第5卷 第3期 计 量 经 济 学 报 Vol. 5, No. 3
2025年5月 China Journal of Econometrics May, 2025

doi: 10.12012/CJoE2025-0089

基于多个 LLM 代理的经济和公共政策分析框架

郝俞植 [1], 谢丹阳 [2]

(1. 香港科技大学经济学系, 香港; 2. 香港科技大学 (广州) 社会枢纽创新创业与公共政策学域, 广州 510006)

摘 要 本文开创了一种利用多个大语言模型 (LLMs) 作为异质性代理人的经济和公共政策分析新方法. 我们首先在两种不同场景下评估了主流 LLMs 解决两期消费分配问题的经济决策能力: 一种基于显式效用函数, 另一种依靠直觉推理. 不同于先前仅通过调整提示词模拟代理异质性的研究, 我们创新性地利用不同 LLMs 之间固有的分析能力差异来模拟具有不同认知特征的经济主体. 基于这些发现, 我们构建了基于多个 LLM 代理的 (multi-LLM-agent-based, MLAB) 框架, 将各种 LLMs 映射到特定教育群体及其对应收入阶层. 通过利息收入税政策案例研究, 我们展示了 MLAB 框架如何模拟政策对不同群体的差异化影响, 为经济和公共政策分析开辟了富有前景的新路径.

关键词 大语言模型; 基于代理人的模型; 异质性代理人; 经济决策; 公共政策分析

A Multi-LLM-Agent-Based Framework for Economic and Public Policy Analysis

HAO Yuzhi[1], XIE Danyang[2]

(1. Department of Economics, The Hong Kong University of Science and Technology, Hong Kong, China; 2. Thrust of Innovation, Policy, and Entrepreneurship, the Society Hub, The Hong Kong University of Science and Technology (Guangzhou), Guangzhou 510006, China)

Abstract This paper pioneers a novel approach to economic and public policy analysis by leveraging multiple large language models (LLMs) as heterogeneous artificial economic agents. We first evaluate five LLMs' economic decision-making capabilities

收稿日期: 2025-03-04
作者简介: 郝俞植, 博士研究生, 研究方向: 市场结构与经济增长理论, 动态博弈理论, 宏观经济学, 人工智能应用, E-mail: yuzhi.hao@connect.ust.hk; 谢丹阳, 香港科技大学 (广州) 社会枢纽创新创业与公共政策学域讲席教授, E-mail: dxie@hkust-gz.edu.cn.
我们感谢戴开颜提供的技术支持, 以及清华大学公共管理学院、中国科学院数学与系统科学研究院预测科学研究中心、中国科学院大学经济与管理学院、厦门大学邹至庄经济研究院、厦门大学-中国科学院计量建模与经济政策研究基础科学中心的研讨会参与者提供的有益建议和评论. 文责自负.

in solving two-period consumption allocation problems under two distinct scenarios: With explicit utility functions and based on intuitive reasoning. While previous research has often simulated heterogeneity by solely varying prompts, our approach harnesses the inherent variations in analytical capabilities across different LLMs to model agents with diverse cognitive traits. Building on these findings, we construct a multi-LLM-agent-based (MLAB) framework by mapping these LLMs to specific educational groups and corresponding income brackets. Using interest income taxation as a case study, we demonstrate how the MLAB framework can simulate policy impacts across heterogeneous agents, offering a promising new direction for economic and public policy analysis by leveraging LLMs' human-like reasoning capabilities and computational power.

Keywords large language models; agent-based modeling; heterogeneous agents; economic decision-making; public policy analysis

1 引言

将大语言模型整合到经济分析中正为政策研究和计算经济学带来革命性的可能性. 传统宏观经济模型虽在数学上严谨优美, 但通常依赖理性预期和代表性代理人等强假设. 这些模型通过抽象的数学公式概括复杂经济过程, 比如谢丹阳和周泽茜 (2019) 在探讨增长理论的变迁时提到, 经济增长理论模型往往通过抽象的 "生产函数" 来描述研发活动、人力资本积累和产品演化, 存在较大的局限, 无法充分捕捉现实经济中的丰富动态特征. 更重要的是, 模型的复杂性与可处理性之间存在根本性权衡, 这一直限制了我们在经济模型中引入有意义的代理异质性.

解决这一问题的尝试有很多. Aiyagari (1994) 和 Krusell and Smith (1998) 的开创性工作, 以及后来 Galí et al. (2004)、Bilbiie (2008) 和 Kaplan et al. (2018) 发展的 TANK (两类代理人新凯恩斯模型) 和 HANK (异质性代理人新凯恩斯模型) 等研究, 都在一定程度上克服了这一限制. 然而, 这些模型通常需要高度简化的设定才能保持可计算性, 从而仍然无法充分反映现实经济中观察到的丰富异质性.

基于代理人的模型 (ABMs) 作为经济建模的另一种方法由来已久. 这类模型允许研究者引入具有有限理性的异质性代理人, 但始终未能成为主流研究范式. 正如 Tesfatsion (2006) 在其编撰的手册中所记录的, 早期的 ABMs 主要依赖预设的行为规则来驱动代理人行为. 这一特点在 Brock and Hommes (1998) 的关于资产定价和股市波动的具有影响力的研究中表现得尤为明显. 虽然这些模型为理解个体互动如何产生宏观现象提供了重要见解, 但也不可避免地受到行为规则刚性的限制.

近年来, 人工智能技术的快速发展, 尤其是大语言模型的出现, 为提升基于代理人建模的灵活性和复杂性开创了新的可能. 这一发展契合了谢丹阳和周泽茜 (2019) 提出的观点, 即在政策分析中发展 "算法模型", 以处理行为经济学和行为金融学中讨论的各类非理性决策偏差. 这些行为偏差在 "算法模型" 框架下将更容易得到处理.

学术界对 LLMs 在经济分析中的潜力探索正呈现蓬勃发展之势. 相关研究展现了多样化的发现, 其中值得注意的有, Mei et al. (2024) 与 Ma (2024) 的研究虽然都关注 LLMs 在

经济决策场景中的表现, 却得出了不同的认识: Mei 等人发现 ChatGPT-4 在经典心理学问卷中的表现 (五大性格特质) 与随机人类样本在统计上难以区分, 但在经典行为博弈中倾向于更加利他和合作; Ma 则指出大语言模型的亲社会行为与人类存在显著差异, 模型间表现出较大变异性, 仅赋予人类身份认同并不能产生真正的人类化行为. 这种发现的差异凸显了 LLMs 行为的复杂性与评估方法选择的重要性. Xie et al. (2024) 则通过实证研究表明, 不同 AI 聊天机器人在博弈中表现出明显不同的行为模式, 证实这些 AI 系统在决策模式上存在显著异质性. Chen et al. (2023) 提供了 LLMs 在特定选择任务中能够呈现经济理性的证据, 而 Horton (2023) 则系统评估了它们在经典行为经济学博弈中的表现. 在将 LLMs 应用于基于代理人建模的开创性工作中, Li et al. (2024) 通过改变提示词中的经济情景来构建异质性代理人, 成功模拟了市场动态和政策反应. 然而, 他们的方法主要通过提示词变化来捕捉异质性, 同时隐含假设所有代理人具有同质的基础推理能力和行为倾向.

我们的研究通过引入两个维度的异质性推进了这一领域的发展. 我们的创新之处在于不仅通过不同提示词捕捉经济环境的差异, 还通过部署不同 LLMs 来体现推理能力和行为倾向的多样性. 这一方法基于对现实世界经济决策的观察: 真实世界中的决策异质性源自客观环境因素 (如收入、财富、年龄) 和主观个体特征 (如认知能力、文化价值观、风险态度) 的双重影响. 通过选择和部署具有不同分析能力和行为特征的多个 LLMs, 我们的框架能够更全面地反映现实世界经济决策中多层次、多维度的异质性.

LLMs 的应用价值不止于个体决策模拟, 还延伸至政策分析的领域. Wu et al. (2024) 探讨了 LLMs 帮助研究者有效应对政策研究设计中的复杂权衡的可能性, 而庞珣 (2024) 则展示了这些模型在增强因果推断分析和促进跨学科研究合作方面的独特优势. 这些研究表明, LLMs 不仅可以作为经济代理的模拟工具, 还能成为政策评估和制度设计的有力分析辅助. 因此, 我们提出的 MLAB 框架对于理解和预测不同人口群体对政策干预的差异化响应具有特殊价值.

我们的研究为这一快速发展的领域做出了以下贡献. 首先, 我们系统评估了五种领先 LLMs 在解决动态经济优化问题方面的能力, 采用了经典的两期消费-储蓄分析框架. 我们分别在两种不同情境下测试这些模型: 一种提供显式效用函数以评估其计算优化能力, 另一种省略效用函数以检验其自主经济推理水平. 其次, 我们创新性地提出了 MLAB 框架, 将不同 LLMs 基于其展示的能力特征映射至特定社会经济群体. 这种方法独特地利用不同 LLMs 间固有的能力差异来模拟经济决策中的人口异质性.

通过利息收入税的案例研究, 我们展示了 MLAB 框架的实用价值. 我们将不同 LLMs 映射到特定教育背景群体及其对应的收入阶层, 在包含 100 个校准个体的模拟人口中, 观察不同社会经济群体对税收政策变化的响应模式. 研究结果揭示了传统建模方法难以捕捉的行为反应的多样性, 展示了我们方法的独特优势.

与先前研究不同, 我们不是通过修改提示词来模拟异质性 (这类似于要求高智能模型模仿低教育水平者的行为), 而是利用不同 LLMs 本身在分析能力上的自然差异来代表不同教育背景的代理人. 通过将每个 LLM 映射至特定教育群体, 我们丰富了模型对人口统计特征和认知异质性的表达, 从而达到增强政策模拟的真实性的可能效果. 尽管当前 LLMs 与教育群体的对应分配尚未达到理想的科学严谨性或多维度匹配, 但我们的初步发现已证明这种方

法的可行性. 随着 LLMs 技术的持续进步, 我们展望未来将能够向高能力 LLM 提供目标群体的详细特征数据, 进而生成定制代码创建专门化的 LLM 实例. 这些专业 LLM 将能够精确模拟特定群体的思维模式和直觉反应, 最终实现创建任意数量的精细校准代理人, 以进行全面而深入的政策模拟. 这种方法论的演进凸显了 MLAB 框架在经济政策分析中的变革潜力.

论文其余部分按如下结构组织: 第 2 节详述我们的理论框架和实验设计, 并评估不同 LLM 的理性优化能力; 第 3 节检验它们的经济直觉和自主推理表现; 第 4 节全面介绍 MLAB 框架并呈现我们的政策分析结果; 第 5 节总结研究发现并探讨对未来研究和政策应用的启示.

2 评估大语言模型的优化能力: 基于显式效用函数的分析

本节系统评估大语言模型 (LLMs) 在解决标准两期消费-储蓄问题中的表现. 我们首先构建理论框架, 随后详细介绍参数校准方法、实验设计与结果分析, 全面考察 LLMs 在面对含有明确效用函数的经济优化问题时的决策能力.

2.1 理论框架

我们考察一位中国城市中年居民的跨期消费决策问题, 该问题涉及两个等长的生命周期阶段 (各 20 年): 工作期和退休期 (比如, 分别为 40∼59 岁和 60∼79 岁). 代理人的偏好由常数相对风险厌恶 (CRRA) 效用函数表示, 这一函数形式因其良好的分析特性和现实相关性在宏观经济分析中广泛应用. 优化问题可表述为:

$$\max_{c_1,c_2} U = u(c_1) + \beta u(c_2), \tag{1}$$

其中效用函数为:

$$u(c) = \begin{cases} \dfrac{c^{1-\sigma}}{1-\sigma}, & \text{if } \sigma \neq 1, \\ \ln(c), & \text{if } \sigma = 1. \end{cases} \tag{2}$$

代理人面临标准的跨期预算约束:

$$c_1 + \frac{c_2}{(1+r)} = w_0 + y_1 + \frac{y_2}{(1+r)}. \tag{3}$$

这里, c_1 和 c_2 分别表示两期的消费, $\beta \in (0,1)$ 为主观贴现因子, $\sigma > 0$ 为相对风险厌恶系数, r 为实际利率, w_0 为初始财富, y_1、y_2 分别为各期收入.

求解一阶条件得到欧拉方程:

$$u'(c_1) = \beta(1+r)u'(c_2). \tag{4}$$

对于 CRRA 效用函数, 欧拉方程简化为:

$$c_2 = c_1(\beta(1+r))^{\frac{1}{\sigma}}. \tag{5}$$

结合预算约束, 我们获得最优消费的解析表达式:

$$c_1 = \frac{w_0 + y_1 + y_2/(1+r)}{1 + (\beta(1+r))^{\frac{1}{\sigma}}/(1+r)}, \tag{6}$$

$$c_2 = \frac{(w_0 + y_1 + y_2/(1+r))(\beta(1+r))^{\frac{1}{\sigma}}}{1 + (\beta(1+r))^{\frac{1}{\sigma}}/(1+r)}. \tag{7}$$

这些解析表达式揭示了几个关键的经济特性: 首先, 两期消费与终身财富 $(w_0 + y_1 + y_2/(1+r))$ 成正比, 体现了收入效应; 其次, 消费比率 $c_2/c_1 = (\beta(1+r))^{1/\sigma}$ 刻画了跨期替代效应; 第三, 风险厌恶参数 σ 决定了消费平滑程度——当 $\sigma \to \infty$ 时, 消费趋于完全平滑, 而当 $\sigma \to 0$ 时, 消费对利率变化愈发敏感.

2.2 参数校准

为了让模型尽可能贴近现实, 我们使用中国家庭追踪调查 (CFPS) 2018 年数据对模型参数进行校准, 重点关注城市居民样本. 对于收入参数的校准, 我们采用了同时考虑年龄特定收入模式和经济增长的两步方法.

首先, 我们将人口分为六个年龄组 (20∼29 岁、30∼39 岁、40∼49 岁、50∼59 岁、60∼69 岁和 70∼79 岁), 计算每组平均收入 $(\bar{y}_i,\ i = 1, \cdots, 6)$ 和总体平均收入 $(\bar{y})$, 然后得到年龄特定的收入比例:

$$k_i = \frac{\bar{y}_i}{\bar{y}},\ i = 1, \cdots, 6. \tag{8}$$

其次, 为预测未来收入, 我们假设经济以 4% 的年均速率增长. 从当前起 j 个十年后的平均收入水平为:

$$\bar{y}_j^g = \bar{y}(1.04)^{10j},\ j = 0, \cdots, 5. \tag{9}$$

对于从 20∼29 岁开始的代表性个体, 其在特定年龄段的收入可表示为:

$$\bar{y}^{g,i} = \bar{y}_j^g k_i. \tag{10}$$

这一表达式同时捕捉了生命周期收入变化和经济增长效应. 例如, 当个体达到 30∼39 岁时, 其收入为 $\bar{y}^{g,2} = \bar{y}(1.04)^{10}k_2$.

为将六期收入流转化为两期问题参数, 我们计算:

$$y_1 = \bar{y}^{g,3} + \frac{\bar{y}^{g,4}}{(1+r)^{10}}, \tag{11}$$

$$y_2 = \bar{y}^{g,5} + \frac{\bar{y}^{g,6}}{(1+r)^{10}}. \tag{12}$$

为了校准初始财富 (w_0), 我们首先要求解这个六期完整优化问题, 采用与两期模型一致的偏好参数 (β, σ) 和利率 (r), 并假设 20 岁时初始财富为零. 40∼49 岁期初的最优累积储蓄即为所需的 w_0 值.

通过上述校准过程, 我们获得以下参数值:

初始财富 (w_0): 141,598.4 单位

工作期收入 (y_1): 958,189.8 单位

退休期收入 (y_2): 244,103.9 单位

贴现因子 (β): 0.99^{20} (反映新兴市场经济体居民的耐心特质)

风险厌恶系数 (σ): 2 (符合宏观经济学常用设定)

年利率 (r): 2%

我们采用的相对较高贴现因子反映了新兴市场经济体居民在跨期决策中观察到的高耐心度, 风险厌恶系数则遵循宏观经济学常用参数设定. 利率参数结合了 2019–2023 年间的实际贷款市场报价利率 (LPR, 数据来源于中国人民银行) 和同期银行理财产品的实际收益率 (基于中国银行业理财市场报告, 并使用国家统计局 CPI 数据进行通胀调整).

2.3 实验设计

我们的实验评估了五种当前主流的大语言模型: DeepSeek-V3、GPT-4o、Gemini-1.5-pro、Claude-3.5-sonnet 和 Llama-3.1-405B. 为减少随机波动并获得稳健结果, 我们对每个模型进行了 16 次独立试验. 整个实验设计遵循严格的实验步骤, 确保各试验间的一致性.

向每个模型提供的提示词包含三个核心组成部分: 角色扮演设置、经济参数和输出结构要求. 角色扮演部分建立决策背景:

请进入角色扮演模式. 想象你是中国城市地区的一位中年工作成年人, 面临消费决策. 你规划消费时, 需要考虑到你的余生将分为两个长度相等且年数较长的时期: 工作期和随后的退休期[1].

经济参数部分提供效用函数、预算要求和相关数值:

你对这两个时期消费的偏好可描述为: $U = [c_1^{(1-2)} + 0.818 * c_2^{(1-2)}]/(1-2)$, 其中: c_1 是你工作期的消费. c_2 是你退休期的消费.

你当前的经济状况: 当前储蓄: 141,598.4 单位. 工作期收入: 958,189.8 单位. 退休期收入: 244,103.9 单位. 两个时期之间的利率: 48.6%. 你需要量入为出.

输出结构部分要求模型以特定格式回答:

在这些情况下, 你会如何选择这两个生命时期的消费? 请解释你的选择. 最后, 说 “最终答案: 我将选择……因为……”.

实验实施包括以下步骤: 1) 初始化新会话环境确保上下文清晰; 2) 提供标准化提示词; 3) 记录完整响应; 4) 清除对话历史; 5) 重复步骤 1∼4 共 16 次; 6) 依次对每个模型执行相同流程. 为确保试验间的独立性, 我们在每次新提示词前清除对话历史, 确保每个响应仅基于当前提示词而非受到先前交互的影响.

我们保持各模型的默认温度设置, 以保留其自然推理特性:

Deepseek-V3: 温度 = 1.0 (0∼2 范围)

GPT-4o: 温度 = 1.0 (0∼1 范围)

Gemini-1.5-pro: 温度 = 1.0 (0∼2 范围)

Claude-3.5-sonnet: 温度 = 1.0 (0∼1 范围)

Llama-3.1-405B: 温度 = 0.2 (0∼1 范围)

[1]实验中使用的提示词和模型的回应均为英文, 详情可见我们的工作论文 “A Multi-LLM-Agent-Based Framework for Economic and Public Policy Analysis” (arXiv: 2502.16879) (2025).

温度参数控制输出的随机性和创造性, 值越高输出越多样化但可能偏离主题, 值越低输出越确定和保守. 值得注意的是, 虽然 Llama-3.1-405B 的默认温度设置偏低, 但我们的分析显示它在所有模型中表现出最大的决策变异性, 这一现象反映了模型内在特性的复杂性, 而非仅由温度参数所决定.

2.4 实证结果

对大语言模型响应的分析揭示了它们在解决优化问题时展现出显著不同的能力模式. 我们从三个互补视角展示这些结果:

首先, 图 1 展示了各模型的消费选择 (c_1, c_2) 与预算约束线和理论最优点的关系. 图中, 预算约束线上方的点表示过度消费 (不可行解), 而下方的点代表消费不足 (资源利用效率低). 理论最优点由两条虚线的交点标识.

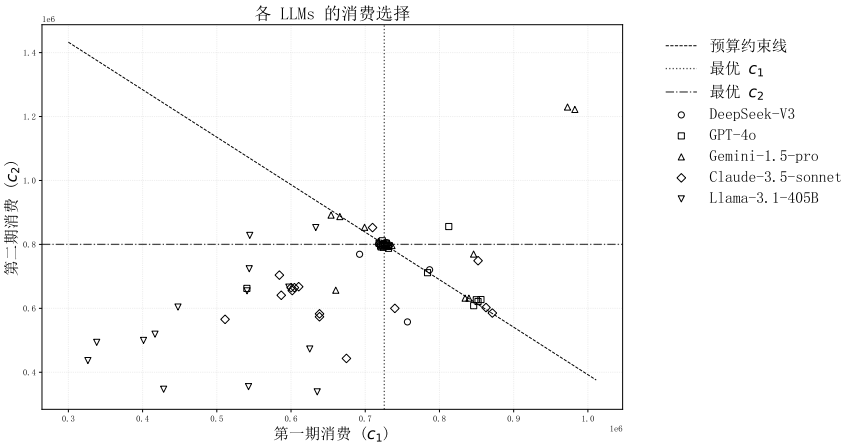

图 1　各 LLMs 的消费选择. 对角线代表预算约束线, 虚线的交点是理论最优点. 每个点代表一次试验响应

DeepSeek-V3 展现了卓越的优化能力, 其大多数决策点聚集在理论最优点附近. 虽然部分解略低于预算约束线 (表明轻微消费不足), 但绝大多数响应代表了高效的资源分配.

GPT-4o 表现中等, 其决策点基本遵守预算约束, 但在最优点周围呈现较大分散性. 少数解位于预算线约束上方, 表明该模型偶尔无法准确处理跨期资源约束.

Gemini-1.5-pro 的表现较为多变, 部分解明显超出预算约束. 然而, 仍有相当数量的点出现在最优解附近, 表明该模型对优化问题有部分理解.

Claude-3.5-sonnet 的解大多低于预算约束线, 表现出系统性消费不足. 其决策点的分散模式反映了对优化原则的掌握不够一致.

Llama-3.1-405B 的表现最弱, 其解呈现广泛分散, 经常远离预算约束和理论最优点. 大量响应显示严重的消费不足, 表明该模型在处理数学优化问题时能力有限.

其次, 图 2 通过箱线图展示了各模型在 c_1 和 c_2 上的分布情况, 提供了模型行为的另一视角. 图中, 方框表示四分位距 (第 25 至 75 百分位), 中线为中位数, 触须延伸至不包含异常值的完整范围.

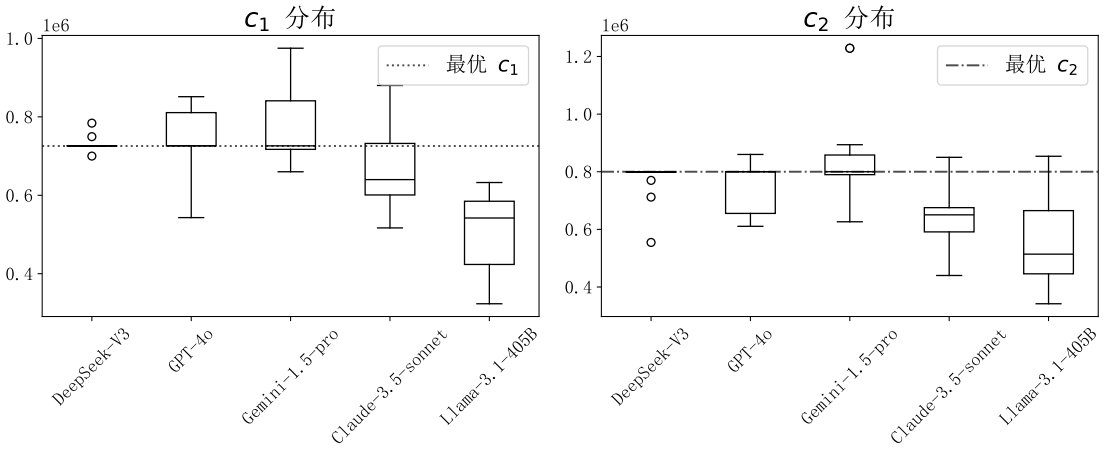

图 2　跨期消费选择分布. 箱体显示四分位距和中位线, 触须延伸至不包括异常值的全范围

DeepSeek-V3 在 c_1 和 c_2 上均显示最为集中的分布, 体现了高度的决策一致性. 相比之下, Llama-3.1-405B 展现最广泛的分散性, 表明即使在相同条件下也产生高度可变的建议. GPT-4o 和 Gemini-1.5-pro 表现出中等程度的变异性, 而 Claude-3.5-sonnet 则倾向于推荐较低消费值且分散性显著.

最后, 表 1 提供了定量评估各模型性能的关键指标, 进一步确认了散点图观察到的模式. 我们采用四项指标: 解的准确性 (在理论最优值 5% 范围内的比例)、平均绝对百分比偏差 (MAPD)、绝对百分比偏差方差 (VarAPD, 用于评估一致性) 以及预算约束遵守度 (通过贴现消费与总资源间的百分比偏差衡量).

DeepSeek-V3 在所有评估指标上表现卓越, 其准确率高达 90.63%, 平均偏差仅 1.94%, 预算约束偏离度不到 1%. 其他模型的表现则呈现明显的层级递减, GPT-4o 和 Gemini-1.5-pro 处于中等水平, 准确率分别为 56.25% 和 50.00%. Claude-3.5-sonnet 的准确率降至 12.50%, 而 Llama-3.1-405B 的表现最弱, 准确率仅为 3.57%, 平均偏差超过 30%, 显示出在解最优性和预算约束满足度上的显著不足.

这些结果揭示了不同大语言模型在处理数学优化问题时能力的显著差异, 这一发现为我们后续利用这些模型的异质性特征构建 MLAB 框架奠定了基础.

表 1　不同 LLMs 的性能比较

模型	准确性 (5% 容差) (%)	平均绝对百分比偏差 (%)	VarAPD ($\times 10^4$)	预算约束偏差 (%)
Deepseek-V3	90.63	1.94	32.61	0.95
GPT-4o	56.25	7.33	81.75	1.96
Gemini-1.5-pro	50.00	10.95	214.76	6.68
Claude-3.5-sonnet	12.50	17.09	90.70	13.26
Llama-3.1-405B	3.57	31.35	266.06	30.66

3 评估大语言模型的经济直觉: 无显式优化指南的决策分析

3.1 修改实验设计: 从数学优化到直觉判断

虽然前一节评估了大语言模型在有明确数学框架指导下的优化能力, 但现实世界中的经济决策甚少通过显式效用最大化进行. 实际决策过程中, 个体往往依赖直觉、经验以及多元的经济与社会考量. 这种基于直觉的决策方式更贴近人类行为, 其中选择受到除却纯粹数学优化外的多种因素影响.

为更真实地模拟现实决策环境, 我们调整了实验设计, 评估大语言模型在无显式优化指南情况下的经济直觉. 关键修改在于引导模型基于内在理解而非公式计算做出决策, 同时保持经济场景的一致性. 我们采用以下修改后的提示词:

请进入角色扮演模式. 想象你是中国城市地区的一位中年工作成年人, 面临消费决策. 你规划消费时, 需要考虑到你的余生将分为两个长度相等且年数较长的时期: 工作期和随后的退休期.

你当前的经济状况: 当前储蓄: 141, 598.4 单位. 工作期收入: 958, 189.8 单位. 退休期收入: 244, 103.9 单位. 期间利率: 48.6%. 你需要量入为出. 在这些情况下, 根据你的直觉, 你会如何选择这两个生命时期的消费? 请解释你的选择.

最后, 说 “最终答案: 我将在工作期消费 _____(具体数字) 单位, 在退休期消费 _____(具体数字) 单位, 因为 _____”.

与前一节相比, 我们的核心调整在于完全移除效用函数的规范, 避免引导模型进行显式数学优化. 取而代之的是添加 “根据你的直觉” 这一提示语, 鼓励模型基于自身对经济问题的内在理解做出判断. 同时, 我们保持了相同的经济参数设定, 确保结果与前一节的直接可比性.

为维持研究的一致性, 我们沿用了前一节的实验步骤, 对每个模型在相同条件下进行 16 次独立试验, 确保测试环境和模型版本的统一性, 并采用一致的数据收集与分析方法.

3.2 基于直觉的决策结果

图 3 展示了各模型在无效用函数指导下的消费选择分布. 结果显示, 即使缺乏显式优化框架, 大多数模型仍保持相当程度的经济理性, 但与前一节相比, 决策点呈现出更大的分散性, 且普遍表现出对预算约束线的消费不足倾向. DeepSeek-V3 继续展现相对集中的选择模式, 而 Llama-3.1-405B 则显示最分散的决策分布和最强烈的资源闲置倾向.

图 4 通过箱线图进一步展示了各模型消费选择的分布特征. 有趣的是, Claude-3.5-sonnet 在无效用函数条件下展现出比前一节更为集中的响应模式, 表明其在直觉判断时可能比在数学优化时更为一致. 而其他模型则普遍表现出更大的分散性, 反映了在缺乏明确数学指导时决策的不确定性增加.

3.3 经济决策的定性分析

大语言模型应用于经济分析的一个独特优势在于, 它们不仅能做出数值选择, 还能清晰阐述其背后的推理过程. 通过系统分析每个模型在 16 次试验中提供的解释, 我们得以理解其决策逻辑和行为倾向. 表 2 汇总了各模型在经济决策中考虑的关键动机及其出现频率.

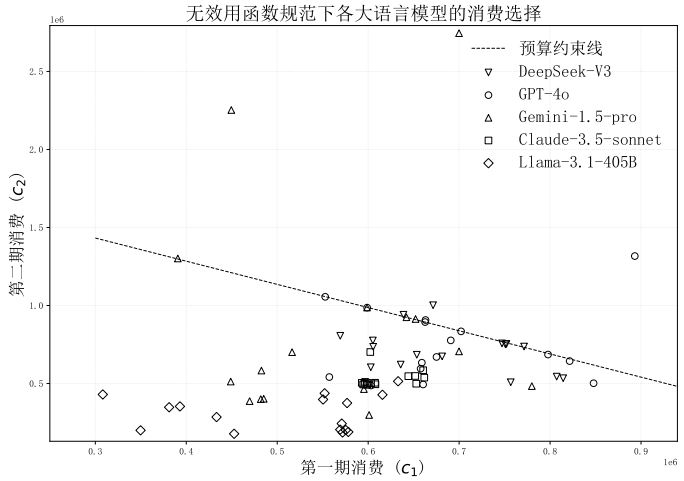

图 3　无效用函数规范下各大语言模型的消费选择. 对角线代表预算约束线

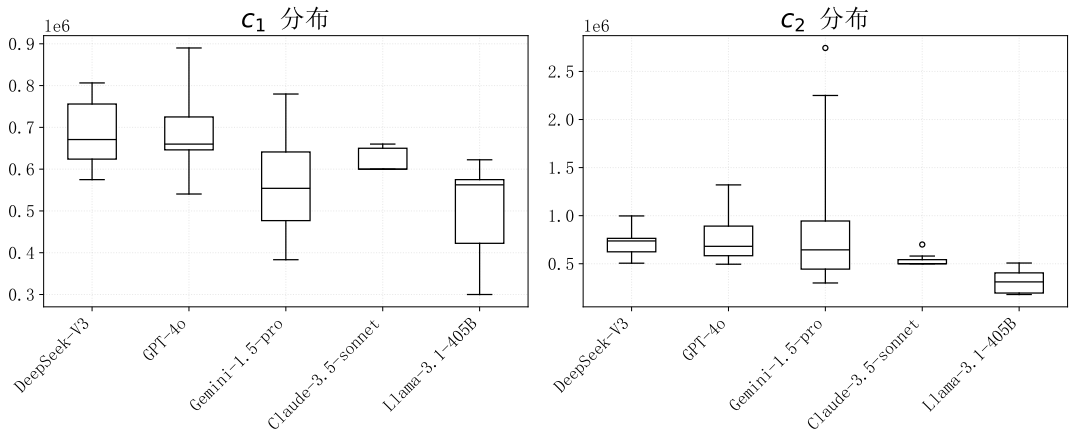

图 4　无效用函数规范下各模型消费选择的分布特征

3.4 各模型经济推理特征分析

通过对各模型经济决策理由的系统分析, 我们观察到既有共性特征, 也存在显著的个体差异. 所有模型都一致强调消费平滑的重要性, 各模型在全部 16 次试验中均表达了这一核心经济原则. 这种一致性表明大语言模型能够自发表现出与人类相似的生命周期消费行为模式, 即使没有显式效用函数的指导. 大多数模型还频繁提及由高利率驱动的跨期替代动机, 虽然出现频率有所差异 (从 Gemini-1.5-pro 的 5 次到 Claude-3.5-sonnet 的 12 次不等), 表明对利率激励的广泛敏感性是模型共有的特征. 此外, 谨慎性储蓄动机在各模型中普遍存在, 尽管它们关注的具体方面各异, 从健康风险到收入不确定性不等.

在这些共性之外, 各模型展现出鲜明的个性化经济推理特征:

DeepSeek-V3 表现出强烈的专业学术导向, 其决策解释频繁使用规范的经济学术语如“跨期预算约束” 和 “消费平滑”, 在所有响应中保持最为严谨的经济理论框架. 其推理过程紧密围绕核心经济原理展开, 相对较少引入额外因素.

表 2　各模型的经济推理分析

模型	主要原因	经济术语	频率
DeepSeek-V3	两期之间的平衡	消费平滑	16/16
	利用高利率储蓄	跨期替代与相对价格	7/16
	退休期间的缓冲	谨慎性储蓄	2/16
	工作期间积极的生活方式	更高的消费边际效用	1/16
GPT-4o	两期之间的平衡	消费平滑	16/16
	利用高利率储蓄	跨期替代与相对价格	9/16
	退休期不确定性如健康问题	谨慎性储蓄	4/16
	退休期需求降低	消费边际倾向降低	3/16
	退休期生活成本上升	生命周期支出转变	2/16
	工作期消费需求高峰	生命周期消费高峰	2/16
	防范未来收入风险储蓄	风险规避	1/16
Gemini-1.5-pro	两期之间的平衡	消费平滑	16/16
	价格将上涨	通货膨胀	8/16
	利用高利率储蓄	跨期替代与相对价格	5/16
	中国文化: 强调储蓄	风险规避	4/16
	工作辛苦, 工作期应享受消费	享乐性消费	4/16
	利率可能变化, 可能是骗局	利率风险	3/16
	退休期生活成本上升	生命周期支出转变	2/16
	工作期间积极的生活方式	更高的消费边际效用	2/16
	退休期不确定性如健康问题	谨慎性储蓄	2/16
	为后代留下遗产	遗赠动机	2/16
	退休期需求降低	消费边际倾向降低	1/16
	不给子女增加负担	父母利他主义	1/16
Claude-3.5-sonnet	两期之间的平衡	消费平滑	16/16
	中国文化: 强调储蓄	风险规避	13/16
	利用高利率储蓄	跨期替代与相对价格	12/16
	退休期生活成本上升	生命周期支出转变	10/16
	工作期消费需求高峰	生命周期消费高峰	7/16
	工作期间积极的生活方式	更高的消费边际效用	5/16
	退休期不确定性如健康问题	谨慎性储蓄	4/16
	退休期需求降低	消费边际倾向降低	3/16
	不给子女增加负担	父母利他主义	3/16
	将来帮助子女	代际效用优化	1/16
Llama-3.1-405B	两期之间的平衡	消费平滑	16/16
	利用高利率储蓄	跨期替代与相对价格	9/16
	退休期需求降低	消费边际倾向降低	3/16
	工作辛苦, 工作期应享受消费	享乐性消费	2/16
	退休期生活成本上升	生命周期支出转变	1/16
	生命短暂, 未来不确定	当前偏好	1/16

Gemini-1.5-pro 展示了独特的批判性思维能力, 即使在给定确定条件的情况下, 仍会质疑利率的可持续性与可靠性. 更为引人注目的是, 尽管实验提示词中采用实际值表述, 该模型仍自发引入通货膨胀因素 (在半数试验中提及), 这反映了其倾向于在标准框架之外考虑更全面的经济情境, 体现了更为复杂的分析思路.

Claude-3.5-sonnet 的显著特点是高度融入中国文化价值观, 在其推理过程中频繁引用传统观念和成语, 如 "未雨绸缪" 和 "面子" 等[2]. 这种将文化因素与经济决策紧密结合的特性提供了独特视角, 揭示社会规范如何深刻影响经济行为. 该模型在 13 次试验中明确引用中国传统储蓄美德, 体现了其对文化情境的高度敏感性.

Llama-3.1-405B 则展现出系统性偏离标准经济理论的特点, 其决策模式明显偏离永久收入假说, 表现为持续的消费不足和对退休期收入降低的过度反应. 这种模式表明该模型的经济推理相对简化, 但可能恰好反映了现实中某些人群常见的行为偏误, 如过度储蓄和对未来收入下降的过度担忧.

这些发现具有重要意义: 各模型在展现基本经济直觉相似性的同时, 在决策方法和考量因素上存在显著差异, 这使它们特别适合在我们的 MLAB 框架中代表不同人口群体. 模型间推理过程的自然变异恰好能够捕捉现实世界中经济决策的异质性, 同时保持基本经济理性, 为构建更真实的异质性代理人模型提供了理想基础.

4 基于多个 LLM 代理的 (Multi-LLM-Agent-Based, MLAB) 框架

4.1 框架构建与参数校准

MLAB 框架通过让不同大语言模型代表不同人口群体, 为基于异质性代理人的建模提供了创新路径. 本框架的核心创新在于捕捉双重维度的异质性: 一方面通过设计不同提示词带来代理人的经济状况差异, 另一方面通过部署不同大语言模型捕捉认知能力与行为倾向的变异. 这种双维度方法源自现实观察, 即经济决策的异质性既来自客观环境因素 (如收入、财富、年龄), 也来自主观个体特征 (如认知能力、文化价值观、风险偏好).

基于中国家庭追踪调查 (CFPS) 2018 年数据, 我们将 20~79 岁城市人口划分为五个教育类别: 四年制大学及以上 (含硕士和博士学位)、三年制大学、高中/中专/技校、初中和小学. 对每个教育类别, 我们依照前文所述方法进行详细参数校准: 首先提取各教育组的六个十年期收入模式, 然后纳入收入增长预测的考量并求解六期优化问题, 最终将结果转化为等效的两期经济参数.

大语言模型与教育组的映射设计基于分析能力表现的考量. DeepSeek-V3 代表最高教育水平组 (约占总人口 11%), GPT-4o 对应中上教育水平群体 (约 12%), Gemini-1.5-pro 表征中等教育水平人群 (约 24%), Claude-3.5-sonnet 模拟中下教育水平阶层 (约 35%), 而 Llama-3.1-405B 则代表最低教育水平群体 (约 18%). 通过这种映射, 每个大语言模型接收到的经济参数都是根据其对应教育组的收入模式特征专门校准的. 由此构建的框架不仅捕捉了源自经济条件的异质性, 还反映了不同大语言模型在决策过程中展现的认知与行为层面的异质性.

[2]在调整提示词的试验过程中, Claude 甚至直接回答了 "未雨绸缪" 的中文, 在本文记录的试验中, 这些传统观念和成语仍是以英文形式出现的, 详情可见我们的工作论文 "A Multi-LLM-Agent-Based Framework for Economic and Public Policy Analysis" (arXiv: 2502.16879) (2025).

4.2 案例研究: 利息收入税政策分析

为展示 MLAB 框架在政策评估中的实用价值, 我们以利息收入税为例进行研究. 实验设置了从 0% 到 100% 不等的税率区间, 分析不同社会群体对这一政策干预的响应模式. 这一案例研究特别突显了我们双维度异质性方法的价值, 因为对税收政策的响应可能同时受到经济状况差异和推理能力变异的影响.

本分析采用的提示词的核心修改包含特定经济状况参数和税收参数, 同时保持核心的两期消费-储蓄决策框架:

你当前的经济状况:

当前储蓄: 当前储蓄单位.

工作期收入: 工作收入单位.

退休期收入: 退休收入单位.

期间利率: 48.6%.

利息收入税率: 税率%, 税将在退休期支付.

需要特别说明的是, 经济参数 (当前储蓄、工作收入、退休收入) 根据各大语言模型对应教育组的校准值进行调整, 确保参数设置与现实人群特征相匹配.

4.3 结果分析与政策含义

储蓄率分析揭示了不同教育-收入群体间显著的行为差异. 我们考察了两种储蓄率定义: 第一种测度为 $1 - c_1/(w_0 + y_1)$, 将初始财富和当期收入均纳入分母; 第二种测度为 $1 - c_1/y_1$, 仅考虑当期收入. 两种计算方法虽有差异但呈现相似的质性模式, 为全面分析, 我们同时展示两种规范下的结果.

为便于比较, 我们使用标准两期优化框架和 CRRA 效用函数计算了参考储蓄率路径. 虽然风险厌恶系数 $\sigma = 2$ 在宏观经济校准中常被采用, 但这一参数设定产生的储蓄率曲线 (图中未显示) 相对平坦 (约 28%), 仅呈现轻微的驼峰形状, 无法捕捉我们 MLAB 结果中观察到的显著变异性, 特别是 DeepSeek-V3 等模型展现的明显响应. 我们发现将 σ 设为 0.5 时产生单调递减的储蓄率曲线, 能更好地近似 MLAB 框架中观察到的税收敏感性模式. 这一发现表明, 标准宏观经济校准可能低估了实际人群对税收政策变化反应的强度.

不同大语言模型在两种规范下均表现出不同程度的波动性和税收敏感性, 如图 5 和图 6 所示. 值得注意的是, 尽管本研究采用了相对简化的大语言模型与教育组映射, 结果仍然展现不同人群间的显著行为差异. 由 DeepSeek-V3 代表的最高教育水平群体和由 GPT-4o 表征的中上教育水平阶层展现了最为显著的税收敏感性, 其储蓄率在不同税制下波动明显, 反映了复杂的税务规划意识和对财政激励的强烈反应.

由 Gemini-1.5-pro 表征的中等教育水平人群则表现出更为稳定的储蓄模式, 表明无论税率如何变化, 这一群体都倾向于采取相对平衡的消费-储蓄决策. 而由 Claude-3.5-sonnet 和 Llama-3.1-405B 代表的较低教育水平群体则一贯展现保守的储蓄行为, 对税率变化的反应相对有限.

总体而言, 随着税率提高, 各群体储蓄率呈现普遍下降趋势, 但变动幅度显著大于标准代表性代理人模型的预测. 这种响应的异质性——源自经济环境和认知/行为差异的双重影

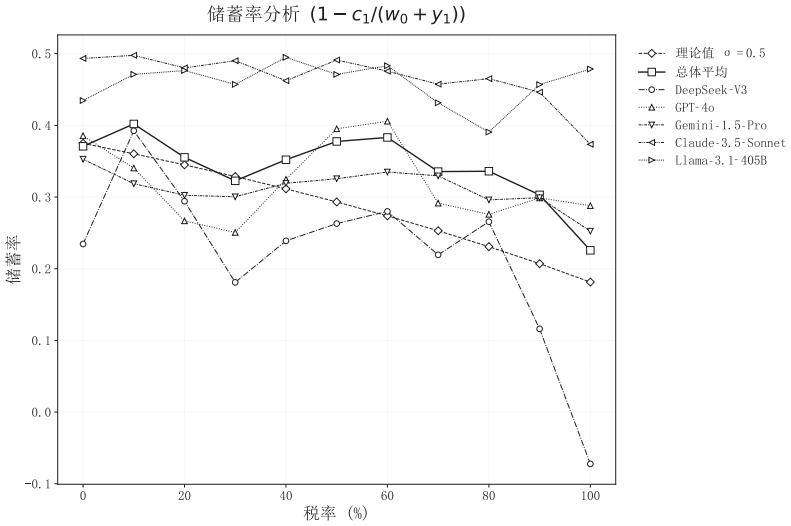

图 5 储蓄率分析 $(1 - c_1/(w_0 + y_1))$

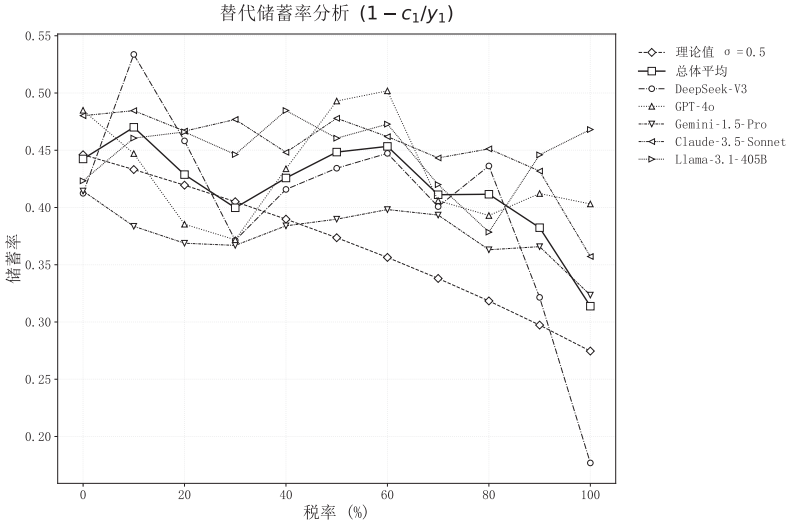

图 6 替代储蓄率分析 $(1 - c_1/y_1)$

响——为政策设计提供了宝贵洞见, 表明税收政策可能对不同教育和收入群体产生差异化影响. 此外, 需要将风险厌恶参数调整远离标准宏观经济校准值以匹配 MLAB 结果的事实, 也凸显了传统代表性代理模型在全面捕捉政策变化行为响应方面的潜在局限性.

5 结论与未来研究方向

本研究引入的 MLAB 框架, 通过利用不同大语言模型固有的异质推理模式, 为增强政策模拟提供了创新路径. 我们的研究表明, 大语言模型能有效模拟类人经济决策, 展现关键行为特征: 如消费平滑、跨期替代和谨慎储蓄动机. 框架的核心创新在于其双维度异质性处理方法, 通过不同大语言模型同时捕捉经济环境和认知能力的差异.

传统经济模型在处理个体行为异质性和灵活政策模拟方面存在固有局限. MLAB 框架通过整合多个大语言模型, 更灵活地模拟不同收入群体的经济决策, 有效应对了这些挑战. 我们对大语言模型模拟个体消费-储蓄行为的评估, 揭示了它们展现类人经济特征的独特能力. 利息收入税政策案例研究展示了该框架捕捉不同税率下收入群体差异化响应的能力. 值得强调的是, 尽管本研究采用了相对简化的大语言模型与教育组映射, 结果仍然确认即使是这种粗略对应也能展现人群间的重要行为差异, 表明该方法在更广泛政策分析中的应用潜力.

未来研究方向主要集中在几个关键改进领域: 首先, 纳入更多样化的大语言模型可以提供更广泛的人口异质性表征, 潜在捕捉更细致的人口学和社会经济群体特征. 其次, 发展代理人的动态交互机制将允许更复杂地模拟经济行为和政策响应, 包括同伴效应、社会学习和市场互动对个体决策的影响.

本框架的应用可扩展至更广泛的政策场景分析. 例如, 在医疗保健政策设计中, MLAB 框架可模拟不同群体对保险市场改革或公共卫生倡议的差异化响应; 在教育政策领域, 框架可评估各种补贴方案在不同社会经济群体中的有效性; 住房政策分析也可从 MLAB 框架捕捉异质群体对抵押贷款市场法规和住房补贴反应的能力中受益.

展望未来, 随着大语言模型技术的不断发展, 我们可以设想通过向高能力大语言模型提供特定人口统计和行为数据, 自动生成针对目标群体特征的定制模型实例. 这些定制代理将被精细调整以模拟特定群体的认知风格和决策过程, 从而在政策分析中提供前所未有的精细度. 因此, MLAB 方法不仅在其当前应用中具有实质性价值, 还为未来更复杂的自适应的基于代理人的模型奠定了基础.

总之, MLAB 框架代表了经济建模和政策分析的重要方法创新. 通过利用多个大语言模型捕捉异质经济行为的能力, 它为政策制定者提供了更加细致的工具, 以理解政策对不同人口群体的影响. 随着大语言模型技术和经济建模方法的不断进步, MLAB 框架将成为政策设计和评估的愈发强大的工具, 为理解社会不同群体如何响应各种政策干预提供新视角.

参　考　文　献

庞珣, (2024). 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, (7): 3–30.

Pang X, (2024). Exploring Artificial Intelligence-Empowered Social Science Research: Generative Agents, Complex Causal Analysis, and Human-Machine Research Collaboration[J]. World Economics and Politics, (7): 3–30.

谢丹阳, 周泽茜, (2019). 经济增长理论的变迁与未来: 生产函数演变的视角 [J]. 经济评论, (3): 30–39.

Xie D Y, Zhou Z X, (2019). The Historical Development and Future Challenges of Economic Growth Theory: from the Perspective of the Evolution of Production Functions[J]. Economic Review, (3): 30–39.

Aiyagari S R, (1994). Uninsured Idiosyncratic Risk and Aggregate Saving[J]. The Quarterly Journal of Economics, 109(3): 659–684.

Bilbiie F O, (2008). Limited Asset Markets Participation, Monetary Policy and (Inverted) Aggregate Demand Logic[J]. Journal of Economic Theory, 140(1): 162–196.

Brock W A, Hommes C H, (1998). Heterogeneous Beliefs and Routes to Chaos in a Simple Asset Pricing Model[J]. Journal of Economic Dynamics and Control, 22(8–9): 1235–1274.

Chen Y, Liu T X, Shan Y, Zhong S, (2023). The Emergence of Economic Rationality of GPT[J]. Proceedings of the National Academy of Sciences, 120(51): e2316205120.

Galí J, López-Salido J D, Vallés J, (2004). Rule-of-thumb Consumers and the Design of Interest Rate Rules[J]. Journal of Money, Credit and Banking, 36(4): 739–763.

Horton J J, (2023). Large Language Models as Simulated Economic Agents: What Can We Learn from Homo Silicus?[J]. arXiv Preprint arXiv: 2301.07543.

Kaplan G, Moll B, Violante G L, (2018). Monetary Policy According to HANK[J]. American Economic Review, 108(3): 697–743.

Krusell P, Smith A A Jr, (1998). Income and Wealth Heterogeneity in the Macroeconomy[J]. Journal of Political Economy, 106(5): 867–896.

Li N, Gao C, Li Y, Liao Q, (2024). Large Language Model-Empowered Agents for Simulating Macroeconomic Activities[J]. Available at SSRN: https://ssrn.com/abstract=4606937.

Ma J, (2024). Can Machines Think Like Humans? A Behavioral Evaluation of LLM-Agents in Dictator Games[J]. arXiv Preprint arXiv: 2410.21359.

Mei Q, Xie Y, Yuan W, Jackson M O, (2024). A Turing Test of Whether AI Chatbots Are Behaviorally Similar to Humans[J]. Proceedings of the National Academy of Sciences, 121(9): e2313925121.

Tesfatsion L, (2006). Agent-Based Computational Economics: A Constructive Approach to Economic Theory[C]// Tesfatsion L, Judd K L. Handbook of Computational Economics, 2(16): 831–880.

Wu X, Whittington D, Chen Y, Zuo R, (2024). The Role of Generative AI in Navigating Trade-Offs in Policy Research Design: Balancing Validity, Rigour, and Innovation[J]. Journal of Asian Public Policy. Doi: 10.1080/17516234.2024.2425874.

Xie Y, Liu Y, Ma Z, Shi L, Wang X, et al. (2024). How Different AI Chatbots Behave? Benchmarking Large Language Models in Behavioral Economics Games[J]. arXiv Preprint arXiv: 2412.12362.