务不同，社会调查具有标准化题项、明确选项、抽样设计和真实数据基准，因而能够为评估大语言模型的受访者模拟能力提供相对清晰的检验框架。特别是对于生活满意度、主观健康、社会信任、公平感和政策态度等主观变量而言，问卷调查真正关心的并不是某一个孤立个体在某道题上的单次选择，而是特定总体或特定群体在不同选项之间形成的比例分布。换言之，社会调查中的回答没有唯一正确答案，研究者更关心的是不同性别、年龄、教育程度、城乡身份或历史状态的受访者，各有多少比例选择某一选项。

在硅基受访者研究中，提示词工程是最常见的做法。研究者通常在提示词中加入性别、年龄、教育、党派、种族、国家或文化背景等信息，要求模型“作为某类人”回答问题。跨国主观意见评估任务考察了模型回答是否能够代表不同国家和文化群体的主流意见[33]；OpinionQA 将美国民调问题转化为模型回答任务，用以评估语言模型是否能够反映不同人口子群体的意见分布[34]；关于人口学角色提示的系统比较也表明，提示词形式本身会显著影响模型输出[35]。这些研究说明，人口学身份提示可以在一定程度上改变模型回答，但也存在根本限制：提示词并不改变模型内部的知识结构和概率配置，结果容易受到措辞、语言、顺序和角色设定影响；同时，角色提示也可能激活模型对“某类人”的刻板化先验，而不一定反映真实调查中的经验比例[33,36-38]。对 ChatGPT 生成公共意见数据的评估表明，合成回答可能在平均水平上接近某些结果，却在方差、回归关系和群体差异方面与真实调查不一致[39]。

中文语境下的相关研究也显示出类似问题。已有研究发现，大模型生成的“硅基样本”虽然形式完整，但往往存在代表性不足、异质性不足和结构性偏差，容易强化主流话语、忽略边缘声音，并将复杂社会关系简化为更线性、更类型化的“合理预期”[4-5,40]。更重要的是，社会调查回答的本质是分布性的，单纯用准确率衡量模型可能遮蔽真实差异[6]。如果仅仅通过提示词工程要求大模型扮演某类“硅基受访者”，大模型即使命中了多数人的答案，也可能忽略少数意见、群体内部异质性和真实社会的分布尾部信息[7]。

近期研究开始使用真实社会调查回答分布直接约束大语言模型。基于世界价值观调查的研究构建了国家层面的回答分布任务，用真实国家人群在各选项上的比例对模型进行微调，并检验模型对未见国家和未见问题的泛化能力[41]。SubPOP 数据集则将题项、子群体和回答分布整理为训练样本，使用选项词元概率上的 KL 散度训练模型，使其输出分布接近真实公共意见分布[42]。这些研究表明，分布微调代表了社会调查中大语言模型应用的重要转向：评价大语言模型生成社会科学数据的有效性，不能只看个体层面的预测是否准确，而应关注人口层面的统计真实性，即模型生成的整体回答分布是否接近真实调查中的群体分布[8]。不过，现有研究仍未充分解决两个问题：第一，已有研究多从国家、文化或较粗粒度人口子群体层面评估模型回答分布，较少在真实社会内部更细粒度的人口学群体和历史状态之间学习回答差异；第二，现有研究大多依赖横截面调查或静态公共意见数据，较少利用面板调查中的历史信息预测未来主观回答分布。由此，如何在真实追踪调查框架下，将个体或群体的历史回答、人口学特征与未来回答分布相连接，仍是大语言模型应用于社会调查研究需要进一步推进的问题。

三、数据来源与实验方法

（一）数据来源与描述性统计

本文使用中国家庭追踪调查（China Family Panel Studies，CFPS）作为实验基准数据。选择 CFPS 作为本文基准数据，主要基于三个方面的考虑。第一，CFPS 具有较强的全国代表性和丰富的人口学信息，能够为不同性别、年龄、教育程度、城乡身份和地区群体的回答分布构造提供数据基础。第二，CFPS 连续覆盖多个调查波次，并通过跨年个人识别码持续追踪同一受访者，区别于 CGSS、CSS 等一般横截面调查，不仅能够描述某一时点不同群体的主观回答分布，也能够观察同一批受访者在后续年份中的变化轨迹。第三，CFPS 包含生活满意度、主观健康等稳定重复测量的主观变量，适合构建“历史状态—人口学特征—未来回答分布”的面板预测任务。谢宇[43]在介绍 CFPS 的理念与实践时指出，社会现象具有多层次、多维度和时间性的特点，CFPS 正是通过个体、家庭和社区多个层面的连续调查，为研究中国社会变化提供经验基础。因此，如果大语言模型要成为可校准的“硅基受访者”，它不仅应能模拟某一时点的回答，还应当能够在真实面板数据约束下学习受访群体主观变量随时间变化的经验规律。

本文使用 CFPS2010、2012、2014、2016、2018、2020 和 2022 年共七个调查波次的个人层面追踪数据，并利用跨年个人识别信息匹配同一受访者在不同年份中的记录。实验变量方面，本文选择生活满意度和主观健康状况作为预测目标。生活满意度反映受访者对自身生活状况的总体评价，是社会科学研究中常用的主观福利指标；主观健康状况则反映受访者对自身健康水平的总体判断，容易受到年龄增长、疾病冲击、劳动状态和家庭环境等因素影响。客观变量方面，本文选取性别、教育程度、城乡、地区和年龄组等人口学信息，并将其统一重编码为可跨年比较的类别变量。由此，本文既能够考察不同人口学

群体的主观回答分布差异，也能够进一步利用同一受访者的历史主观状态预测其所在群体在未来波次中的回答分布。

本文的训练样本基于 CFPS 的面板结构构造。具体而言，本文首先提取受访者上一波次或历史波次的主观题回答，并将其与人口学变量单独或交叉组合，形成群体定义；随后，在每一组转移波次内，统计属于同一群体的受访者在目标年份五个选项上的回答人数，并以有效回答总人数为分母进行归一化，得到和为 1 的目标回答分布。由此，本文最终使用的样本单位不再是单个受访者，而是“历史状态+群体属性+目标年份”所对应的回答分布。例如，在生活满意度任务中，一条真实样本可以表示为：2020 年生活满意度为“比较满意”、居住在东部地区、年龄为 30—44 岁的一组受访者，在 2022 年对“您对自己生活的满意程度”这一题的五个选项上的回答分布。在该样本中，有效受访者共 662 人，其中选择“很不满意”的有 3 人，占 0.45%；选择“不太满意”的有 8 人，占 1.21%；选择“一般”的有 143 人，占 21.60%；选择“比较满意”的有 405 人，占 61.18%；选择“非常满意”的有 103 人，占 15.56%。在主观健康任务中，样本构造方式相同，只是将历史状态和预测目标替换为“上一波主观健康状况”和“目标年份主观健康状况”。例如，2020 年主观健康状况为“比较健康”、居住在中部地区、年龄为 45—59 岁的一组受访者，在 2022 年对“您认为自己的健康状况如何”这一题的五个选项上的回答分布为：有效受访者共 546 人，其中选择“非常健康”的有 49 人，占 8.97%；选择“很健康”的有 48 人，占 8.79%；选择“比较健康”的有 343 人，占 62.82%；选择“一般”的有 53 人，占 9.71%；选择“不健康”的有 53 人，占 9.71%。

为检验模型对未来波次回答分布的预测能力，本文按照目标年份划分训练集、验证集和测试集。目标年份不晚于 2018 年的样本进入训练集，目标年份为 2020 年的样本进入验证集，目标年份为 2022 年的样本进入测试集。这样的划分方式可以避免同一目标年份的真实回答分布同时出现在训练集和测试集中，从而使测试集更接近真实的未来预测场景。本文进一步构造两类面板回答分布预测任务。第一类为连续两波之间的短期预测任务，例如 2018→2020 和 2020→2022，用于检验模型对短期主观状态变化的预测能力。第二类为任意较早波次预测较晚波次的长期预测任务，例如 2010→2022、2014→2022 和 2020→2022，用于检验模型能否利用不同时间间隔下的历史主观状态信息预测未来回答分布。表 1 概括了本文用于训练、验证和测试的样本规模。

表 1 重新构建后的 CFPS 面板回答分布样本

目标变量	总样本数	训练集	验证集	测试集	最小有效样本量
生活满意度（任意较早波次预测较晚波次）	9759	4996	2232	2531	80
生活满意度（只看连续两波之间的变化）	2796	2020	405	371	80
主观健康（任意较早波次预测较晚波次）	11332	5364	2713	3255	30
主观健康（只看连续两波之间的变化）	3348	2210	574	564	30

由表 2 可见，本文构造的群体分布样本具有较为稳定的样本基础。在生活满意度任务中，各子样本的中位群体人数大多介于 369 至 503 人之间；在主观健康任务中，各子样本的中位群体人数介于 334.5 至 533 人之间，说明多数群体分布并非由少量个体计算得到，具有一定统计稳定性。同时，第二类任务的平均时间间隔明显长于第一类任务，不仅用于考察连续波次之间的短期主观变量变化，也进一步纳入较长时间跨度下的主观状态转移，从而能够检验模型对不同时间尺度下回答分布变化的预测能力。

表 2 CFPS 分样本的描述性统计

数据集	样本数	波次间隔（年）	中位群体人数	最小群体人数	最大群体人数
生活满意度（任意较早波次预测较晚波次）					
训练集	4996	3.99	460.0	80	10348
验证集	2232	6.07	377.5	80	6511
测试集	2531	7.18	369.0	80	5837
生活满意度（只看连续两波之间的变化）					
训练集	2020	2.00	495.0	80	10348
验证集	405	2.00	503.0	80	6511
测试集	371	2.00	466.0	80	5465
主观健康（任意较早波次预测较晚波次）					
训练集	5364	3.92	487.0	30	12343
验证集	2713	5.82	376.0	30	9174
测试集	3255	6.80	340.0	30	8466
主观健康（只看连续两波之间的变化）					
训练集	2210	2.00	533.0	30	12343
验证集	574	2.00	405.0	30	9174
测试集	564	2.00	334.5	31	7958

注：群体人数指构造某条回答分布样本时可用于计算目标分布的有效受访者人数。波次间隔以年为单位。第一类任务的时间间隔固定为 2 年，第二类任务包含多个历史年份到同一目标年份的组合，因此验证集和测试集平均时间间隔更长。

（二）实验变量

在 CFPS 变量构造方面，由于不同年份变量编码形式不同，本文对原始问卷变量进行了统一重编码，如表 3 所示。生活满意度题项统一为五级有序选项，包括“很不满意”“不太满意”“一般”“比较满意”和“非常满意”；主观健康题项统一为五级有序选项，包括“非常健康”“很健康”“比较健康”“一般”和“不健康”。人口学变量方面，性别划分为男、女；教育程度划分为小学及以下、初中、高中/中职、大专及以上；城乡划分为城镇、乡村；地区按照省份代码划分为东部、中部、西部和东北；年龄根据调查年份与出生年份计算，并划分为 16—29 岁、30—44 岁、45—59 岁和 60 岁及以上。对于原始数据中的缺失值、负值，以及 79、98、99 等无效编码，本文均在构造分布前予以剔除，以保证每条回答分布均来自有效受访者的真实回答。

在群体定义方面，本文将“上一轮主观状态”与不同人口学属性组合为 16 类群体。具体包括：上一轮主观状态与性别、教育程度、城乡、地区、年龄组的组合，以及上一轮主观状态与性别×教育、性别×地区、性别×城乡、性别×年龄组、教育×地区、教育×城乡、教育×年龄组、地区×城乡、地区×年龄组、年龄组×城乡等二重人口学交叉组合。同一受访者可能同时被计入“上一轮生活满意度×性别”“上一轮生活满意度×地区”“上一轮生活满意度×性别×年龄组”等不同分布样本，在保留样本

规模的同时，以尽可能捕捉不同社会属性条件下的主观状态转移差异，关注不同社会属性条件下的回答分布规律。

对于每一个“历史状态—人口学分组—目标年份”的组合群体，其在目标年份选项上的有效回答人数，记为：

$$(n_1,n_2,n_3,n_4,\cdots,n_i) \tag{1}$$

其中，n_i表示该群体中选择第i个选项的有效受访者人数。将该群体在该题项上的有效回答总人数为分母进行归一化：

$$p_i = \frac{n_i}{\sum_{j=1}^{5} n_j} \tag{2}$$

由此得到目标回答分布：

$$P=(p_1,p_2,p_3,p_4,\cdots,p_i) \tag{3}$$

该分布天然满足：

$$\sum_{i=1}^{5} p_i = 1 \tag{4}$$

为避免小群体导致分布估计不稳定，生活满意度每条分布样本的有效回答人数不少于 80 人，主观健康变量由于有效样本规模和变量分布不同，有效回答人数不少于 30 人，低于该门槛的群体组合不进入训练、验证和测试数据。

表 3 CFPS 实验变量及构造方式

变量名称	构造方式	说明
	个人识别变量	
pid	跨波次个人识别码	用于匹配同一受访者在不同波次中的记录
	目标变量	
生活满意度	五级有序变量	很不满意、不太满意、一般、比较满意、非常满意
主观健康	五级有序变量	非常健康、很健康、比较健康、一般、不健康
	历史状态	
上一波/历史波次生活满意度	与目标变量同口径	用于构造生活满意度转移任务
上一波/历史波次主观健康	与目标变量同口径	用于构造主观健康转移任务
	人口学变量	
性别	男、女	来自跨年个人信息
教育程度	小学及以下、初中、高中/中职、大专及以上	对原始教育变量合并重编码
城乡	城镇、乡村	按各波次城乡变量重编码
地区	东部、中部、西部、东北	按省份代码划分
年龄组	16—29、30—44、45—59、60 岁及以上	由调查年份减出生年份得到

（三）实验设计

当前多数社会科学研究使用大语言模型模拟问卷受访者时，主要依赖提示词工程，即通过设定人口学身份、社会背景或角色描述，让模型“扮演”某类受访者并生成回答（周穆之等，2025）。仅依靠提示词让模型模拟受访者，仍然可能面临代表性不足、异质性压缩和结构性偏差等问题。

因此，本文设计两类实验，一类是无需参数更新的零样本人口学提示词方法，另一类是基于真实调查回答分布进行 KL 散度约束的低秩适应有监督微调方法（FT-KL）。前者代表既有硅基受访者研究中最常见的“身份提示”思路，即通过提示词告诉模型受访者的年份、历史状态、性别、教育、城乡、地区和年龄等信息，再观察模型如何作答；后者则进一步把真实 CFPS 面板调查中的回答分布作为监督信号，直接约束模型在问卷选项词元上的概率分布。换言之，实验包括四种，其中三种为零样本提示词，即 QA 式零样本、BIO 式零样本和 PORTRAY 式零样本。最后一种，为基于真实调查回答分布进行 Kullback-Leibler 散度约束下的低秩适应有监督微调（FT-KL）。

在提示词构造方面，我们采用三类零样本提示词基线。包括：QA 式零样本、BIO 式零样本和 PORTRAY 式零样本。三者都不给模型任何历史真实回答分布示例，只提供调查年份、历史主观状态、群体特征、问卷题干和选项；差异仅在于身份信息的书写方式。QA 式提示采用问答方式呈现受访者信息，BIO 式提示采用第一人称自传式身份描述，PORTRAY 式提示则采用第三人称角色描摹方式。示例如表 4 所示：

微调实验采用低秩适应（LoRA）方法对模型参数进行高效更新，秩设置为 4，缩放系数为 16，随机失活率为 0.05，更新范围覆盖模型自注意力机制中的查询、键、值与输出投影层，以及前馈网络中的门控与上下行投影层。优化器选用 AdamW，学习率为 2×10^{-5}，权重衰减为 0，每步有效批量大小为 32，每设备批量 8，梯度累积 4 步。学习率采用线性预热后线性衰减的调度方式，预热比例占总训练步数的 3%。训练上限设为 10 轮，以验证集 JSD 为早停监控指标，连续 2 轮未改善时终止训练，并以验证集 JSD 最低轮次对应的模型权重作为最终评估模型。全部实验固定随机种子为 42 以保证可复现性。选项概率的提取方式为：将完整提示词输入模型后，取最后一个词元位置的输出对数概率，抽取 A 至 E 五个选项对应抽取词元的得分，经 softmax 函数归一化为预测概率分布；KL 散度计算中引入极小平滑项（$\varepsilon=10^{-12}$）以保证数值稳定，未额外设置生成温度参数。

表 4 三种零样本人口学提示词的构造方式与实际示例

提示词类型	实际示例
QA 式零样本	问：调查年份是哪一年？ 答：2022 年。 问：这名受访者上一轮生活满意度如何？ 答：2020 年回答为“比较满意”。 问：这名受访者的群体特征是什么？ 答：女性，30—44 岁，城镇户口，大专及以上学历，居住在东部地区。 现在请回答调查问题：您对自己生活的满意程度如何？ A.很不满意　B.不太满意　C.一般　D.比较满意　E.非常满意 请只输出一个选项字母。

BIO 式零样本	我是一名 2022 年的中国家庭追踪调查受访者。 我在 2020 年回答自己的生活满意度为“比较满意”。 我的群体特征是：女性，30—44 岁，城镇户口，大专及以上学历，居住在东部地区。 现在我需要回答下面这个调查问题： 您对自己生活的满意程度如何？ A.很不满意 B.不太满意 C.一般 D.比较满意 E.非常满意 请只输出一个选项字母。
PORTRAY 式零样本	请模拟一名 2022 年的中国家庭追踪调查受访者。 这名受访者在 2020 年回答自己的生活满意度为“比较满意”。 这名受访者是一名女性，年龄为 30—44 岁，城镇户口，大专及以上学历，居住在东部地区。 请根据这名受访者的背景，回答下面的调查问题： 您对自己生活的满意程度如何？ A.很不满意 B.不太满意 C.一般 D.比较满意 E.非常满意 请只输出一个选项字母。

注：三类提示词均不提供任何历史真实回答分布，只提供调查年份、历史主观状态、人口学特征、问卷题干和选项；区别仅在于受访者身份信息的表达方式。

（四）评价指标

本文将模型输出设计为一种接近问卷作答过程的概率化预测，大语言模型在生成每个词元时都会给出在“A、B、C、D、E”等选项词元上的概率，将其归一化为五项回答分布后，可以得到模型对该群体回答分布的预测结果。本文使用四类分布距离指标评价模型预测分布与真实调查分布之间的接近程度，以 JSD 作为主指标，同时报告 KL、L1 和 EMD。同时呈现整体分布差异、训练目标一致性、比例误差和有序等级偏移。设模型预测分布为：

$$Q=(q_1,q_2,q_3,q_4,\cdots,q_k,) \tag{5}$$

其中K表示选项数量。本文生活满意度和主观健康均为五级有序变量，$K=5$。在评估指标计算前，本文对真实回答分布P和Q加入极小平滑项并重新归一化，以避免零概率导致对数不可定义。本文采用四个评估指标评价模型的生成表现，四个指标均为距离型指标，数值越小，表示模型预测分布与真实调查分布越接近。

第一个评价指标是 Jensen–Shannon 散度（JSD），用于衡量真实分布和预测分布之间的整体差异。设：

$$M=\frac{1}{2}(P+Q) \tag{6}$$

则：

$$JSD(P,Q)=\frac{1}{2}D_{KL}(P\|M)+\frac{1}{2}D_{KL}(Q\|M) \tag{7}$$

JSD 具有对称性，适合用作本文的主评价指标。JSD 越小，说明模型预测分布越接近真实调查分布。

第二个评价指标是 KL 散度，采用真实分布到预测分布的方向：

$$D_{KL}(P\|Q)=\sum_{i=1}^{K}p_i\log\frac{p_i}{q_i} \tag{8}$$

如果真实调查分布是P，但模型给出的分布是Q，KL 散度衡量了二者之间存在多大信息偏离。该指标对“真实比例较高、但模型预测概率较低”的选项较为敏感，适合检验模型是否遗漏真实群体中的重要回答倾向。

第三个评价指标是 L1 距离，用于衡量模型在各选项比例上的总偏差：

$$L1(P,Q)=\sum_{i=1}^{K}|p_i-q_i| \tag{9}$$

它解释了模型在五个选项上的预测比例与真实比例累计相差多少。L1 越小，表示模型对各选项比例的整体拟合越接近真实样本。

第四个评价指标是 EMD，即地球移动距离。由于生活满意度和主观健康都是五级有序变量，仅看各选项比例差异还不够。比如，把“比较满意”预测成“非常满意”和预测成“很不满意”，在社会调查含义上并不相同。EMD 衡量了有序选项上的分布偏移。设 $CDF_P(k)$和 $CDF_Q(k)$分别表示真实分布和预测分布在第k个选项处的累积分布，则：

$$EMD(P,Q)=\sum_{k=1}^{K}|CDF_P(k)-CDF_Q(k)| \tag{10}$$

EMD 越小，说明模型不仅在总体比例上接近真实分布，而且在有序等级上的偏移也更小。

四、实验结果与分析

（一）总体结果

为便于系统比较不同模型规模、任务设定与评价方法之间的表现差异，本文基于千问大模型（Qwen），采用 Jensen–Shannon 散度（JSD）为核心评价指标，绘制总体结果热力图，如图 1 所示。JSD 用于衡量模型预测分布与真实调查分布之间的差异，数值越小，说明模型预测结果越接近真实人群的回答分布；图中颜色越深，也表示预测误差越低。

图 1 展示了不同模型在四类 CFPS 预测任务中的总体表现，具体包括生活满意度与主观健康两个题项、连续波次预测与跨波次预测两类任务，以及三种零样本提示词方法和 FT-KL 微调方法的测试结果。整体来看，FT-KL 在几乎所有模型和任务中都取得了最低的 JSD 值，说明经过真实回答分布约束后的模型，能够明显更好地预测未来波次中的群体回答分布。相比之下，三种零样本提示词方法的 JSD 普遍较高，尤其在跨波次预测任务中误差更为明显，表明仅依靠提示词让模型“扮演受访者”，难以稳定恢复真实调查中的回答比例。

进一步看，模型规模与预测表现之间并不存在严格的单调递增关系。较大的模型并不必然取得更低的 JSD，而部分中小规模模型在经过 FT-KL 微调后，同样能够获得较强的分布预测能力。例如，在多个任务中，Qwen2.5-32B 和 Qwen3-32B 的 FT-KL 结果表现较好，但较小规模模型在 FT-KL 条件下也明显优于其对应的零样本提示结果。这说明，影响模型预测效果的关键并不只是参数规模，而在于是否利

用真实调查分布对模型进行校准。换言之，分布约束微调能够显著提升大语言模型作为“硅基受访者”的统计真实性，使其预测结果更接近真实社会调查中的群体回答分布。

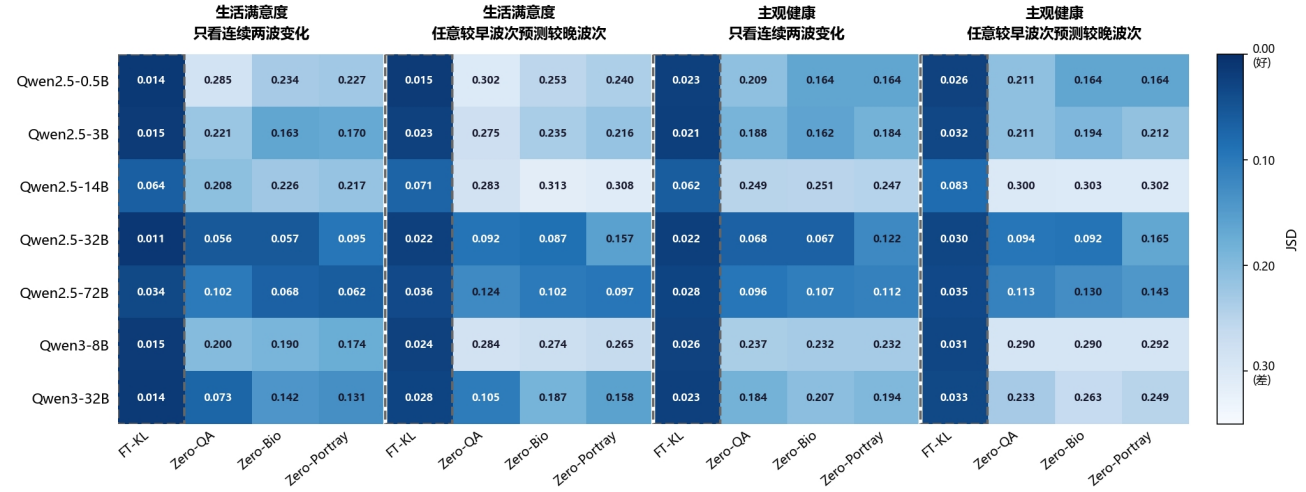


图 1 不同模型规模、任务设定与评价方法的 JSD 热力图

注：行表示基础模型，列表示任务与方法组合。每组中第一列为 FT-KL，后三列分别为 QA 式零样本、BIO 式零样本和 PORTRAY 式零样本提示。单元格数值为 JSD，数值越低表示模型预测分布越接近真实回答分布，颜色越深效果越好。

为进一步考察不同方法是否能够保留真实人群回答中的差异性，本文计算了模型预测分布的归一化熵，并绘制其密度分布。归一化熵可以理解为回答分布的“分散程度”：数值越高，说明模型把概率较均匀地分配到多个选项上，能够反映群体内部存在不同意见；数值越低，说明模型更倾向于把概率集中在少数几个选项上，容易把真实人群中较为分散的回答简化为少数主流答案。

从图 2 可以看出，真实调查数据的熵值整体较高，四个任务中的均值大致位于 0.69—0.79 之间。这说明，在真实社会调查中，受访者的回答并不是高度集中在某一个选项上，而是普遍存在一定程度的分散性和群体内部差异。相比之下，PORTRAY 式零样本的熵值最低，四项任务中的均值仅约为 0.39—0.46，且整体分布明显偏向低值区域。这表明，该方法最容易把回答集中到少数选项上，难以呈现真实人群中存在的多样化回答。QA 式零样本和 BIO 式零样本的熵值处于中间水平，虽然比 PORTRAY 式零样本更接近真实分布，但与真实调查数据相比仍有明显差距，说明单纯依靠提示词让模型扮演受访者，仍难以稳定恢复真实问卷数据中的群体差异。

相比之下，FT-KL 的熵分布整体上最接近真实调查数据。其均值大多位于 0.67—0.75 附近，说明该方法既不会像 PORTRAY 式零样本那样把回答过度集中到少数选项上，也能够较好地保留不同选项之间的比例差异。换言之，FT-KL 不仅提高了模型对真实回答分布的预测精度，也更好地保留了真实问卷回答中的多样性和群体内部差异。这表明，经过真实调查分布约束后的模型，比单纯依靠提示词的方法更适合用于模拟社会调查中的群体回答分布。

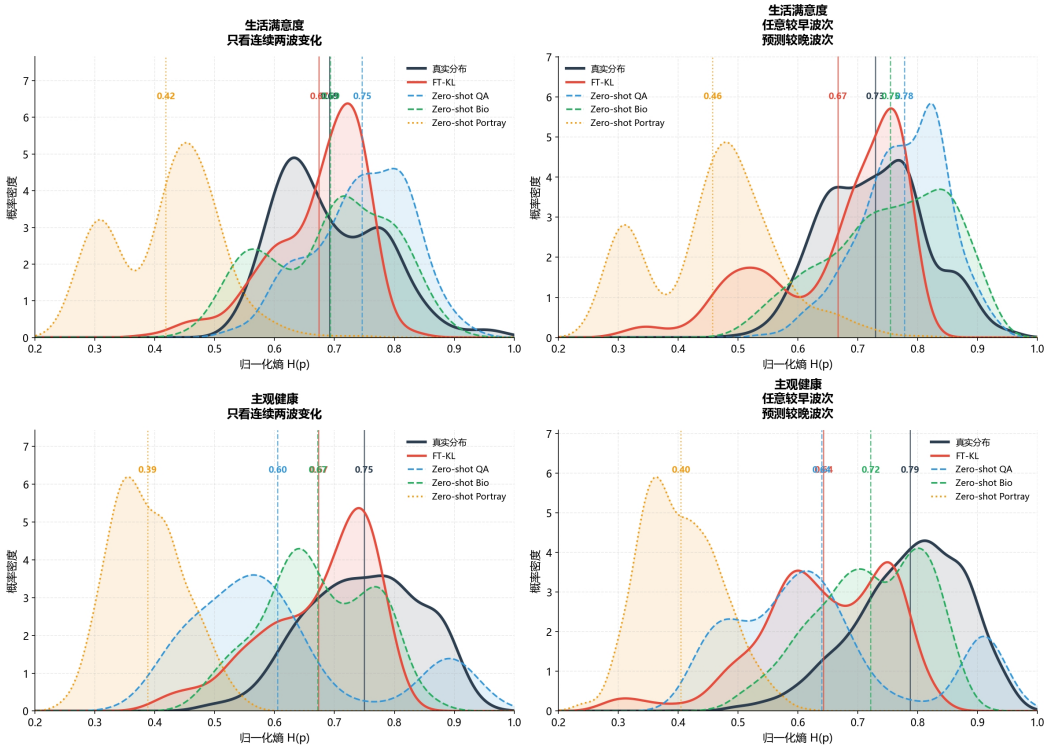


图 2 不同方法下回答分布异质性的保持情况

注：横轴为归一化熵 $H(p)=-\sum_i p_i \log p_i/\log K$，其中 K 为选项数。熵值越高，表示模型在多个选项之间分配的概率越分散，预测分布的异质性越强；熵值越低，表示模型输出更集中，分布更容易发生收缩或坍缩。图中黑色曲线为真实分布，红色曲线为 FT-KL，蓝色、绿色和橙色曲线分别表示 QA、BIO 和 PORTRAY 三类零样本提示词方法，竖线表示各方法的均值位置。

（二）生活满意度

图 3 是以 Qwen2.5-32B 为例，生活满意度（任意较早波次预测较晚波次）任务的分布比较。左图显示，真实分布主要集中在“一般”“比较满意”和“非常满意”三个选项上，其中“比较满意”和“非常满意”占比较高，同时仍保留一定比例的中间评价和低满意度评价。三类零样本提示词虽然能够大体生成向高满意度选项集中的回答倾向，但它们在各选项上的概率分配仍与真实分布存在明显偏离，尤其容易在低满意度选项和中间选项上产生系统性误差。相比之下，FT-KL 的预测曲线与真实分布形状更为接近，说明真实调查分布监督能够有效校准模型在不同选项之间的概率结构。从右图看，FT-KL 不仅平均误差更低，而且逐样本 JSD 分布也更集中，JSD 中位数为 0.007，而三类零样本提示词的中位数分别为 0.049、0.049 和 0.083。分布微调不仅提升了平均拟合效果，也降低了不同测试样本之间的预测波动，可靠的“硅基受访者”不能只在少数样本上偶然接近真实分布，而应在不同群体和不同历史状态下都保持与真实样本的一致。从图 3 可以看出，提示词工程可以影响模型“如何扮演受访者”，却难以真正地模拟“硅基受访者”。FT-KL 以真实回答分布为监督信号，使模型学习不同选项之间的比例关系，可真正地模拟“硅基受访者”。

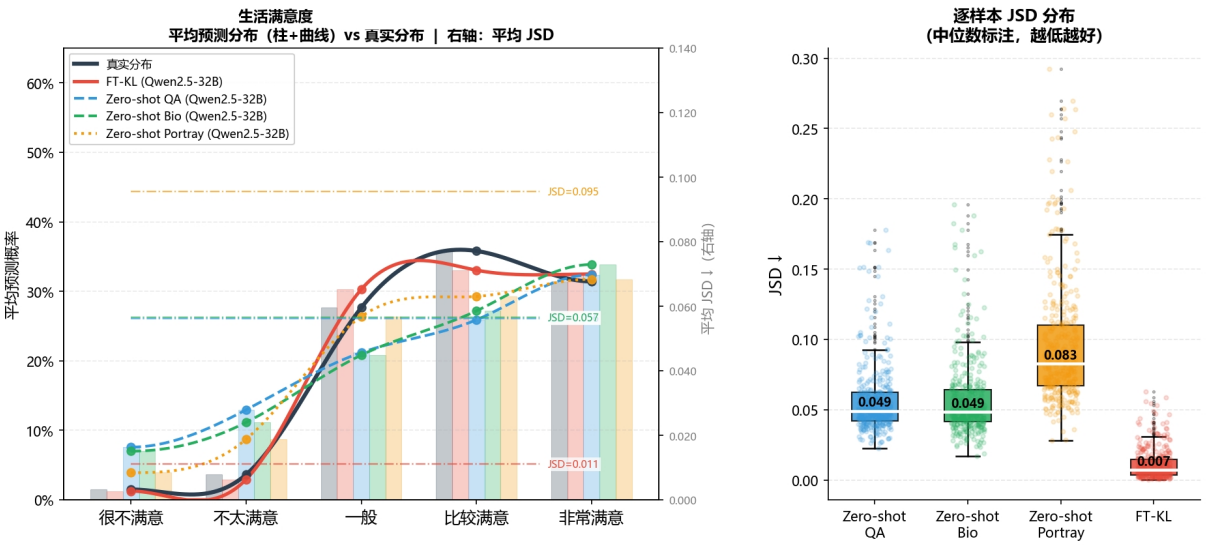


图 3 生活满意度（任意较早波次预测较晚波次）任务中不同方法的比较

表 5 为不同方法不同模型在 CFPS 生活满意度（任意较早波次预测较晚波次）的测试结果。可以看到，Kullback-Leibler 散度约束下的低秩适应有监督微调（FT-KL）在所有模型规模上均取得统计显著的改善（Wilcoxon 配对检验，$p<0.001$），95%bootstrap 置信区间不重叠。无论是 Qwen2.5 系列还是 Qwen3 系列，经过分布约束微调后，模型的 JSD 都明显低于 QA、BIO 和 PORTRAY 三类提示词基线。Qwen2.5-0.5B 的 FT-KL JSD 为 0.0153，而同模型下最好的零样本提示词 JSD 仅为 0.2402；Qwen2.5-3B 的 FT-KL JSD 为 0.0234，而最好的零样本提示词 JSD 仅为 0.2163。即便是零样本表现较强的 Qwen2.5-32B，其最佳提示词 JSD 为 0.0869，仍明显高于 FT-KL 的 0.0222。这说明，真实面板调查分布的监督信号具有实质性作用，提高了模型生成硅基样本的能力。

表 5 生活满意度（任意较早波次预测较晚波次）测试集结果

模型	方法	JSD↓	KL↓	L1↓	EMD↓
Qwen2.5-0.5B	FT-KL	**0.0153**	**0.0623**	**0.2691**	**0.1932**
	QA 式零样本	0.3023	1.4223	1.2447	1.7434
	BIO 式零样本	0.2534	1.1372	1.1504	1.5193
	PORTRAY 式零样本	0.2402	1.0395	1.0845	1.5343
Qwen2.5-3B	FT-KL	**0.0234**	**0.0994**	**0.3391**	**0.2370**
	QA 式零样本	0.2753	1.3346	1.2317	1.4592
	BIO 式零样本	0.2346	1.2873	1.0751	1.1245
	PORTRAY 式零样本	0.2163	1.0599	1.0494	1.1450
Qwen2.5-14B	FT-KL	**0.0707**	**0.3060**	**0.5605**	**0.5598**
	QA 式零样本	0.2832	3.6898	1.1682	0.9971
	BIO 式零样本	0.3127	4.4432	1.2594	1.0303
	PORTRAY 式零样本	0.3077	4.4378	1.2548	1.0215
Qwen2.5-32B	FT-KL	**0.0222**	**0.0934**	**0.3297**	**0.2302**
	QA 式零样本	0.0917	0.3631	0.6529	0.6467
	BIO 式零样本	0.0869	0.3492	0.6412	0.6242
	PORTRAY 式零样本	0.1570	0.7790	0.9386	0.7967
Qwen2.5-72B	FT-KL	**0.0357**	**0.1567**	**0.4477**	**0.3354**
	QA 式零样本	0.1242	0.4605	0.7364	0.9710

	BIO 式零样本	0.1019	0.4380	0.6817	0.6510
	PORTRAY 式零样本	0.0966	0.4382	0.6697	0.6100
	FT-KL	**0.0238**	**0.1033**	**0.3397**	**0.2383**
Qwen3-8B	QA 式零样本	0.2841	2.1124	1.2184	1.1037
	BIO 式零样本	0.2740	2.3144	1.1896	0.9799
	PORTRAY 式零样本	0.2650	2.3401	1.1569	0.9705
	FT-KL	**0.0276**	**0.1185**	**0.3937**	**0.2629**
Qwen3-32B	QA 式零样本	0.1049	0.5334	0.7499	0.6306
	BIO 式零样本	0.1873	1.4383	0.9609	0.7750
	PORTRAY 式零样本	0.1577	1.1254	0.8985	0.7028

注：JSD、KL、L1 和 EMD 均为距离型指标，数值越小表示模型预测分布越接近真实调查分布。加粗表示同一模型内四种方法中的最优结果。三类零样本提示词均不包含历史真实回答分布示例。FT-KL 在本文中指基于低秩适应（LoRA）的 KL 散度约束微调。

从表 6 可以看到，生活满意度（只看连续两波之间的变化）任务中，FT-KL 在所有模型规模上均明显优于三类零样本人口学提示词。这一任务只考察连续两波之间的短期变化，例如 2020 年到 2022 年的生活满意度分布变化，因此它最接近真实面板研究中“根据上一波状态预测下一波状态”的任务形式。在这一较干净的短期追踪场景中，仅依靠 QA、BIO 或 PORTRAY 式身份提示，仍不足以恢复真实受访者群体的未来回答分布。相反，将真实调查分布作为监督信号进行 FT-KL 微调，可以显著降低模型预测分布与真实分布之间的距离。

从具体模型看，Qwen2.5-32B 的 FT-KL 表现最好，JSD 为 0.0111；Qwen2.5-0.5B、Qwen3-32B、Qwen3-8B 和 Qwen2.5-3B 也都处在较低 JSD 区间，分别为 0.0135、0.0138、0.0153 和 0.0154。相比之下，即使是零样本提示词表现最好的 Qwen2.5-32B，其 QA 式零样本 JSD 也为 0.0562，仍显著高于同模型 FT-KL 的 0.0111。对于较小模型，这一差距更为明显：Qwen2.5-0.5B 微调后的 JSD 为 0.0135，而同模型下最好的 PORTRAY 式零样本仍为 0.2270。Qwen2.5-3B 微调后的 JSD 为 0.0154，而同模型下最好的 BIO 式零样本为 0.1632。这说明，分布监督带来的增益并不只是大模型才具备的能力，小模型经过真实调查分布校准后也能显著改善预测结果。

表 6 生活满意度（只看连续两波之间的变化）测试集结果

模型	方法	JSD↓	KL↓	L1↓	EMD↓
	FT-KL	**0.0135**	**0.0559**	**0.2383**	**0.1656**
Qwen2.5-0.5B	QA 式零样本	0.2847	1.3502	1.1604	1.6037
	BIO 式零样本	0.2337	1.0478	1.0593	1.3839
	PORTRAY 式零样本	0.2270	0.9814	1.0142	1.4067
	FT-KL	**0.0154**	**0.0669**	**0.2496**	**0.1669**
Qwen2.5-3B	QA 式零样本	0.2214	1.0269	1.0657	1.2241
	BIO 式零样本	0.1632	0.8811	0.8477	0.8252
	PORTRAY 式零样本	0.1696	0.8130	0.8960	0.9391
	FT-KL	**0.0644**	**0.2724**	**0.4640**	**0.4805**
Qwen2.5-14B	QA 式零样本	0.2081	2.8671	0.9544	0.7251
	BIO 式零样本	0.2260	3.3111	1.0201	0.7322
	PORTRAY 式零样本	0.2174	2.9519	1.0072	0.7157
	FT-KL	**0.0111**	**0.0463**	**0.2083**	**0.1476**
Qwen2.5-32B	QA 式零样本	0.0562	0.2220	0.4690	0.4163

Qwen2.5-72B	BIO 式零样本	0.0565	0.2365	0.4854	0.4028
	PORTRAY 式零样本	0.0955	0.5022	0.7087	0.4936
	FT-KL	**0.0336**	**0.1557**	**0.3775**	**0.2736**
Qwen3-8B	QA 式零样本	0.1022	0.3572	0.6161	0.8161
	BIO 式零样本	0.0683	0.3085	0.5098	0.4145
	PORTRAY 式零样本	0.0617	0.2750	0.4973	0.3823
	FT-KL	**0.0153**	**0.0668**	**0.2537**	**0.1665**
Qwen3-32B	QA 式零样本	0.2004	1.4596	0.9838	0.8051
	BIO 式零样本	0.1897	1.7751	0.9433	0.6594
	PORTRAY 式零样本	0.1740	1.5282	0.8934	0.6447
	FT-KL	**0.0138**	**0.0591**	**0.2382**	**0.1637**
	QA 式零样本	0.0735	0.3999	0.6001	0.4091
	BIO 式零样本	0.1415	1.2497	0.7852	0.5332
	PORTRAY 式零样本	0.1314	1.0473	0.7901	0.5205

备注：测试集为 2020→2022。JSD、KL、L1 和 EMD 均为距离型指标，数值越小表示模型预测分布越接近真实调查分布。加粗表示同一模型内四种方法中的最优结果。三类零样本提示词均不包含历史真实回答分布示例。FT-KL 在本文中指基于低秩适应（LoRA）的 KL 散度约束微调。

这说明，硅基受访者的可靠性不能只建立在角色扮演式提示词上，而需要经验分布校准。QA、BIO 和 PORTRAY 三类提示词虽然都通过不同方式向模型提供受访者身份信息，但这些信息主要改变模型作答时的叙述框架，却难以让模型在 A、B、C、D、E 五个选项之间形成接近真实 CFPS 分布的概率结构。

FT-KL 把真实 CFPS 生活满意度分布作为监督信号，直接约束模型在选项词元上的概率分布。如果真实样本中有相当比例的受访者选择“一般”或“比较满意”，模型就不能只把概率集中到单一最典型选项上；如果真实群体内部存在分歧，模型也必须在概率空间中保留这种分歧。因此 FT-KL 更符合社会调查研究关注“群体回答分布”而非“单个代表性回答”的方法逻辑。

表 5 和表 6 的结果说明，模型规模与分布拟合能力并不存在简单的单调关系。任务中表现最好的不是 72B 模型，而是 Qwen2.5-0.5B，其 JSD 为 0.0153，Qwen2.5-32B 和 Qwen3-8B 也表现较好，JSD 分别为 0.0222 和 0.0238。相反，Qwen2.5-14B 的 FT-KL 的 JSD 为 0.0707，明显弱于多个小模型和中等规模模型。对于这类具有明确监督信号的社会调查任务，数据构造、训练目标、模型适配性和验证集选择比单纯扩大模型规模更关键，问卷分布预测能力不能简单等同于参数规模。

（三）主观健康

主观健康受到年龄增长、疾病冲击和身体状态的变化影响，比生活满意度更难预测。主观健康的跨期变化并不只是一般态度变化，也包含更强的生命历程和身体状态约束。主观健康任务为本文提供了一个比生活满意度更具“客观约束”的主观评价场景。图 4 以 Qwen2.5-32B 为例，展示了主观健康（任意较早波次预测较晚波次）任务的分布比较。从左图来看，相比零样本提示词，FT-KL 的预测分布与真实分布形状更为接近，说明真实调查分布监督能够有效校准模型在健康自评选项上的概率结构。从右

图来看，FT-KL 的平均 JSD 为 0.022，明显低于 QA 式零样本、BIO 式零样本和 PORTRAY 式零样本。这说明分布微调不仅在平均意义上更接近真实分布，也在不同群体样本之间表现出更稳定的误差控制。

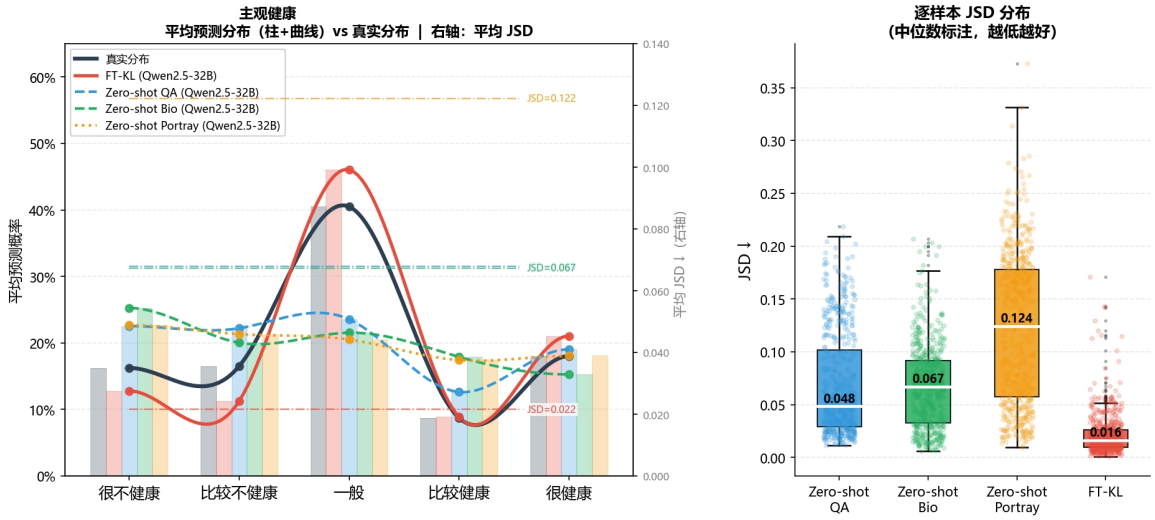


图 4 主观健康（任意较早波次预测较晚波次）任务中不同方法的比较

表 7 表示，在主观健康任意较早波次预测较晚波次任务中，FT-KL 相较三类零样本基线均取得统计显著的改善（Wilcoxon 配对检验，$p<0.001$），95%bootstrap 置信区间不重叠。由此可见，即使在主观健康这一更受身体状态约束的变量上，仅靠人口学身份提示仍不足以恢复真实 CFPS 回答分布，真实调查分布监督仍具有明显优势。

表 7 主观健康（任意较早波次预测较晚波次）测试集结果

模型	方法	JSD↓	KL↓	L1↓	EMD↓
Qwen2.5-0.5B	FT-KL	**0.0258**	**0.1116**	**0.3470**	**0.3267**
	QA 式零样本	0.2111	1.0533	1.0718	1.1970
	BIO 式零样本	0.1636	0.7848	0.9057	0.9671
	PORTRAY 式零样本	0.1636	0.7489	0.9137	1.0389
Qwen2.5-3B	FT-KL	**0.0319**	**0.1395**	**0.3921**	**0.3697**
	QA 式零样本	0.2105	1.1088	1.0843	1.1108
	BIO 式零样本	0.1939	1.1250	1.0162	0.8834
	PORTRAY 式零样本	0.2119	1.1925	1.0763	1.0501
Qwen2.5-14B	FT-KL	**0.0827**	**0.3620**	**0.6578**	**0.6976**
	QA 式零样本	0.2999	5.0975	1.2462	1.0185
	BIO 式零样本	0.3027	4.5715	1.2596	1.0221
	PORTRAY 式零样本	0.3019	4.4353	1.2611	1.0261
Qwen2.5-32B	FT-KL	**0.0303**	**0.1346**	**0.3923**	**0.3360**
	QA 式零样本	0.0943	0.4480	0.6930	0.5532
	BIO 式零样本	0.0924	0.4159	0.6922	0.5372
	PORTRAY 式零样本	0.1651	0.8626	0.9617	0.7456
Qwen2.5-72B	FT-KL	**0.0346**	**0.1518**	**0.4362**	**0.3608**
	QA 式零样本	0.1126	0.5529	0.7529	0.5785
	BIO 式零样本	0.1298	0.6562	0.8181	0.6053
	PORTRAY 式零样本	0.1430	0.7043	0.8763	0.6720
Qwen3-8B	FT-KL	**0.0307**	**0.1307**	**0.3875**	**0.3679**

	QA 式零样本	0.2900	3.0299	1.2332	1.0188
	BIO 式零样本	0.2902	3.2521	1.2338	0.9798
	PORTRAY 式零样本	0.2916	3.4801	1.2328	0.9741
	FT-KL	**0.0330**	**0.1446**	**0.4077**	**0.3824**
Qwen3-32B	QA 式零样本	0.2330	1.9029	1.0854	0.8487
	BIO 式零样本	0.2631	2.6449	1.1493	0.9590
	PORTRAY 式零样本	0.2494	2.1047	1.1205	1.0424

备注：JSD、KL、L1 和 EMD 均为距离型指标，数值越小表示模型预测分布越接近真实调查分布。加粗表示同一模型内四种方法中的最优结果。三类零样本提示词均不包含历史真实回答分布示例。

表 8 进一步表明，在主观健康“只看连续两波之间的变化”任务中，FT-KL 表现更优。该任务只考察连续两波之间的变化，例如 2020 年到 2022 年的健康自评变化，因此更接近短期健康状态追踪。结果显示，Qwen2.5-3B 的 FT-KL 表现最好。这说明，在短期健康自评预测场景中，分布微调并非只改善某一个模型，而是在多个模型规模上都能稳定降低分布误差。

表 8 主观健康（只看连续两波之间的变化）测试集结果

模型	方法	JSD↓	KL↓	L1↓	EMD↓
	FT-KL	**0.0225**	**0.0927**	**0.3008**	**0.2924**
Qwen2.5-0.5B	QA 式零样本	0.2092	1.0477	1.0409	1.1322
	BIO 式零样本	0.1638	0.7909	0.8808	0.9395
	PORTRAY 式零样本	0.1637	0.7517	0.8867	0.9858
	FT-KL	**0.0214**	**0.0907**	**0.2995**	**0.2710**
Qwen2.5-3B	QA 式零样本	0.1879	0.9608	1.0062	1.0022
	BIO 式零样本	0.1624	0.9366	0.8982	0.7413
	PORTRAY 式零样本	0.1837	1.0087	0.9724	0.9223
	FT-KL	**0.0621**	**0.2813**	**0.5231**	**0.5110**
Qwen2.5-14B	QA 式零样本	0.2492	4.4287	1.0996	0.8581
	BIO 式零样本	0.2514	3.9569	1.1092	0.8585
	PORTRAY 式零样本	0.2467	3.6055	1.1030	0.8544
	FT-KL	**0.0216**	**0.0927**	**0.3102**	**0.2765**
Qwen2.5-32B	QA 式零样本	0.0678	0.3190	0.5818	0.4152
	BIO 式零样本	0.0674	0.3025	0.5805	0.3902
	PORTRAY 式零样本	0.1222	0.6346	0.8175	0.5817
	FT-KL	**0.0276**	**0.1216**	**0.3786**	**0.3226**
Qwen2.5-72B	QA 式零样本	0.0956	0.4919	0.6700	0.4718
	BIO 式零样本	0.1071	0.5457	0.7275	0.4871
	PORTRAY 式零样本	0.1120	0.5433	0.7636	0.5216
	FT-KL	**0.0255**	**0.1036**	**0.3341**	**0.3162**
Qwen3-8B	QA 式零样本	0.2368	2.4931	1.0774	0.8538
	BIO 式零样本	0.2319	2.5335	1.0667	0.8022
	PORTRAY 式零样本	0.2323	2.8196	1.0616	0.7919
	FT-KL	**0.0232**	**0.1009**	**0.3278**	**0.2770**
Qwen3-32B	QA 式零样本	0.1838	1.5487	0.9283	0.6829
	BIO 式零样本	0.2071	2.1146	0.9782	0.7524
	PORTRAY 式零样本	0.1944	1.6201	0.9569	0.8400

备注：测试集为 2020→2022。JSD、KL、L1 和 EMD 均为距离型指标，数值越小表示模型预测分布越接近真实调查分布。加粗表示同一模型内四种方法中的最优结果。三类零样本提示词均不包含历史真实回答分布示例。

与生活满意度相比，主观健康任务中的最佳 JSD 整体高于生活满意度任务，说明主观健康回答分布的预测难度更大。这可能是因为，健康自评虽然同样属于主观变量，但它更直接受到年龄增长、疾病状况、身体机能变化和突发生活事件等因素影响，其跨期变化并不完全能够由上一波健康状态和基本人口学变量解释。因此，在主观健康任务中，即使经过分布约束微调，模型预测结果与真实分布之间仍存在一定差距。

主观健康实验也进一步说明，提示词工程的局限并不只存在于生活满意度这一总体评价变量中。QA、Bio 和 Portray 三类零样本提示词虽然能够通过不同方式改变模型对受访者身份和回答情境的描述，但仍难以使模型准确把握健康自评各选项之间的真实比例结构。这表明，单纯依靠提示词让模型“扮演”特定受访者，并不能稳定生成接近真实调查数据的群体回答分布。

图 5 逐选项残差分析显示，零样本方法（尤其是 PORTRAY 式零样本）对中间选项（“一般”“比较满意”）存在系统性高估，对尾部选项（“很不满意”“很不健康”）存在明显低估，呈现出典型的分布坍缩（distributional collapse）模式。FT-KL 在所有选项上的残差显著小于零样本方法。

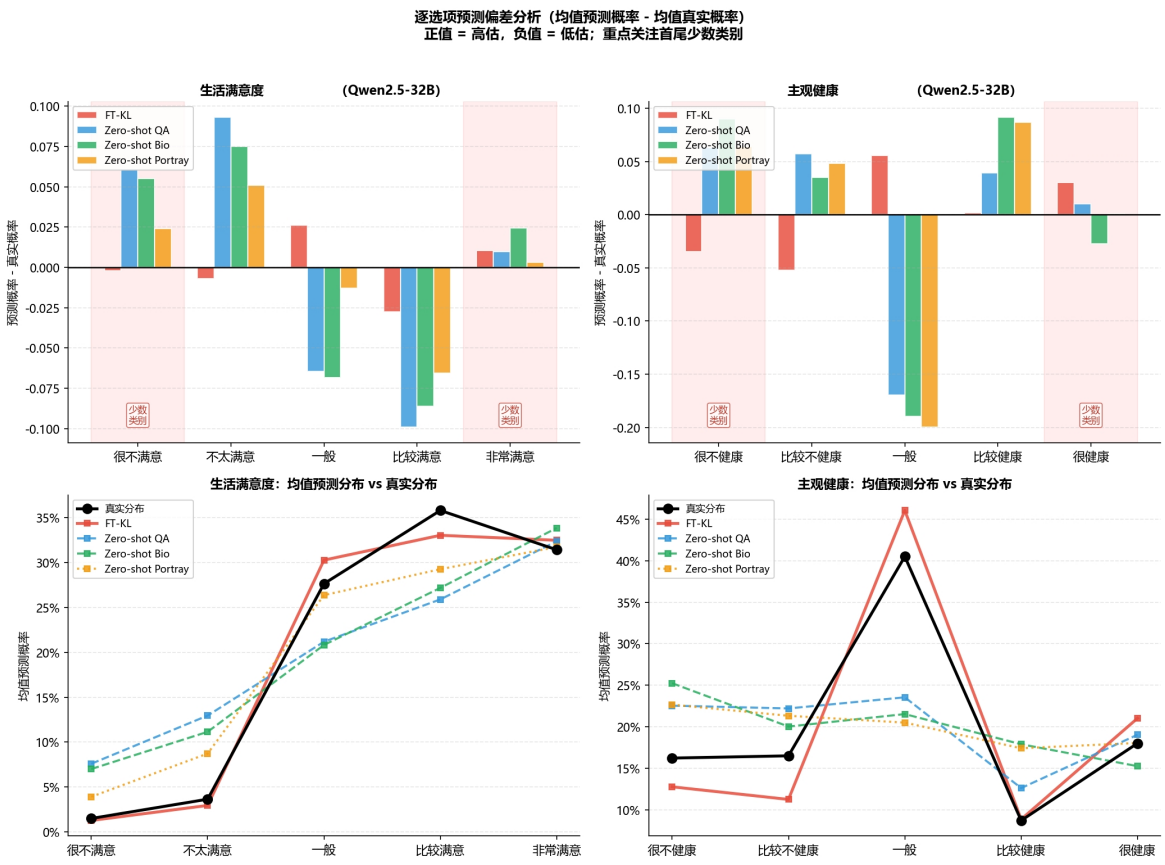


图 5 逐选项预测偏差分析。上方为残差图（均值预测概率–均值真实概率），正值高估，负值低估，红底区域为首尾少数类别；下方为均值预测分布与真实分布的对比。

综合来看，主观健康结果与生活满意度结果方向一致，说明大语言模型在真实面板数据监督下，不仅可以学习总体生活评价的跨期分布规律，也可以在一定程度上扩展到更受生命历程和身体状态影响的主观健康评价。由此可见，大语言模型并不是通过提示词自然获得可靠的“受访者能力”，而是需要在真实调查分布的约束和校准下，才可能在不同主观变量上形成更接近人类群体的回答分布。

五、结论

本文基于 CFPS2010—2022 年真实面板数据，构建生活满意度与主观健康回答分布预测任务，系统比较了三类零样本人口学提示词方法与 FT-KL 分布微调方法的预测表现。研究发现，在“只看连续两波之间的变化”和“任意较早波次预测较晚波次”两类任务结构下，FT-KL 均取得统计显著的改善（Wilcoxon 配对检验，$p<0.001$），95%bootstrap 置信区间不重叠。无论是生活满意度还是主观健康，经过真实调查分布监督后，模型预测分布在 JSD、KL、L1 和 EMD 等指标上均更接近真实 CFPS 回答分布。这一结果说明，在真实个人面板场景中，历史主观状态与人口学属性所包含的信息，可以通过真实调查分布监督转化为较稳定的未来回答分布预测能力，从而更好地保留不同选项之间的比例结构和群体内部异质性。

本文的理论与方法意义主要体现在两个方面。第一，既有研究普遍认为，未经校准的大语言模型在生成硅基样本时容易存在代表性不足、异质性压缩和结构性偏差等问题。本文进一步证明，如果引入真实调查分布监督，并在分布层面对模型进行校准，大语言模型可以在明确边界内生成更接近真实调查分布的硅基受访者。第二，本文将大模型问卷模拟的评价标准从“个体回答是否像人”推进到“群体回答分布是否接近真实调查”，为硅基样本研究提供了一种可量化、可比较和可复现的评估路径。

从循证决策视角看，经真实社会调查分布校准的大语言模型并不能替代正式调查和真实民意收集，但可以在证据尚不充分、政策情境需要快速判断时，为研究者和决策者提供一种可评估、可迭代的先验参照。对政府治理而言，经真实调查数据校准的硅基受访者主要可以服务于两个环节：一是在问卷预实验中，提前模拟不同群体对题项和选项的回答分布，帮助研究者和政府部门发现题目理解偏差、选项设置不当和群体差异被遮蔽等问题；二是在政策试点前，基于既有调查数据对不同群体的可能态度反应进行预评估，为试点方案优化、重点群体识别和后续正式调查设计提供辅助参考。

本文仍存在一定局限。第一，本文主要以生活满意度和主观健康两个主观变量为预测对象，尚未覆盖社会信任、公平感、政策态度等更多类型的调查题项，未来研究可进一步检验该方法在不同议题和量表结构中的适用性。第二，本文关注的是群体回答分布预测，而非单个受访者未来回答的精确预测，因此其结论主要适用于问卷预实验、群体态度模拟和政策预评估等场景。第三，本文模型训练依赖历史调查数据，当未来出现重大政策变化、公共事件或社会结构冲击时，历史分布规律可能难以完全解释新的

态度变化。因此，硅基受访者应被定位为真实社会调查的辅助工具，而不能被直接用作代表公众意见的最终证据。

【注释】

① 在微调方法方面，本文采用KL散度约束的低秩适应有监督微调。具体而言，在模型训练框架上，本文采用有监督微调（Supervised Fine-Tuning，SFT）将真实调查中的群体回答分布作为监督信号对预训练大语言模型进行进一步训练，使模型学习特定任务的输入—输出规律，真实更新模型参数，使模型内部概率结构适应目标任务。

② 在参数更新方式上，本文采用低秩适应（Low-Rank Adaptation，LoRA）实现参数高效微调。LoRA会在模型部分权重矩阵旁引入低秩可训练矩阵，仅训练少量新增参数，而冻结原始大模型主体参数，从而显著降低训练成本。

③ 在优化目标方面，本文采用Kullback–Leibler（KL）散度作为损失函数，对模型在问卷选项词元上的完整概率分布进行约束。损失函数决定模型“朝什么方向学习”，即训练过程中如何衡量模型预测与真实数据之间的差异。设某一题项共有K个选项，真实群体回答分布为$P = (p_1, ..., p_K)$，模型预测分布为$Q = (q_1, ..., q_K)$。在实际计算中，为避免零概率导致对数不可定义，本文对P和Q加入极小平滑项后重新归一化，得到P^*和Q^*。本文使用的分布约束为：

$$D_{KL}(P^* \| Q^*) = \sum_{i=1}^{K} p_i^* \log \frac{p_i^*}{q_i^*} \tag{11}$$

该目标函数即 KL 散度，用于约束模型预测分布接近真实调查回答分布，把真实面板调查中的回答分布转化为可学习的监督信号。KL 散度强调如果真实人类群体中有较高比例选择某一选项，模型就不应对该选项赋予过低概率；如果真实回答分布在多个选项之间存在分歧，模型也应在概率空间中保留这种分歧，而不是仅集中于单一最大概率选项，使模型不仅学习“应该输出哪个选项”，还学习“不同选项之间应当如何分配概率”。模型训练过程中以验证集表现选择最佳训练轮次，最终只在测试集上报告结果。

④ 在基础模型方面，本文选取Qwen2.5和Qwen3系列不同参数规模模型进行实验，包括Qwen2.5-0.5B、Qwen2.5-3B、Qwen2.5-14B-Instruct、Qwen2.5-32B、Qwen2.5-72B，以及Qwen3-8B和Qwen3-32B。以检验分布微调效果是否依赖某一个特定规模模型，以及模型规模扩大是否一定带来更好的问卷回答分布拟合能力。

⑤ 在计算资源方面，本文实验在16卡H100 GPU 环境下完成。0.5B、3B、8B和14B模型主要使用单卡H100，32B模型使用二卡H100，72B模型使用四卡H100 GPU。推理和微调过程中采用bf16/fp16精度与自动设备映射，以保证不同参数规模模型能够在统一实验框架下完成训练和比较。

⑥（Wilcoxon配对检验，$p < 0.001$），95% bootstrap置信区间不重叠。括号内为5000次bootstrap重采样计算的95%置信区间。对所有任务和方法组合，采用单尾Wilcoxon符号秩检验验证FT-KL与各零样本基线的差异，在全部12组任务中均达到$p < 0.001$，置信区间与零样本基线完全不重叠。

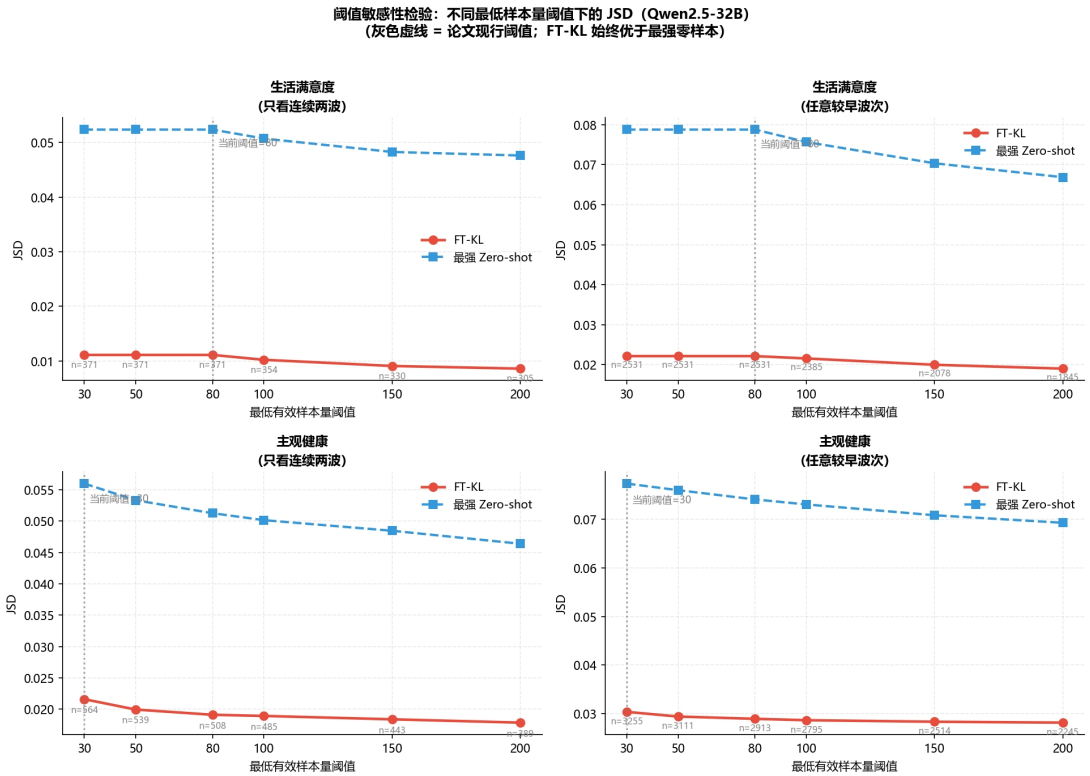
【参考文献】

[1]王天夫. 数字时代的社会变迁与社会研究[J]. 中国社会科学, 2021(12).
[2]梁玉成. 基于生成式大语言模型的“测试社会学”[J]. 探索与争鸣, 2024(11): 17-21.
[3]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[J]. 社会学研究, 2024(4).
[4]周穆之,于璐,耿笑敏,罗岚,刘烨. 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[J]. 智能社会研究, 2025, 4(2): 158-178+251-252.
[5]龚为纲, 黄思源. 整体事实还是有偏样本——基于大语言模型生成数据的测量[J]. 学术月刊, 2025, 57(6): 123-136.
[6]Baan J, Aziz W, Plank B, Fernandez R. Stop measuring calibration when humans disagree[C]//Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. 2022: 1892-1915.
[7]Baan J, Fernández R, Plank B, Aziz W. Interpreting predictive probabilities: Model confidence or human label variation?[C]//Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics. 2024: 268-277.

[8]Xie Y Q, Liang L M, Li S Z, Lu Y F, Xiao Z W, Shi M D, Huang J M, Wang M D, Xie Y. Evaluating the statistical realism of LLM-generated social science data[J]. Proceedings of the National Academy of Sciences, 2026, 123(19): e2538145123.
[9]兰天.数字人格：数字智能时代的人格研究[J].全球传媒学刊,2023,10(3):47-65.
[10]周涛,高馨,罗家德.社会计算驱动的社会科学研究方法[J].社会学研究,2022,37(5):130-155+228-229.
[11]谢天,邱林,李雨暄,罗殷,刘盼.大模型时代的社会科学，何去何从？[J].图书情报知识,2023,40(6):6-9+30.
[12]严洁,林胤志.计算政治学中的生成式人工智能——研究方法、主题与展望[J].世界社会科学,2025,(1):116-144+245.
[13]郑若婷,于文轩,赵昊雪,等.“AI 驱动的社会科学研究与公共治理新范式的构建”高端学术论坛综述[J].公共管理学报,2024,21(1):161-166+175-176.
[14]许安明.计算社会科学引入大模型:驱动机制、理论逻辑与范式新变[J].求索,2025,(3):149-157.
[15]杜娟,李晓义.从大数据到大模型：人工智能如何重塑计算社会科学的研究范式？[J].自然辩证法研究,2026,42(5):42-52.
[16]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[J].世界经济与政治,2024,(7):3-30+153.
[17]华杰,李婷.生成式人工智能与社会科学研究范式拓展[J].南开学报(哲学社会科学版),2026,(1):44-58.
[18]Argyle L P, Busby E C, Fulda N, Gubler J R, Rytting C, Wingate D. Out of one, many: Using language models to simulate human samples[J]. Political Analysis, 2023, 31(3): 337-351.
[19]胡安宁.洞中流影：基于大语言模型的硅基样本在社会科学研究中的应用反思[J].社会发展研究,2025,12(1):22-39+242-243.
[20]Aher G V, Arriaga R I, Kalai A T. Using large language models to simulate multiple humans and replicate human subject studies[C]//Proceedings of the 40th International Conference on Machine Learning. PMLR, 2023, 202: 337-371.
[21]刘泽民,陈向东.教育研究的硅基样本——基于大模型技术路线的分析[J].远程教育杂志,2025,43(1):19-32+45.
[22]林喜芬,付张祎.民众对司法公正的主观感受——基于硅基样本的实验研究[J].浙江社会科学,2025,(5):46-60+157-158.
[23]梁玉成,马昱堃.生成式人工智能实践下的社会观念分化探析——基于自主行动者建模的仿真模拟实验研究[J].济南大学学报(社会科学版),2024,34(6):120-132.
[24]周志忍,李乐.循证决策:国际实践、理论渊源与学术定位[J].中国行政管理,2013,(12):23-27+43.
[25]郁俊莉,姚清晨.从数据到证据：大数据时代政府循证决策机制构建研究[J].中国行政管理,2020,(4):81-87.
[26]吕佳龄,温珂.循证决策的协同模式：面向国家治理体系和治理能力现代化的科学与决策关系建构[J].中国科学院院刊,2020,35(5):602-610.
[27]刘光华,赵幸,杨克虎.循证视角下的大数据法治决策证据转化研究[J].图书与情报,2018,(6):32-38.
[28]王学军,王子琦.从循证决策到循证治理:理论框架与方法论分析[J].图书与情报,2018,(3):18-27.
[29]Bail C A. Can generative AI improve social science?[J]. Proceedings of the National Academy of Sciences, 2024, 121(21): e2314021121.
[30]Kozlowski A C, Evans J. Simulating subjects: The promise and peril of artificial intelligence stand-ins for social agents and interactions[J]. Sociological Methods & Research, 2025, 54(3): 1017-1073.
[31]Hewitt L, Ashokkumar A, Ghezae I, Willer R. Predicting results of social science experiments using large language models[J]. Preprint, 2024.
[32]凌宛莹,崔思瞻.复现的限度——“硅基样本”的可能与可为[J].智能社会研究,2025,4(2):179-201+252-253.
[33]Durmus E, Nguyen K, Liao T I, Schiefer N, Askell A, Bakhtin A, Chen C, Hatfield-Dodds Z, Hernandez D, Joseph N, Lovitt L, McCandlish S, Sikder O, Tamkin A, Thamkul J, Kaplan J, Clark J, Ganguli D. Towards measuring the representation of subjective global opinions in language models[EB/OL]. arXiv:2306.16388, 2024.
[34]Santurkar S, Durmus E, Ladhak F, Lee C, Liang P, Hashimoto T. Whose opinions do language models reflect?[C]//Proceedings of the 40th International Conference on Machine Learning. PMLR, 2023.
[35]Lutz M, Sen I, Ahnert G, Rogers E, Strohmaier M. The prompt makes the person(a): a systematic evaluation of sociodemographic persona prompting for large language models[C]//Findings of the Association for Computational Linguistics: EMNLP 2025. 2025: 23212-23237.

[36]Kwok L, Bravansky M, Griffin L D. Evaluating cultural adaptability of a large language model via simulation of synthetic personas[J]. arXiv preprint arXiv:2408.06929, 2024.
[37]Park P S, Schoenegger P, Zhu C. Diminished diversity-of-thought in a standard large language model[J]. Behavior Research Methods, 2024, 56(6): 5754-5770.
[38]Sun S, Lee E, Nan D, Zhao X Y, Lee W B, Jansen B J, Kim J H. Random silicon sampling: Simulating human sub-population opinion using a large language model based on group-level demographic information[J]. arXiv preprint arXiv:2402.18144, 2024.
[39]Bisbee J, Clinton J D, Dorff C, Kenkel B, Larson J M. Synthetic replacements for human survey data? The perils of large language models[J]. Political Analysis, 2024, 32(4): 401-416.
[40]郭茂灿,李芳.LLM 生成硅基样本的偏差来自何处?——基于中国乡村社会大调查数据的实证分析[J].浙江社会科学,2026,(2):123-136+159.
[41]Cao Y, Liu H, Arora A, Augenstein I, Röttger P, Hershcovich D. Specializing large language models to simulate survey response distributions for global populations[C]//Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies. 2025: 3141-3154.
[42]Suh J, Jahanparast E, Moon S, Kang M W, Chang S. Language model fine-tuning on scaled survey data for predicting distributions of public opinions[C]//Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics. 2025: 21147-21170.
[43]谢宇,胡婧炜,张春泥.中国家庭追踪调查:理念与实践[J].社会,2014,34(2):1-32.

附录 A 稳健性检验



附图 1 阈值敏感性检验。横轴为最低有效样本量阈值，在 30 至 200 范围内，FT-KL 在所有四个任务上均持续优于最强零样本基线，结论稳健。

如附图 1 所示，我们对最低有效样本量阈值 30、50、80、100、150、200 逐一重新计算了主要结果。在所有阈值设置下，FT-KL 表现更优，两条曲线在任何阈值下均不交叉，结论完全稳健。