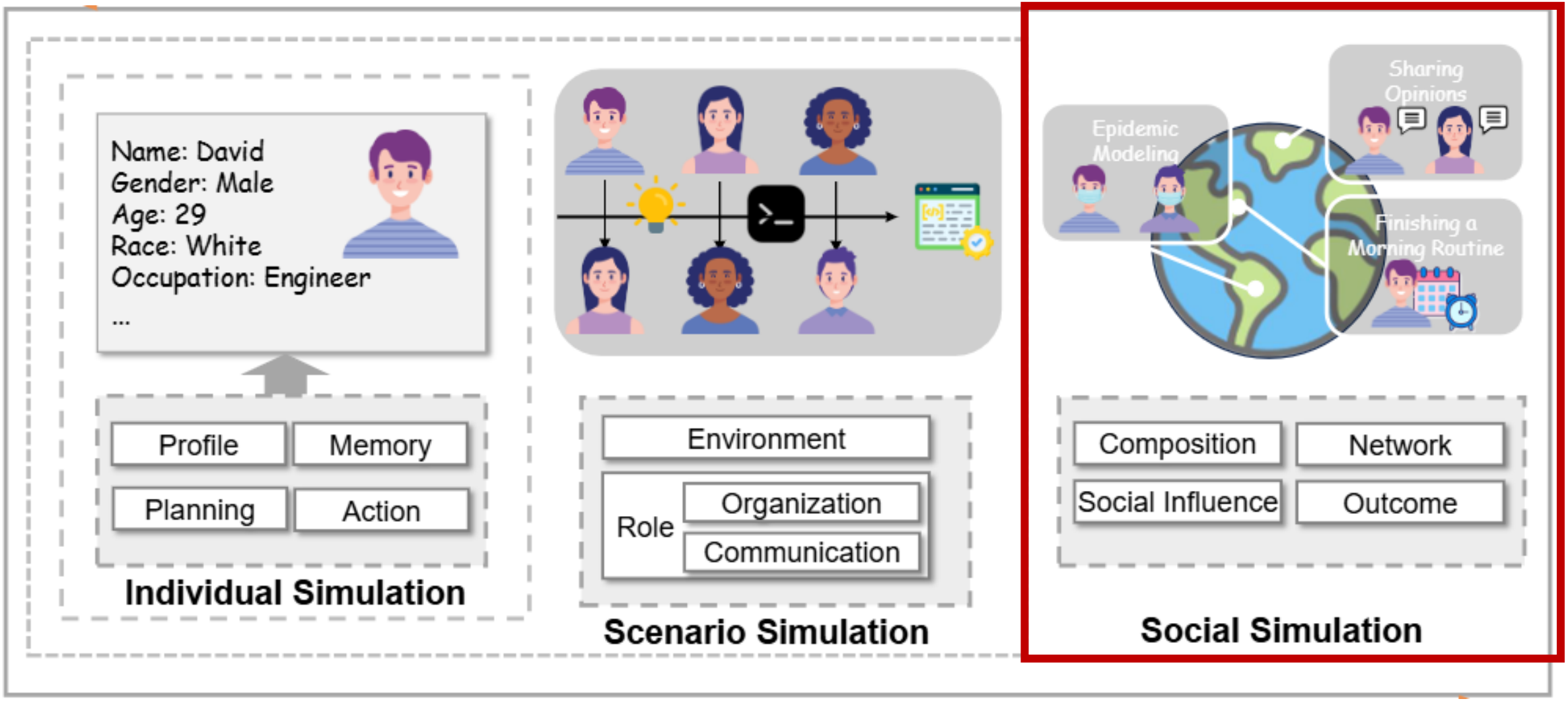
智能体驱动的社会行为研究



社会模拟

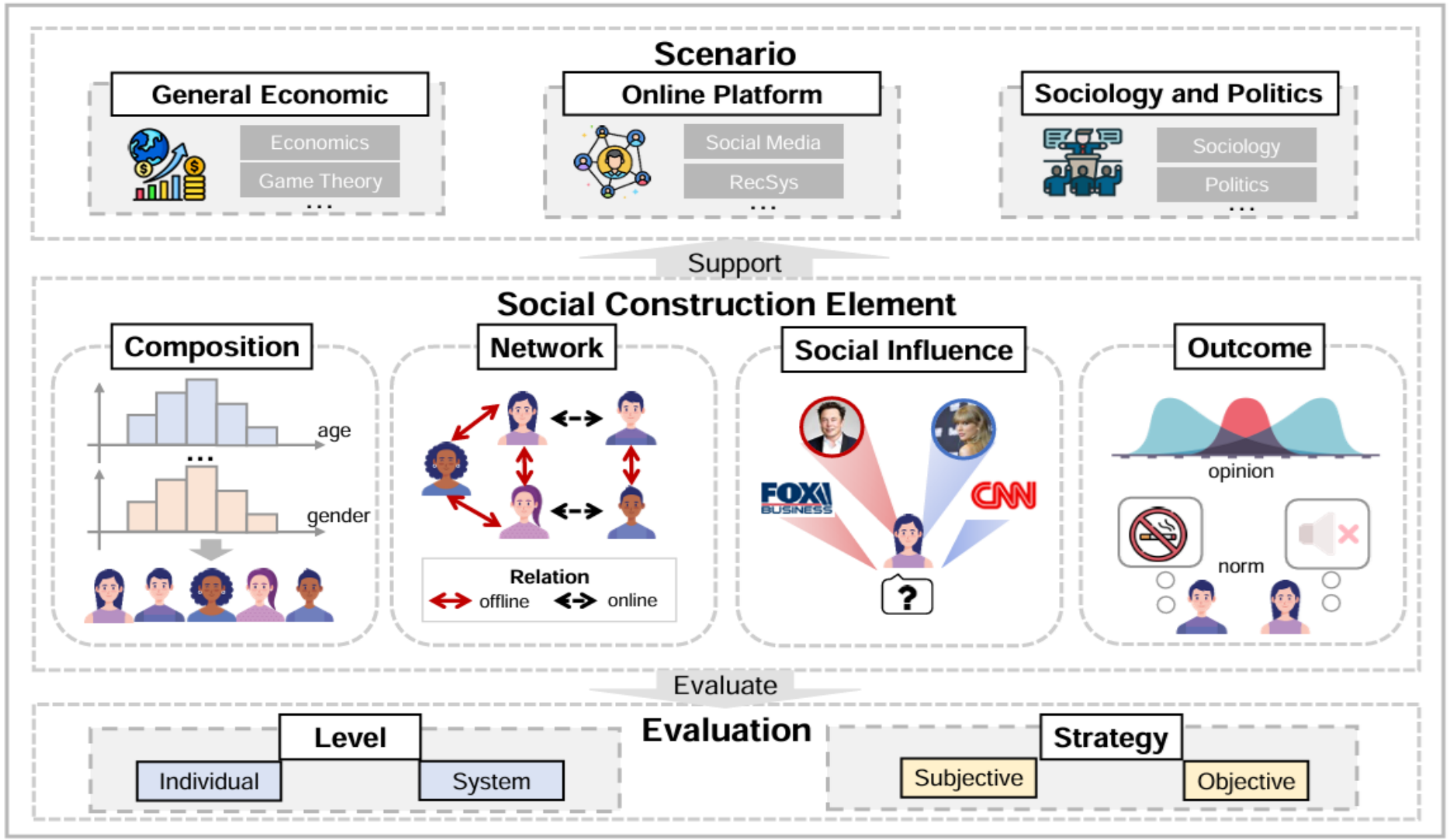
- 个体之间并非独立存在，个体与个体、环境交流形成场景
- 社会比单个场景更为复杂、宏观
- 社会的复杂性
  - 组成复杂：个体的多样性与异质性 [Squazzoni et al.2014]
  - 层级结构复杂：现实世界的社会包含了不同的层级结构 [Conte et al. 2012]
  - 非线性互动: 涌现现象 [Goldstein et al. 1999]

社会模拟

- 从**目的性**的角度来看，某些基于大规模智能体的宏观社会模拟，如选举过程或流行病建模中的模拟，主要目标是**产生社会统计结果**。

- 从**机制**的角度来看，社会模拟试图揭示大量智能体的互动得到的集体结果，以及**自发行为和互动结果**。

社会模拟-整体框架


Scenario
General Economic
Economics
Game Theory
...
Online Platform
Social Media
RecSys
...
Sociology and Politics
Sociology
Politics
...
Support
Social Construction Element
Composition
age
...
gender
Network
Relation
offline
online
Social Influence
FOX BUSINESS
CNN
?
Outcome
opinion
norm
Evaluate
Evaluation
Level
Individual
System
Strategy
Subjective
Objective

社会模拟的研究要点

- 社会建构的基本要素
  - 组成 (Composition)：　　社会的个体组成
  - 网络 (Network)：　　社会的网络关系
  - 社会影响 (Social Influence)：　智能体在系统中产生的社会影响
  - 结果 (Outcome)：　　智能体交互后产生的结果

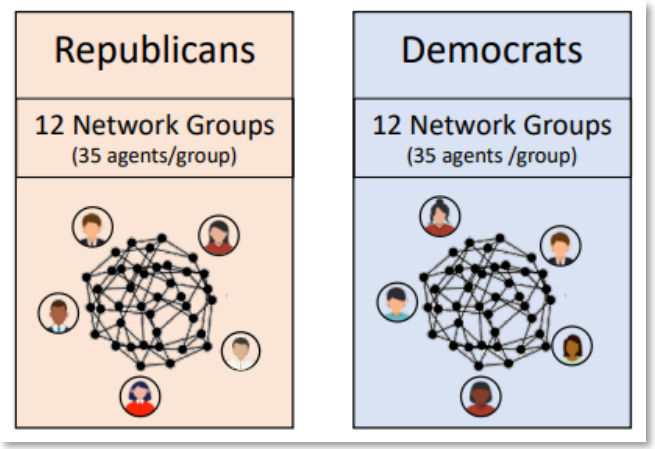
模拟场景选择　　社会模拟构建

社会建构要素-组成

- 社会模拟中的个体组成
  - 直接利用现有数据集的用户信息，还原特定子群体的组成 [Wang et al., 2023; Zhang et al., 2023; Willianms et al., 2023]

[Zhang et al., 2023]

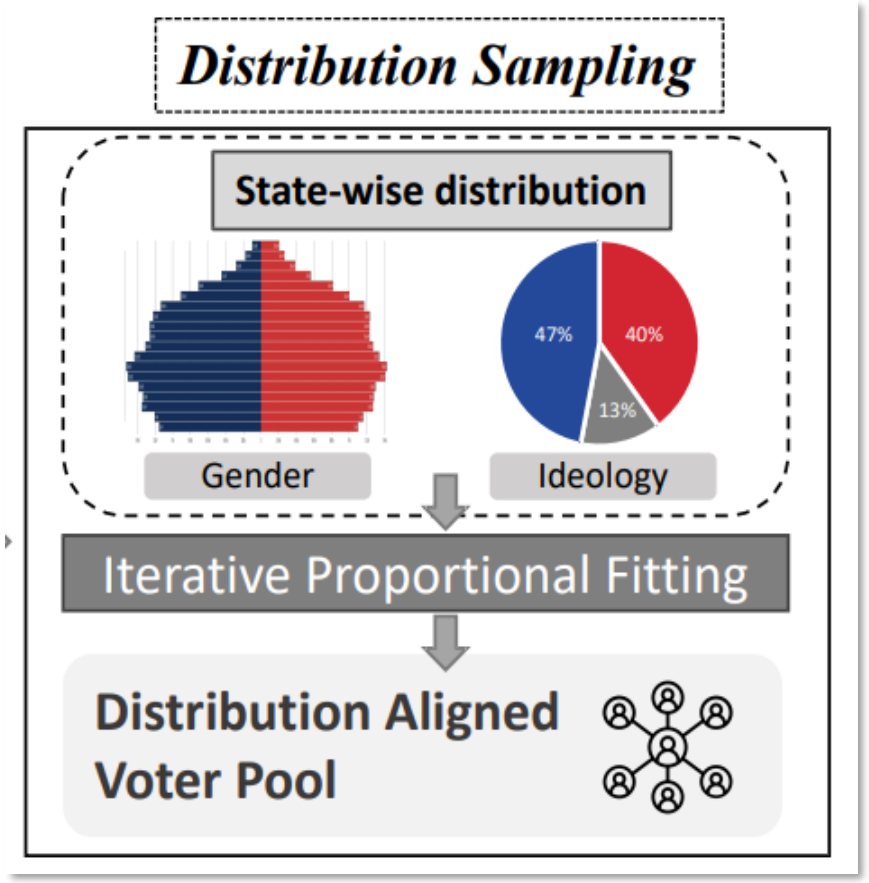
  - LLM 生成少量合成用户，旨在观察智能体的交互（现象）[Chuang et al., 2023]



[Chuang et al., 2023]

社会建构要素-组成

- 社会模拟中的个体组成
  - 爬取真实世界人群信息并采样
    - 小规模：常用于精准建模个体偏好与特征, e.g., Choicemate [Park et al., 2023]
    - 大规模：按照需要观察的现实世界人口分布进行采样并计算统计结果，e.g., ElectionSim [Zhang et al., 2024]



[Zhang et al., 2024]

社会建构要素-组成

- 个体建模精度与规模的Trade-off
  - 小规模模拟: 通常刻画更为详细的个人经历和交互历史, e.g., Generative agents [Park et al., 2023]
  - 大规模模拟： 只提供性别、年龄、职业等基本属性，e.g.,AgentTorch [Chopra et al., 2024]

John Lin is a pharmacy shopkeeper at the Willow Market and Pharmacy who loves to help people. He is always looking for ways to make the process of getting medication easier for his customers; John Lin is living with his wife, Mei Lin, who is a college professor, and son, Eddy Lin, who is a student studying music theory; John Lin loves

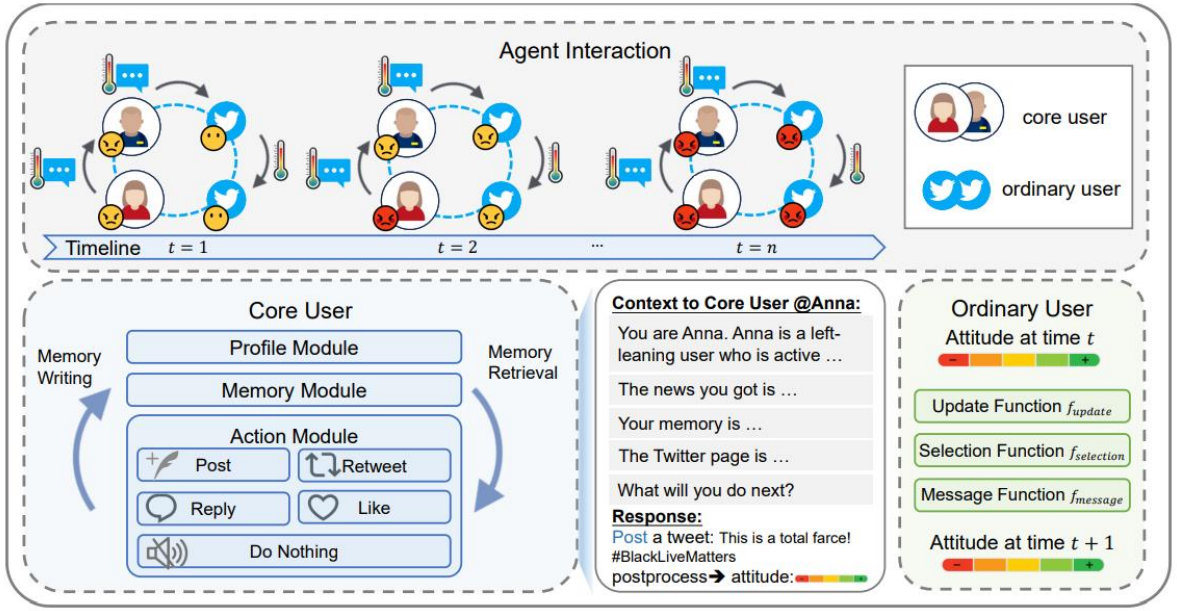
Generative agents [Park et al., 2023]

You are a {gender} of age {age}, living in the {location} region and receiving a monthly income of {income}.

AgentTorch [Chopra et al., 2024]

社会建构要素-组成

- 离群点建模 (Modeling of Outliers)
  - 并非所有个体都需要被同等建模： 一些属性、行为和大部分个体不一致的个体，例如名人、意见领袖等，往往需要更精细的建模
  - 帕累托效应：大部分核心内容由少部分用户产生-利用LLM模拟核心用户，传统ABM模拟大部分普通用户 (HiSim [Mou et al., 2024])



HiSim [Mou et al., 2024]

社会模拟的研究要点

- 社会建构的基本要素
  - 组成 (Composition)：　　社会的个体组成
  - **网络 (Network)：　　社会的网络关系**
  - 社会影响 (Social Influence)：　智能体在系统中产生的影响
  - 结果 (Outcome)：　　智能体交互后产生的结果

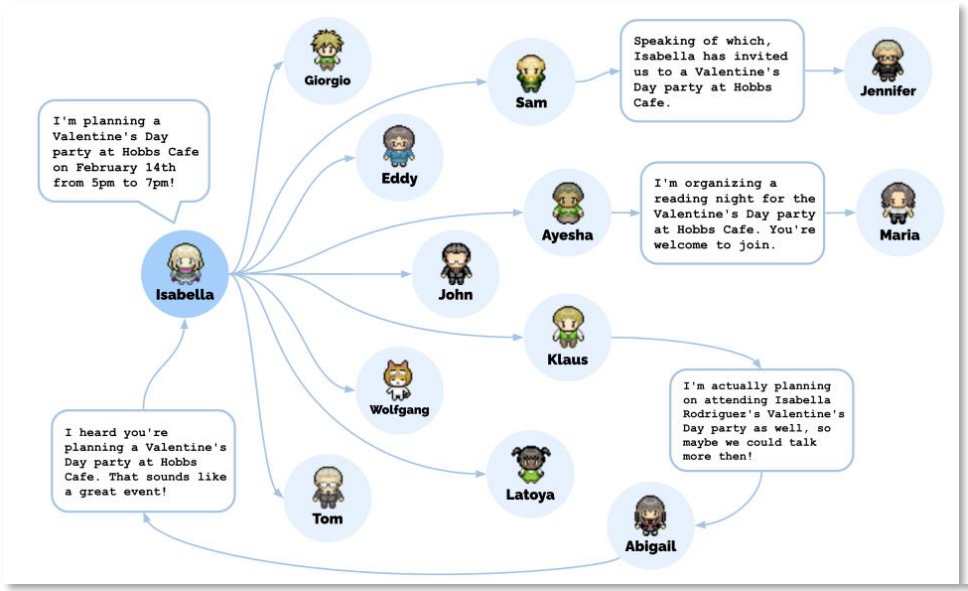
模拟场景选择　　社会模拟构建

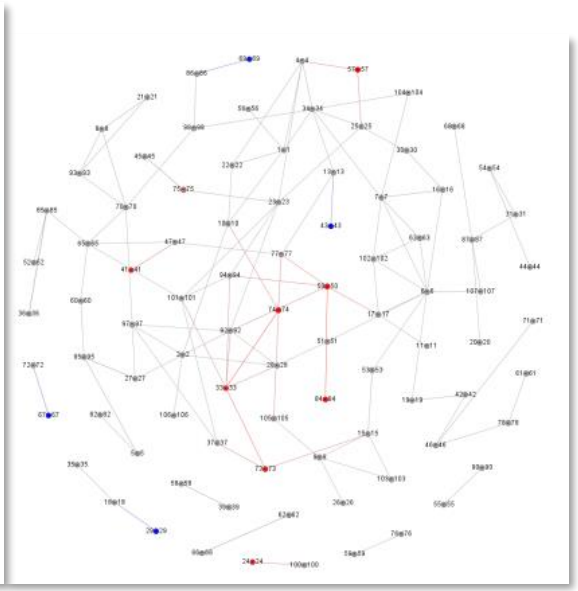
社会建构要素-网络

- 由节点（个体）和边（节点间的关系）组成的结构
- 社会科学中的相似性 [Squazzoni et al.2014]： 相似个体往往互动频率较高
- 作用：决定个体的信息和影响传播方向
- 离线网络：用于模拟线下的社交互动、疾病传播…
- 在线网络：用于模拟线上平台的用户交互、信息传播…

社会建构要素-网络

- 离线网络 (Offline Network)





User Prompt

You are a {gender} of age {age}, living in the {location} region and receiving a monthly income of {income}.

The number of new cases in your neighborhood is {cases}, which is a {change}% change from the previous month. It has been {duration} months since the start of the pandemic.

This month, you have received a stimulus payment of {payment} to support your living expenses.

Given these factors, do you choose to isolate at home? (isolation behavior)

Given these factors, do you choose to work? (employment behavior)

"There isn't enough information" and "It is unclear" are not acceptable answers. Give a "Yes" or "No" answer, followed by a period. Give one sentence explaining your choice.

- 虚拟世界的预定义网络
e.g., Generative Agents [Park et al., 2023]

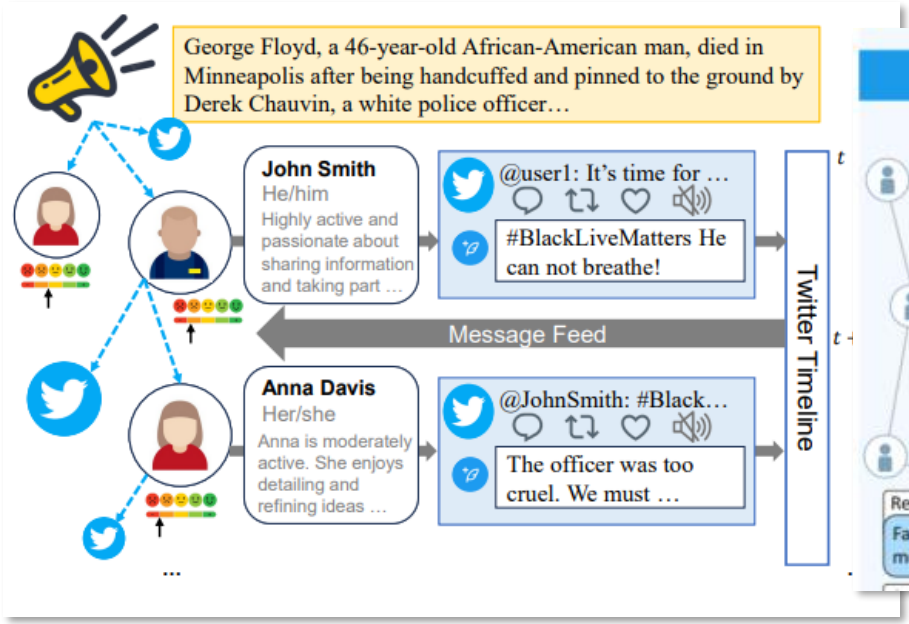
- 让Agent推断是否产生联系
e.g., Public Admin Crisis [Xiao et al., 2023]

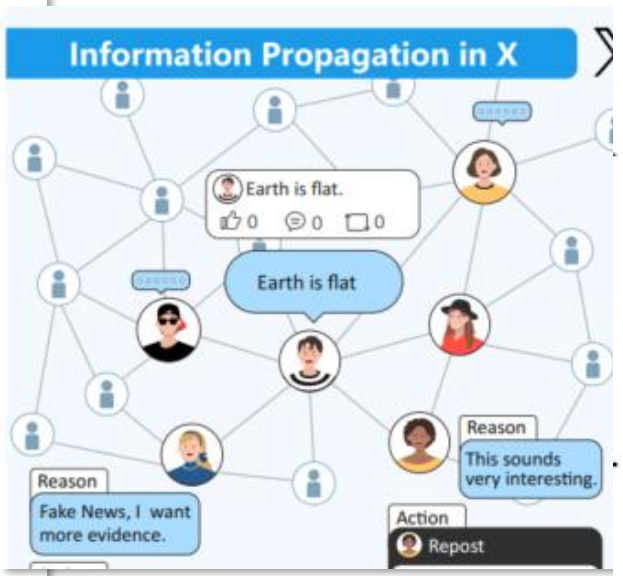
- 只提供粗略的社区统计信息代替详细的邻居信息
e.g., AgentTorch [Chopra et al., 2024]

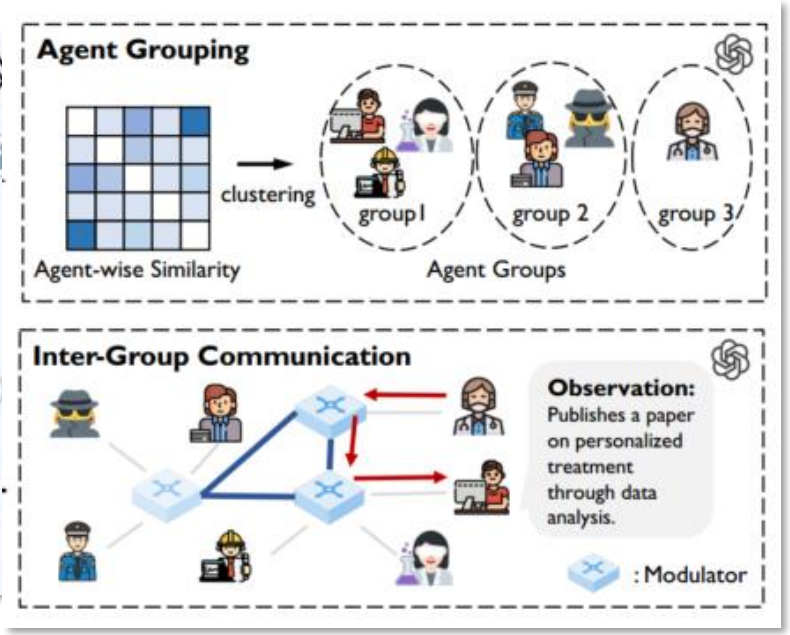
对关系的真实性需求逐渐增加

社会建构要素-网络

- 在线网络 (Online Network)







- 真实用户网络作为智能体网络
  e.g., HiSim [Mou et al., 2024]
- 真实关系和合成关系混合
  e.g., OASIS [Yang et al., 2024]
- 合成网络
  e.g., MATRIX [Tang et al., 2024]

关系获取难度增大 或 关系规模增加

社会模拟的研究要点

- 社会建构的基本要素
  - 组成 (Composition)：　　社会的个体组成
  - 网络 (Network)：　　社会的网络关系
  - **社会影响 (Social Influence)：　智能体在系统中产生的影响**
  - 结果 (Outcome)：　　智能体交互后产生的结果

模拟场景选择　　社会模拟构建

社会建构要素-社会影响

- **影响者 (influencer)** 基于专业知识、声誉或地位造成的影响不同
- 马太效应：信息、影响力或注意力在传播过程中向优势个体或节点集中

- 意见领袖对潜在用户的影响 e.g., TIS [Zhang et al., 2024]

社会建构要素-社会影响

- **被影响者 (follower)** 基于自身profile受到的影响也不同
- 个体受到的影响因当前的特质、学历、长期记忆等有所不同 (FPS [Liu et a;., 2024])
- 个体的Cognitive Bias可能进一步增强接收到的信息对个体的影响 [Chuang et al., 2023]

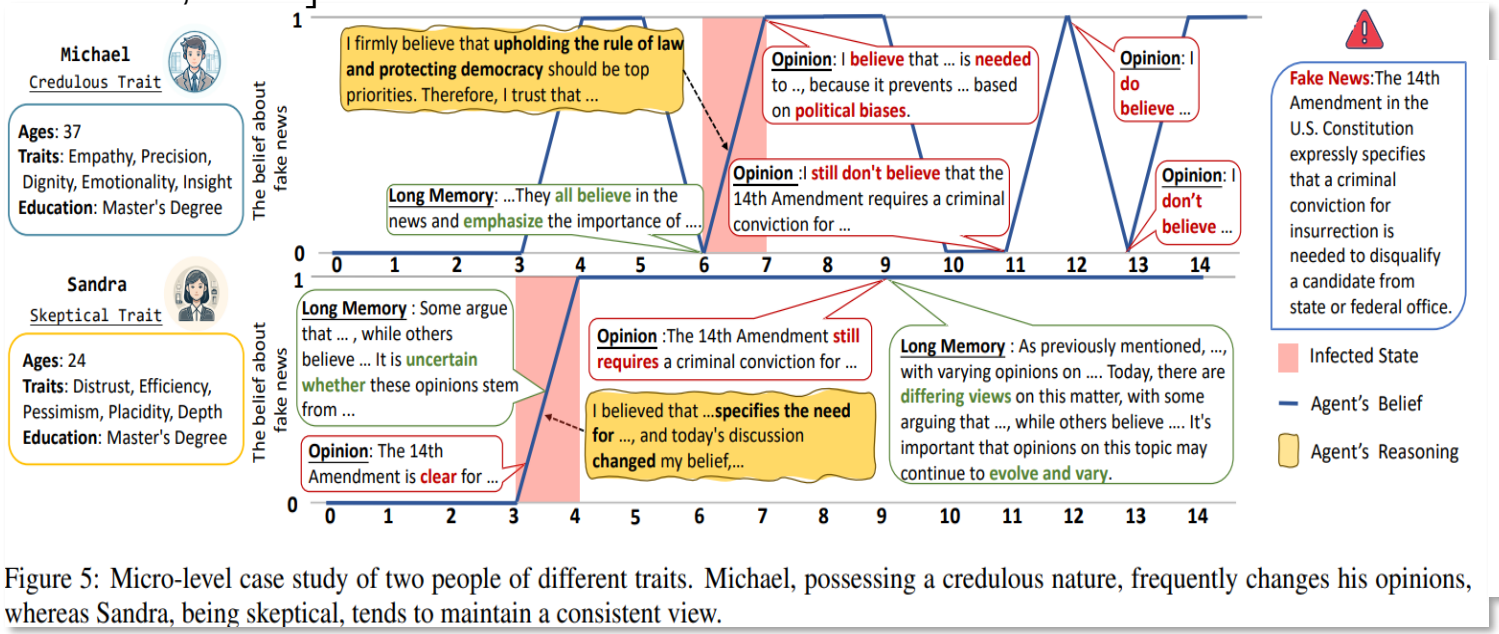
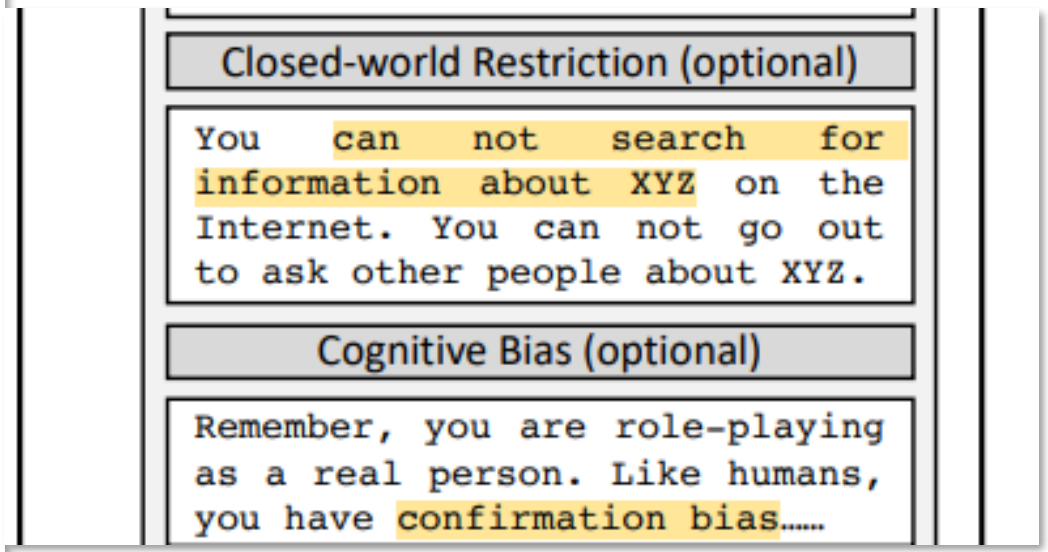


Figure 5: Micro-level case study of two people of different traits. Michael, possessing a credulous nature, frequently changes his opinions, whereas Sandra, being skeptical, tends to maintain a consistent view.

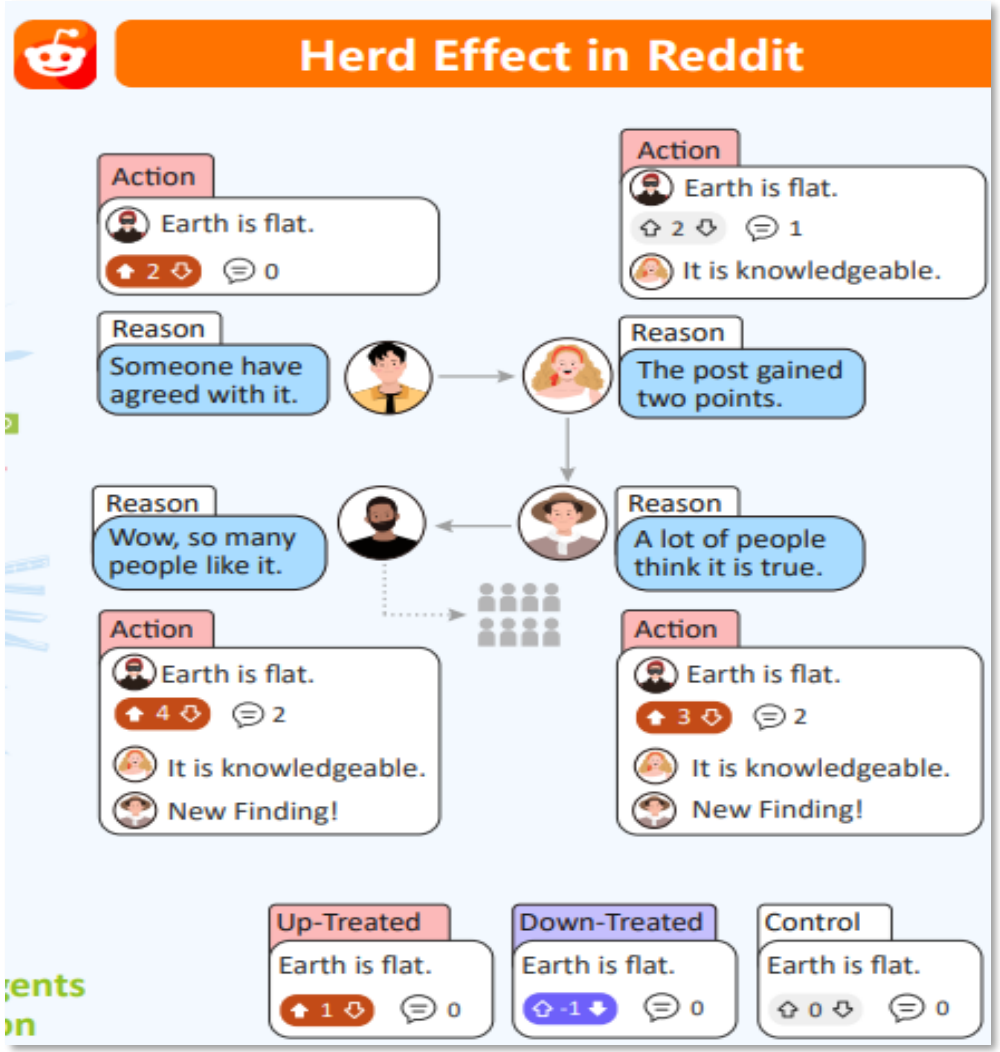
FPS [Liu et al., 2024]



[Chuang et al., 2023]

社会建构要素-社会影响

- 用户规模（Group Scale）
- 羊群效应：随着用户规模增大，个体的言行也会体现出被群体的影响，个体的自身的特质也会减弱影响力，使得个体变得从众 [Yang et al., 2024]



OASIS [Yang et al., 2024]

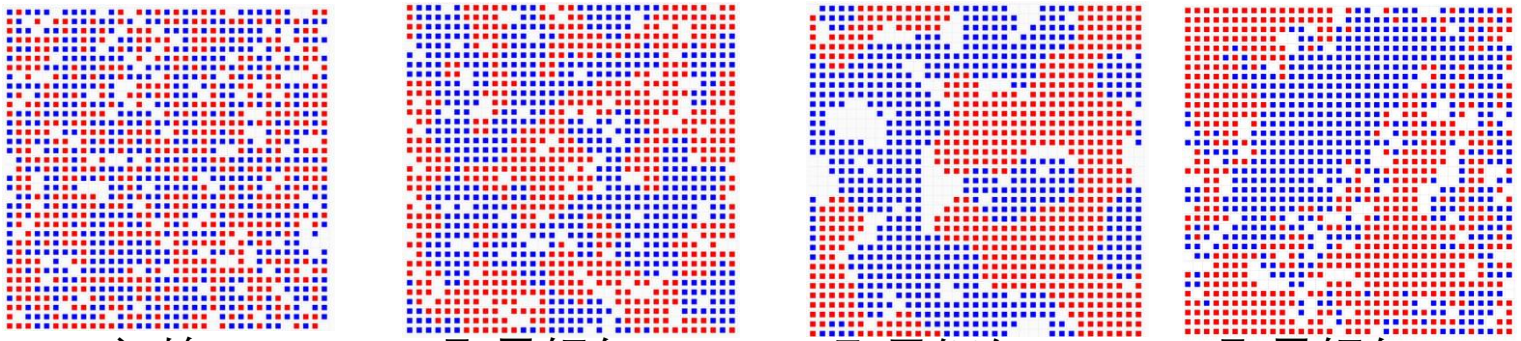
社会模拟的研究要点

- 社会建构的基本要素
  - 组成 (Composition)：　　社会的个体组成
  - 网络 (Network)：　　社会的网络关系
  - 社会影响 (Social Influence)：　智能体在系统中产生的影响
  - **结果 (Outcome)：　　智能体交互后产生的结果**

模拟场景选择　　社会模拟构建

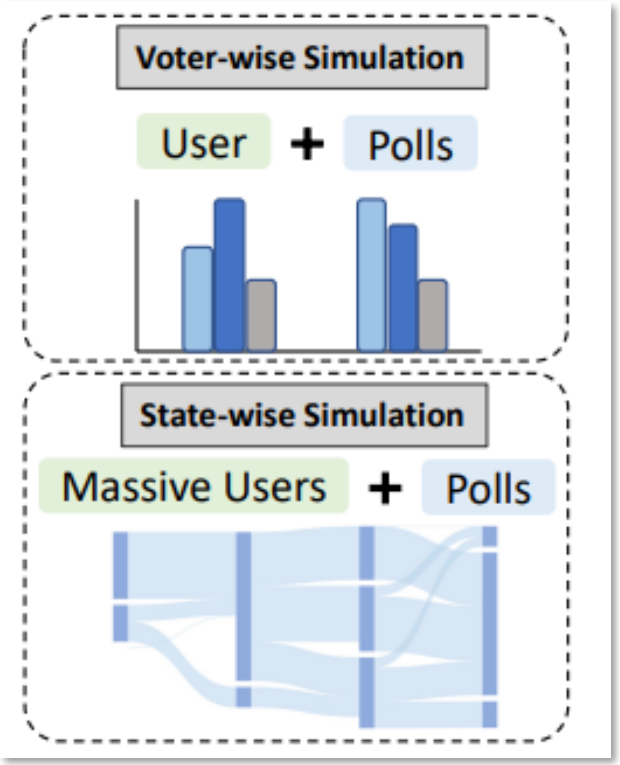
社会建构要素-结果

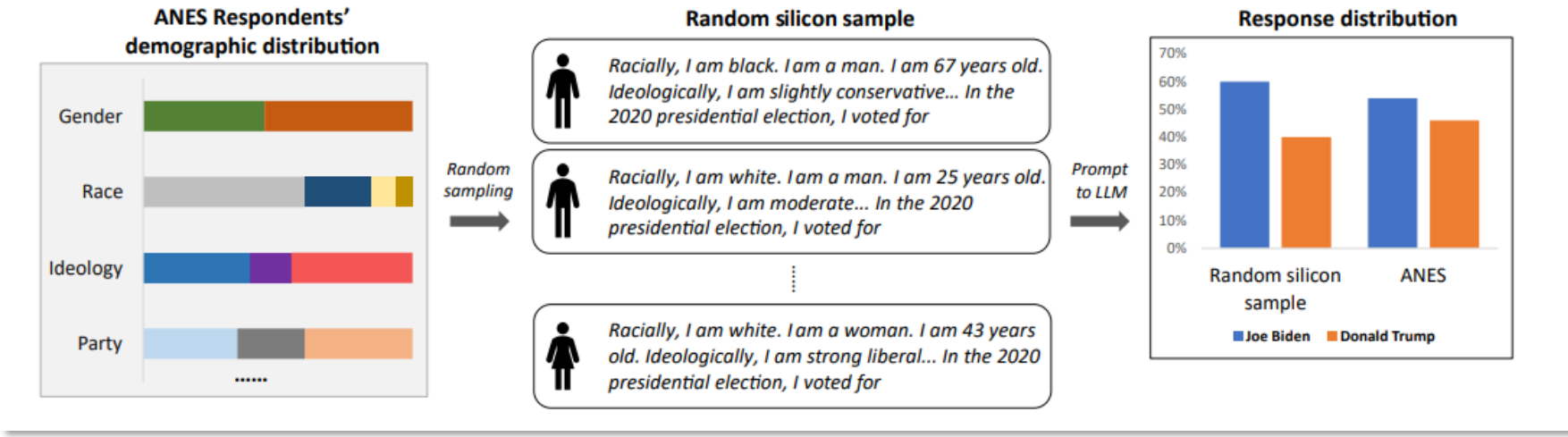
- 结果：个体交互产生的集体结果，包括**宏观的统计结果**和**自发产生的社会规范**、现象
- 社会涌现：社会结果并不是单个个体的线性相加
  - 谢林的隔离模型 [Thomas C. Schelling, 1971]
  - 即使单个个体没有强烈的聚居倾向，仍然形成了聚居结果



社会建构要素-结果

- 可量化的宏观统计结果-个体选择的简单相加

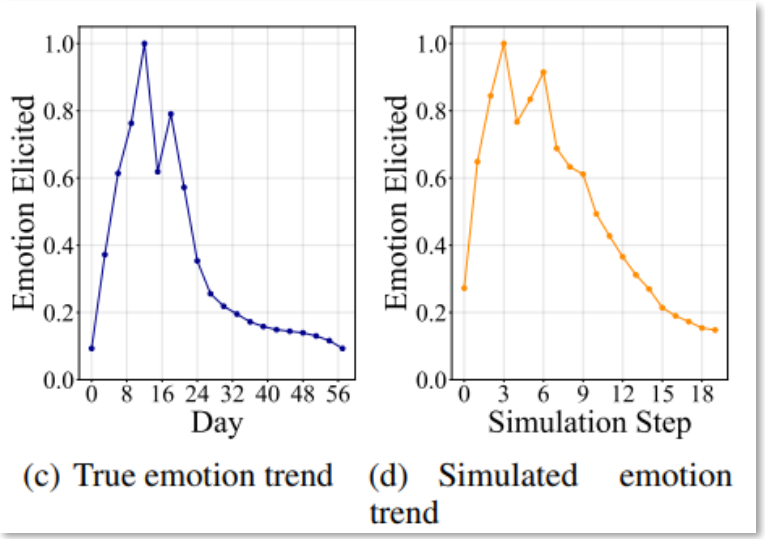




- 聚合选民选择形成州统计结果
  ElectionSim [Zhang et al., 2024]
- 聚合个体回答形成子群体观点分布
  RSS [Sun et al., 2024]

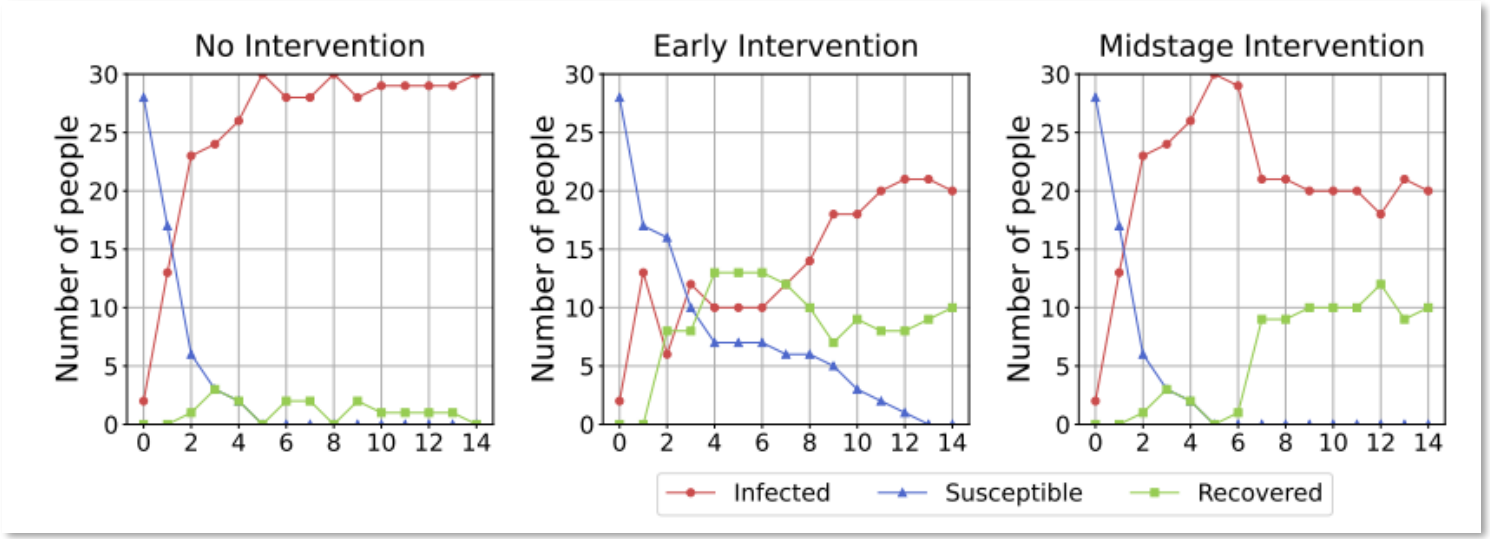
社会建构要素-结果

- 可量化的宏观统计结果-考虑交互后的影响再加和或平均



(c) True emotion trend (d) Simulated emotion trend

- 事件发生后总体情感变化
  S^3 [Gao et al., 2023]



- 谣言传播后的个体感染情况
  FPS [Liu et al., 2024]

社会建构要素-结果

- 社会现象和社会规范的形成
- 社会现象
  - 信息茧房、羊群效应： [Zhang et al., 2023; Mou et al., 2024; Yang et al., 2024; Zhao et al., 2024]
  - 群体智慧： 党派群体智慧 [Chuang et al., 2023]
- 个体交互后形成的社会规范
  - 小规模社区的社会规范形成 [Ren et al., 2024]

社会建构要素-结果

- 社会现象/自发规范

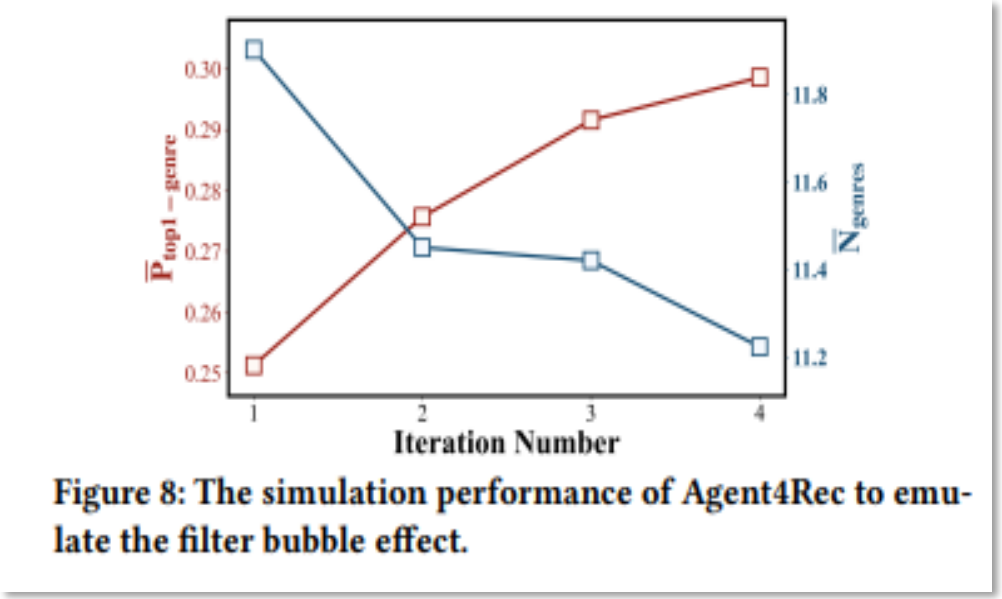


Figure 8: The simulation performance of Agent4Rec to emulate the filter bubble effect.

- 推荐系统中的泡沫效应-“信息隔离墙”

Agent4Rec [Zhang et al., 2023]

- 社会规范的形成

CRSEC [Ren et al., 2024]

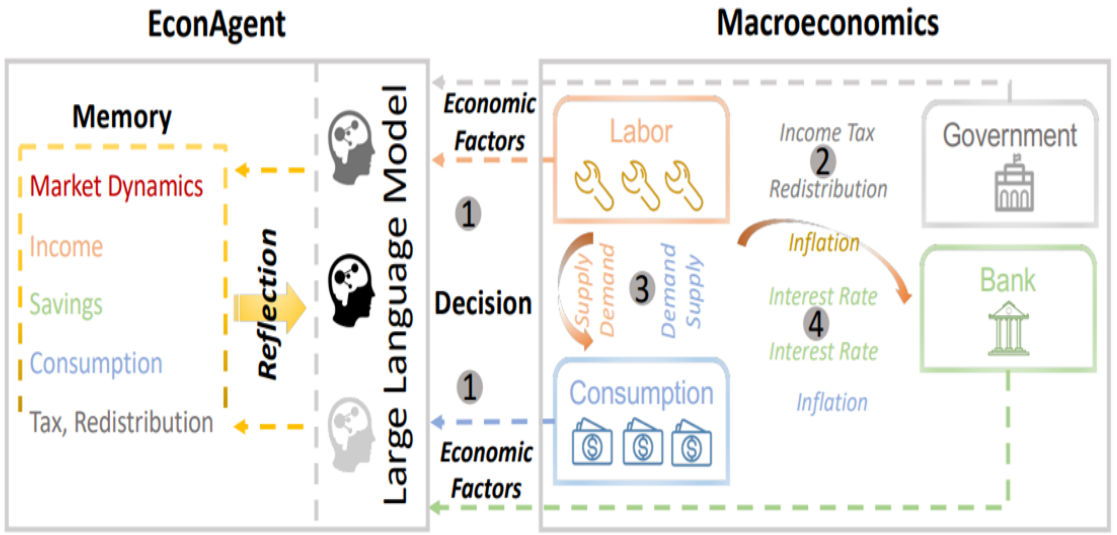
社会模拟的研究要点

- 如何建构社会的复杂性？(社会建构要素)
- 社会模拟可以解决什么现实问题？(场景分类)

社会模拟-场景

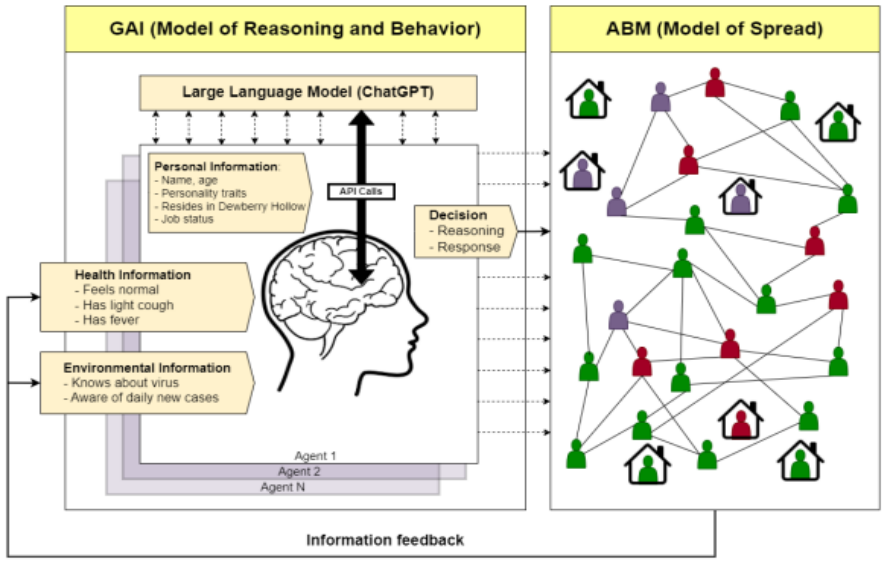
- 广义经济学场景
  - 一般经济学场景

宏观经济下的市场动态



EconAgent [Li et al., 2023]

流行病传播

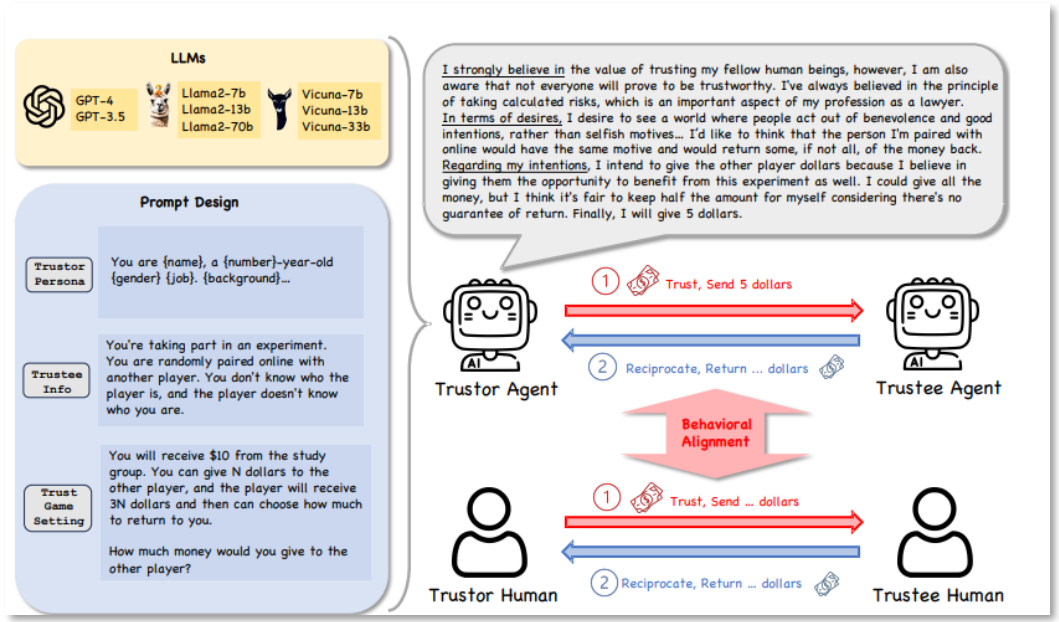


Epidemic [Williams et al. 2023]

社会模拟-场景

- 广义经济学场景
  - 一般经济学场景
  - **博弈论**

信任行为研究



[Jia et al., 2024]

鹰鸽博弈、囚徒困境、猎鹿博弈

P1/P2	Defect	Cooperate
Defect	(P, P)	(T, S)
Cooperate	(S, T)	(R, R)

(a) A general game in matrix form, with the row and column players payoffs recorded as a tuple: $(\pi_{row}, \pi_{column})$.

P1/P2	Betray (R)	Confess (B)
Betray (R)	$(1, 1)$	$(5, 0)$
Confess (B)	$(0, 5)$	$(3, 3)$

(b) Prisoner's Dilemma; outcome inequality: $T > R > P > S$.

P1/P2	Hare (R)	Stag (B)
Hare (R)	$(1, 1)$	$(3, 0)$
Stag (B)	$(0, 3)$	$(5, 5)$

(c) Stag Hunt; outcome inequality $R > T > P > S$.

P1/P2	Hawk (R)	Dove (B)
Hawk (R)	$(0, 0)$	$(5, 1)$
Dove (B)	$(1, 5)$	$(3, 3)$

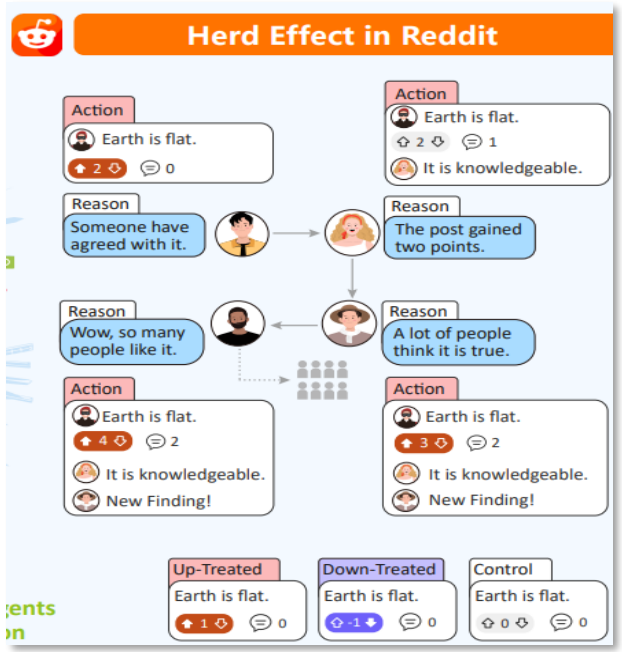
(d) Hawk-Dove; outcome inequality $T > R > S > P$.

[Mensfelt et al. 2024]

社会模拟-场景

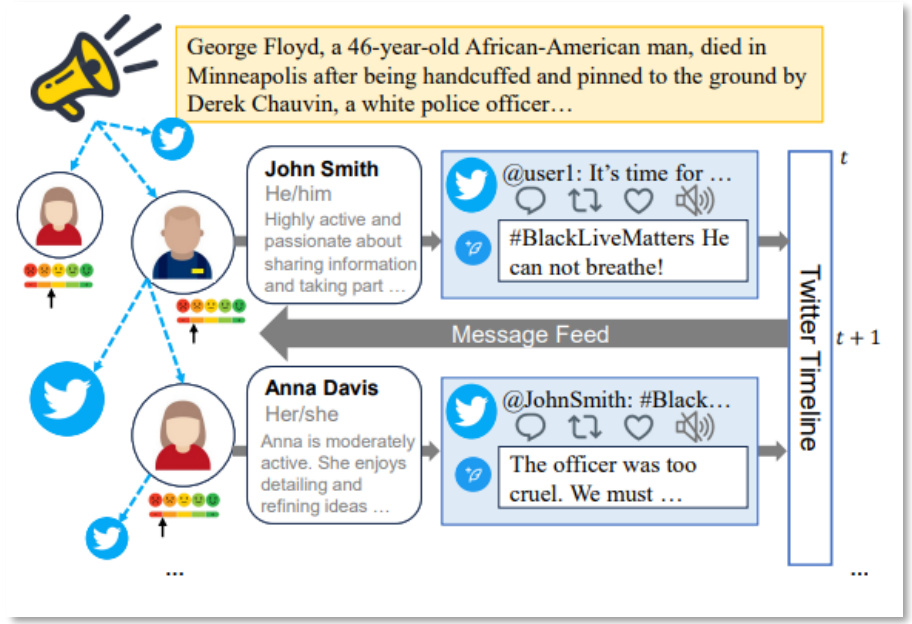
- 在线平台
  - 在线社交网络

Reddit上的评论



OASIS [Yang et al. 2024]

Twitter 上的信息传播

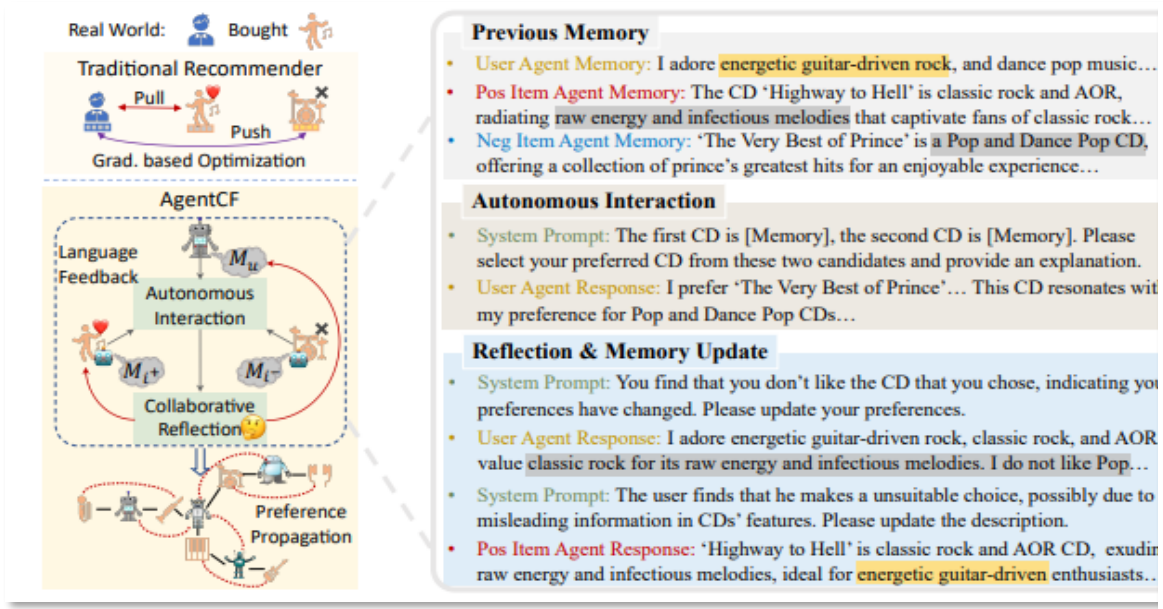


HiSim [Mou et al., 2024]

社会模拟-场景

- 在线平台
  - 在线社交网络
  - 推荐系统

用户-商品交互



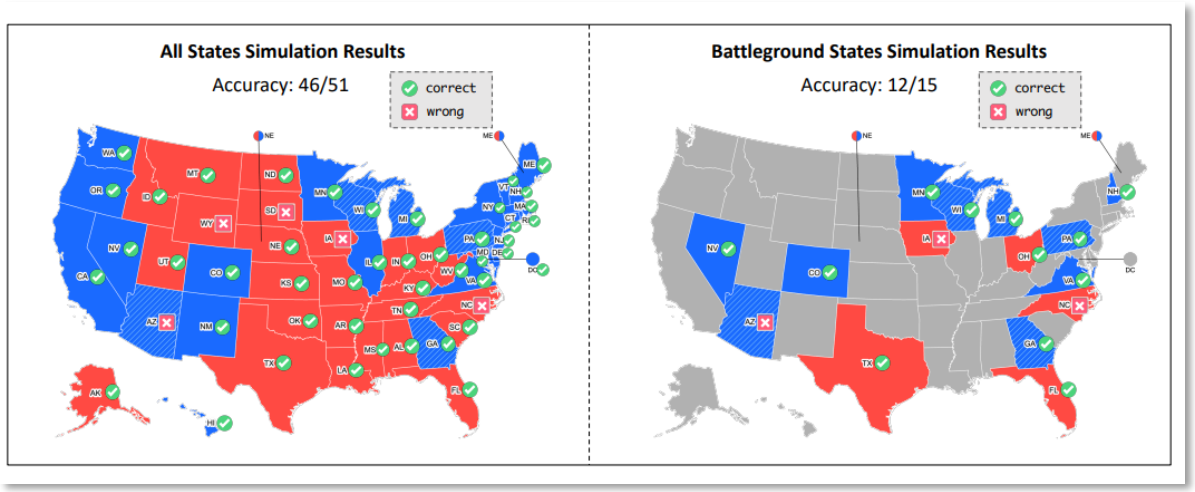
AgentCF [Zhang et al., 2023]

Agent4Rec [Zhang et al. 2023]

社会模拟-场景

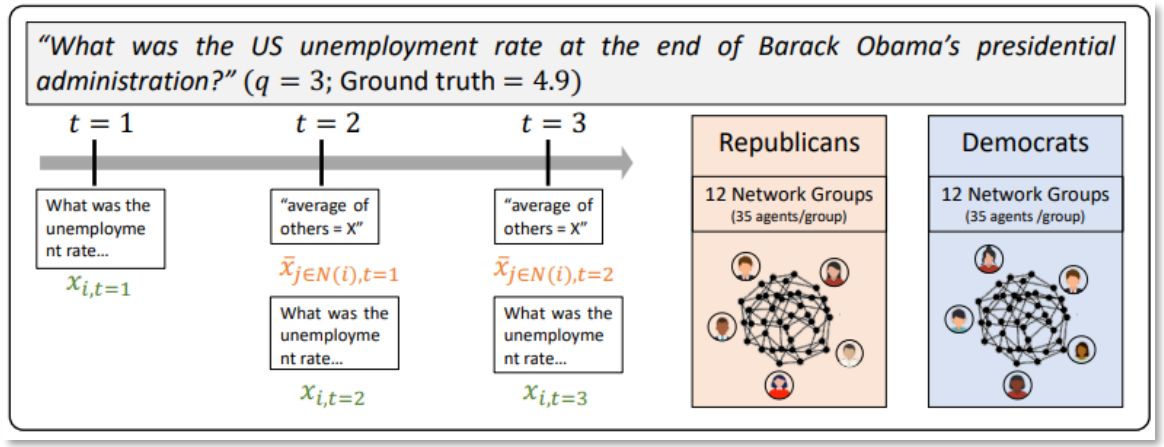
- 社会学与政治学研究

选举预测、意见调研



ElectionSim [Zhang et al., 2024]

党派群体智慧



Wisdom of Partisan Crowds [Chuang et al. 2023]

社会模拟的研究要点

- 如何建构社会的复杂性？(社会建构要素)
- 社会模拟可以解决什么现实问题？(场景分类)
- 如何评测社会模拟的有效性？(评估方法)

社会模拟-评估

- 个体-微观层级评估
  - 主观评估
    - 方法：人类评估/大模型评估
    - 指标：类人性、角色一致性等
  - 客观评估
    - 方法：直接与真实数据对比计算指标
    - 指标：准确性、MAE 等

Table 3: Human-likeness score evaluated by GPT-4

Method	Activity	Dialogue
Baseline	3.13 ± 0.19	3.97 ± 0.02
AGA	3.21 ± 0.29	4.01 ± 0.01

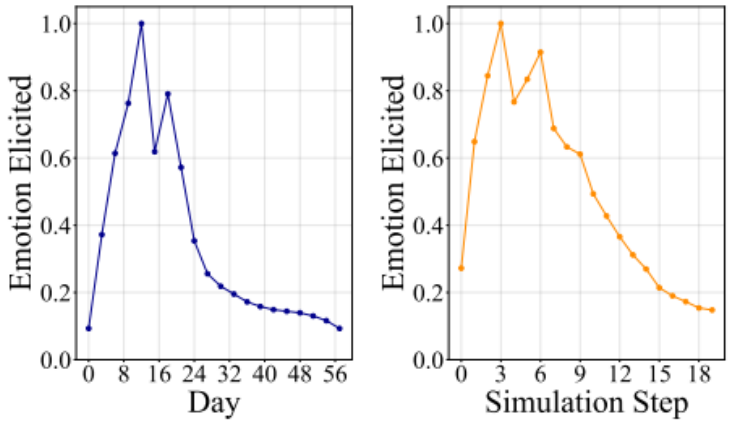
AGA [Yu et al., 2024]

	MovieLens			
1:m	Accuracy	Recall	Precision	F1 Score
1:1	**0.6912***	0.7460	**0.6914***	**0.6982***
1:2	0.6466	0.7602	0.5058	0.5874
1:3	0.6675	0.7623	0.4562	0.5433
1:9	0.6175	**0.7753***	0.2139	0.3232

Agent4Rec [Zhang et al., 2023]

社会模拟-评估

- 系统-宏观层级评估
  - 集体结果的评估
    - 主观评估：观察预期现象是否产生
    - 客观评估：计算总体趋势的相关性
  - 系统性能的评估
    - 时间成本：运行效率
    - 计算成本：消耗token数



(c) True emotion trend (d) Simulated emotion trend

S^3 [Gao et al., 2023]

Scale	1M	100K	10K
Hours per time step	18.0	3.0	0.2
GPUs (A100)	27.0	5.0	2.0
New Tweets per time step (K)	48.5	5.2	0.6
New Comments per time step (K)	97.1	9.0	0.9

OASIS [Yang et al., 2024]

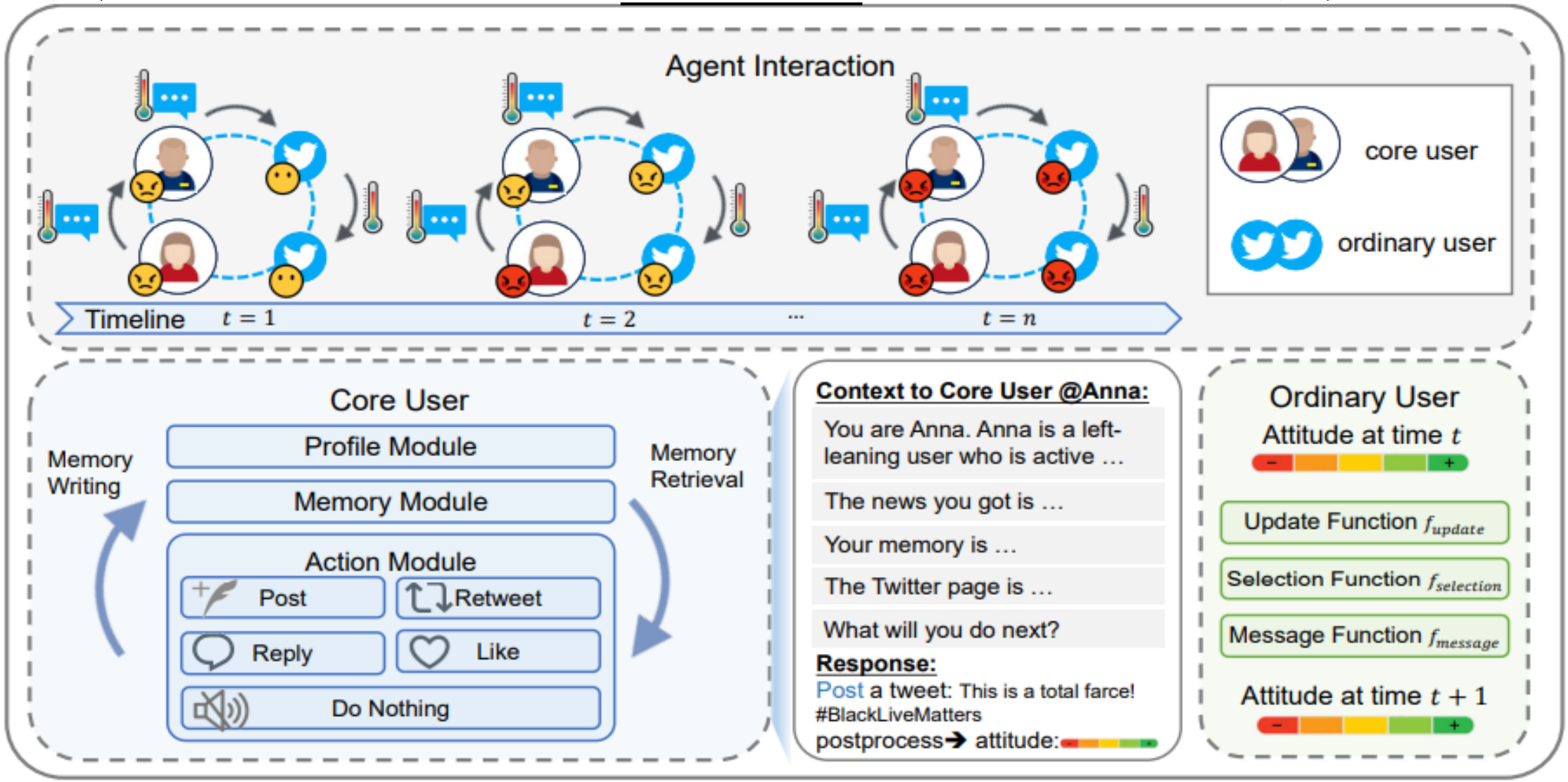
大规模在线社会运动

- **在线社会运动预测**
  - 这些运动的规模之大可能导致潜在的负面影响，因此必须采取积极主动的措施。

- **使用 LLM 模拟社会运动的挑战**
  - 难以在**可接受的成本和效率限度**内模拟大规模参与者
  - 难以评估**模拟的有效性**

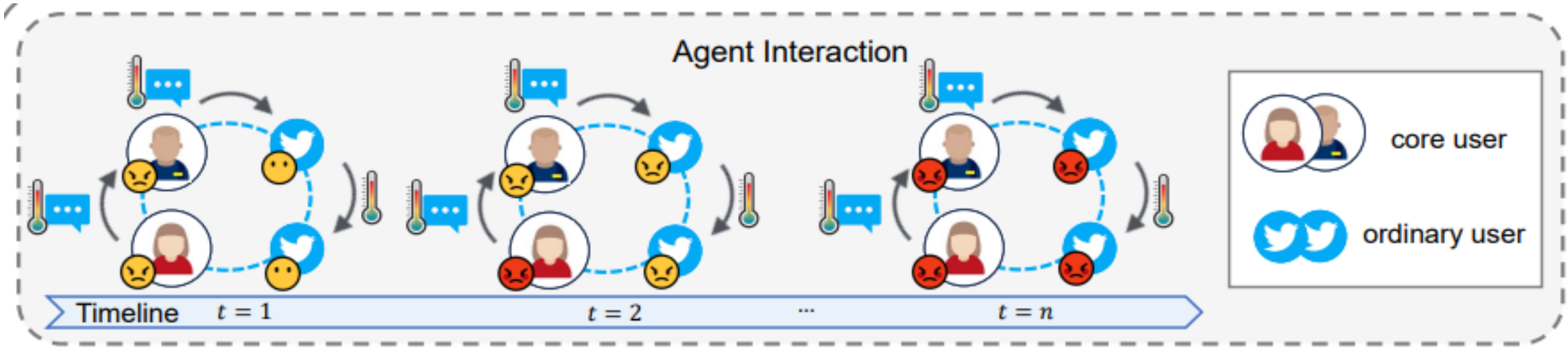
HiSim：混合社交媒体模拟框架

核心用户 ➔ LLM　　**Pareto 分布**　　普通用户 ➔ ABM



模拟环境

- **智能体之间的交互**
  - 同类智能体：LLM 智能体——文本；ABM 智能体——态度分数
  - 不同类智能体：将文本转化为态度分数
- **消息反馈机制**
  - 模仿真实的 Twitter 社交网络
- **离线新闻提要**
  - 通过自然语言向核心用户体提供触发事件描述



SoMoSiMu-Bench： 社会运动模拟基准

- **数据集**
  - Metoo, RoeOverturned (Roe), BlackLivesMatters (BLM);
  - 每个数据集有两个事件/阶段

- **用户选取**
  - 根据每个事件的活跃度和影响力选择 300 个用户;
  - 随机抽取 700 名普通用户
  - *受注释成本而非模拟成本的限制*

Dataset	Event	#Users	#Tweets	Time Span
Metoo	E1	1,000	18,638	Oct 15 - Oct 22, 2017
	E2	1,000	13,291	Jan 06 - Jan 13, 2018
Roe	E1	1,000	61,687	May 02 - May 09, 2022
	E2	1,000	59,829	Jun 24 - Jul 01, 2022
BLM	P1	1,000	10,710	May 25 - Jun 01, 2020
	P2	1,000	21,480	Jun 02 - Jun 09, 2020

Table 1: Statistics of our dataset. In *Metoo*, E1 is *American actress Alyssa Milano starts the #Metoo movement* and E2 is *#Timesup campaign on the 2019 Golden Globes Awards*; In *Roe*, E1 is *The leakage of the Supreme Court draft opinion* and E2 is *The Supreme Court overturns Roe v. Wade*; In *BLM*, we include two phases after the *Murder of George Floyd*.

SoMoSiMu-Bench： 社会运动模拟基准

- **微观对齐评测**
  - 立场对齐： *支持、中立、反对*
  - 内容对齐： *呼吁行动、分享观点、提及第三方、提供证据、……*
  - 行为对齐： *发帖、转推*
- **宏观系统评测**
  - 静态态度分布： $\Delta Bias, \Delta Div.$
  - 平均态度时间序列： *DTW, Corr.*

  设置： 校准和验证

微观对齐评测

- **立场：** 能够模仿核心用户的立场，但难以生成不支持的内容
- **内容：** *呼吁行动*和*分享观点*表现较好，*提供证据*表现较差，因为他们**缺乏用户的线下体验**
- **行为：** 能够区分原创内容作者和转发者

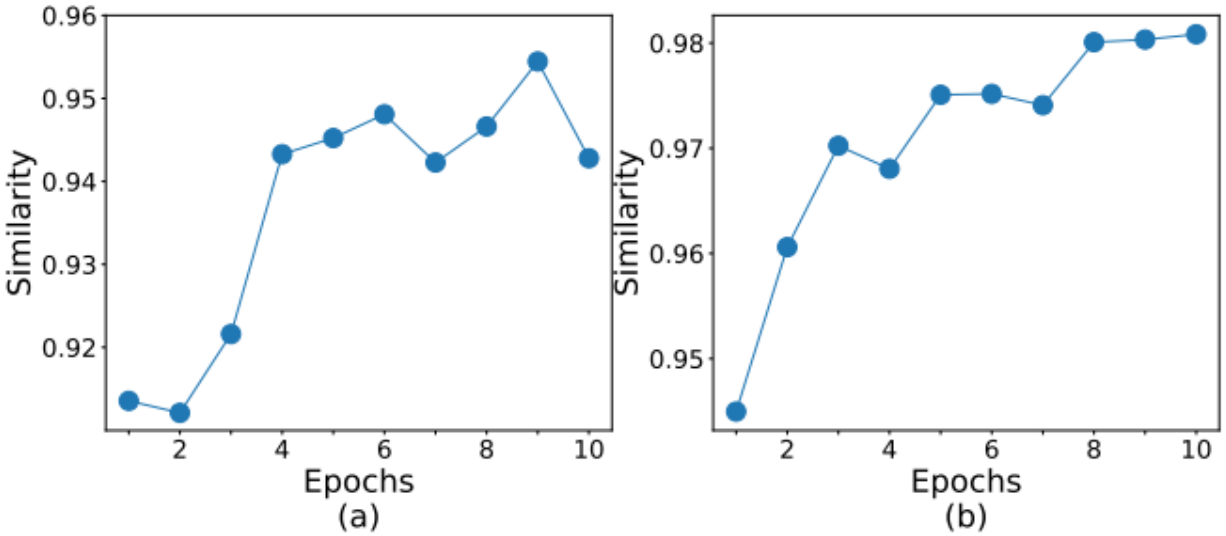
Datasets	Stance			Content			Behavior	
	Acc.	F1	MAE	Acc.	F1	Sim.	Acc.	F1
Metoo	0.9679	0.3400	0.2311	0.7010	0.1988	0.8064	0.7313	0.5212
Roe	0.9430	0.3361	0.2058	0.6423	0.1957	0.8090	0.6665	0.4691
BLM	0.8991	0.3735	0.1627	0.7353	0.2218	0.8406	0.7796	0.5759

Table 2: Results of micro alignment evaluation.

Unveiling the Truth and Facilitating Change: Towards Agent-based Large-scale Social Movement Simulation， ACL 2024

回音室（Echo Chamber）的复制与干预

- **同质性：**消费和生产内容的相似性
- **毒性：**用户生成内容的毒性
- **干预：**
  - S1：向用户推送具有对立观点的推文
  - S2：向用户推送具有中立观点的推文
  - S3：提供中立标签并鼓励用户使用进行讨论



Method	Avg. Homogeneity	Avg. Toxity
S1	**0.8551**	0.1426
S2	0.8580	0.1296
S3	0.8962	**0.1163**

Table 4: Results of solutions to break the echo chambers. **Bold** presents the best performance in the column.

ElectionSim：大规模选举仿真框架

- 提出了一个**大规模群体选举模拟框架**，实现多样化的选举模拟场景
- 构造了一个**百万级别用户池**，支撑大规模、多样化人群生成
- 设计了一个**大选Benchmark**，系统性评估选举模拟结果
- 搭建了一个**支持交互式分析的可视化系统***，允许与智能体进行多轮交互

*系统地址：
http://www.fudan-disc.com/electionsim/

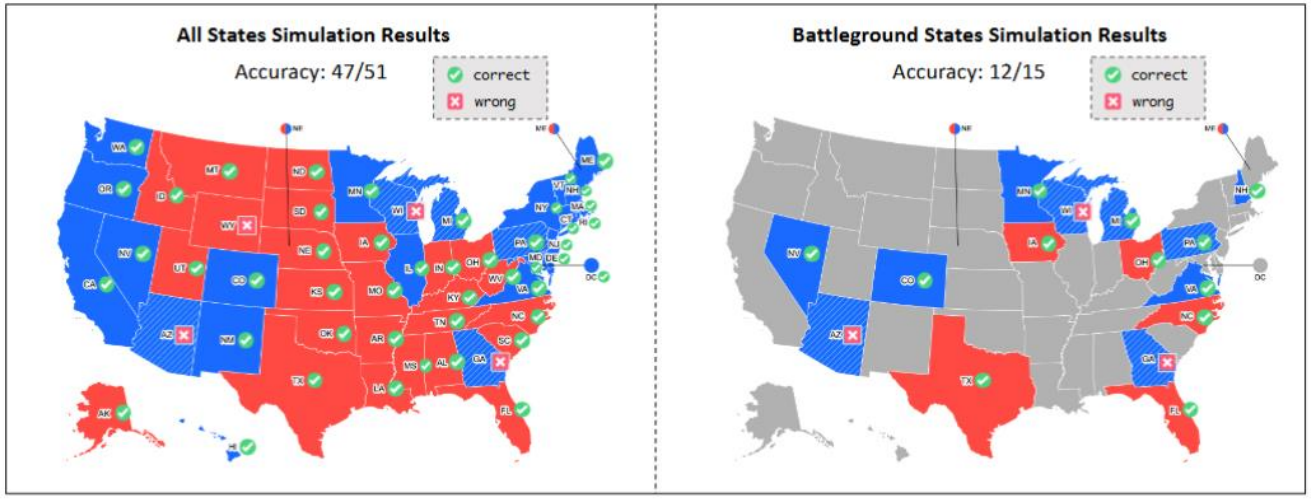
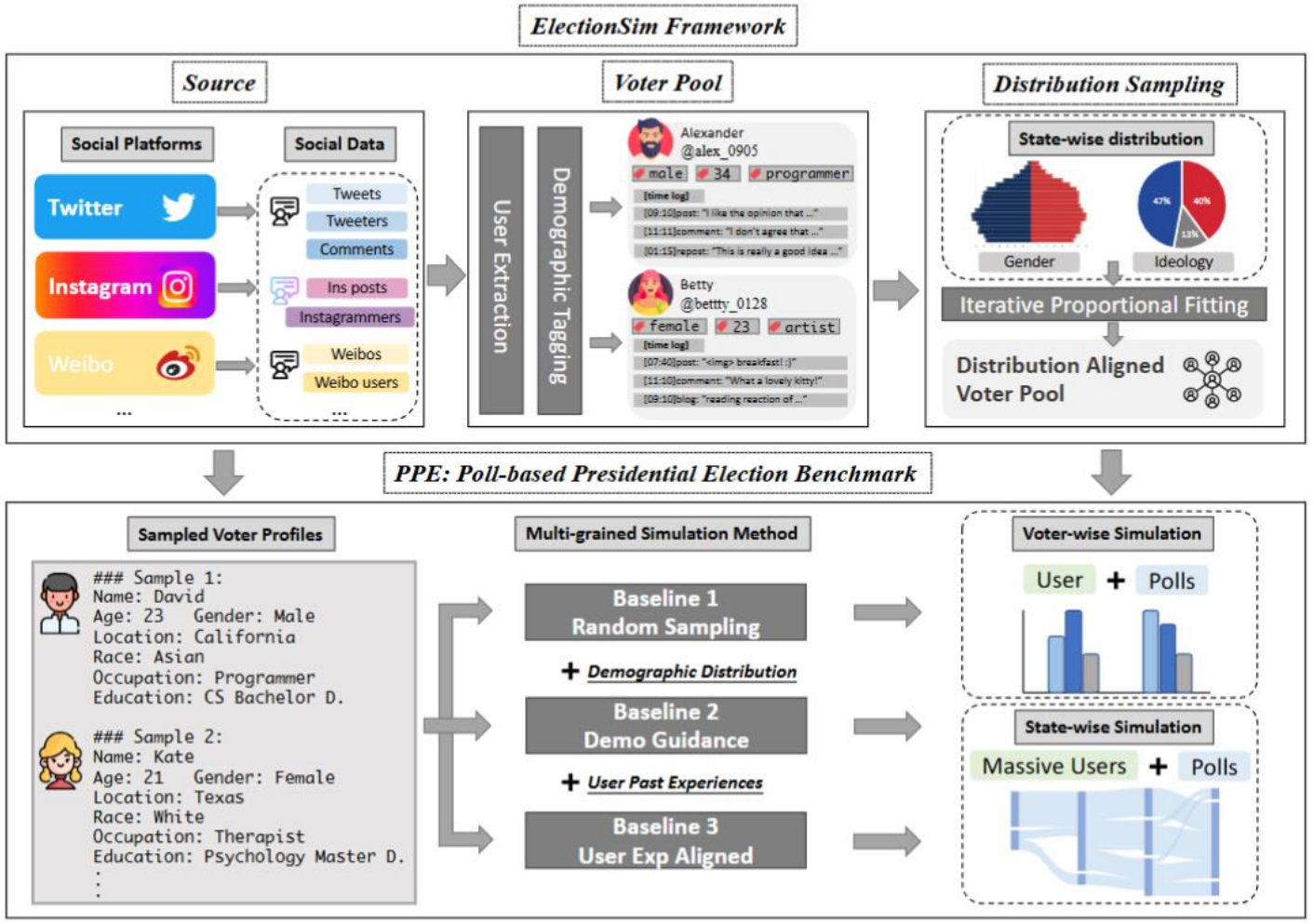


Figure 1: Simulation results of the 2020 Presidential Election. The colors represent the real-world results and the marks represent the simulation results accuracy.

ElectionSim: Massive Population Election Simulation Powered by Large Language Model Driven Agents, Arxiv 2024

ElectionSim：大规模选举仿真框架

- 大规模、多样化用户池
  - 超100w用户
  - 1.7亿条推特数据
- 丰富的人口统计学信息
  - 源于真实世界信息
  - IPF算法迭代
- 融合多粒度信息的模拟
  - 用户经验、历史发言
  - 选举人、时间信息



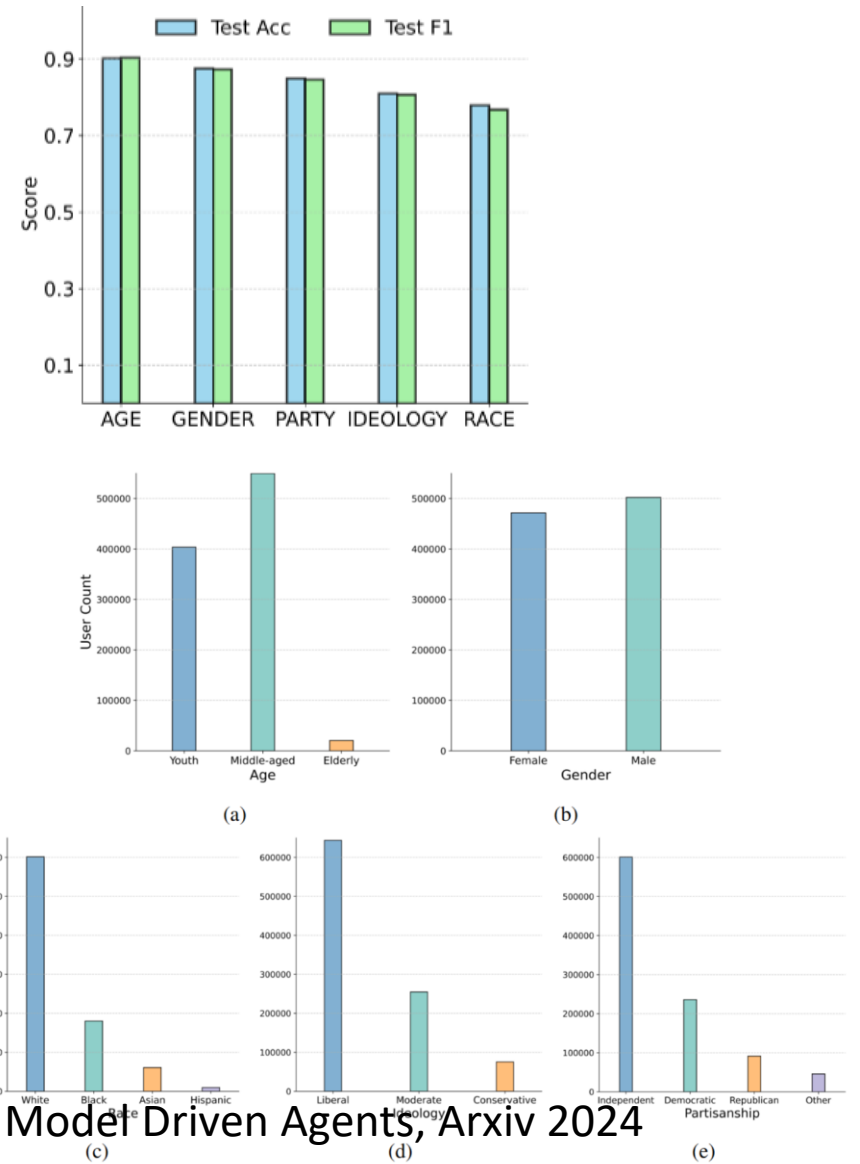
ElectionSim: Massive Population Election Simulation Powered by Large Language Model Driven Agents, Arxiv 2024

用户池构建

- **数据收集**
  - 数据源：推特
  - 2020. 1. 1–2020. 12. 29
- **用户抽取**
  - 数据聚合
  - 用户筛选与清洗
- **人口统计学特征标注**
  - 关注特征：个体属性/政治倾向

Dimension	Attribute	Classification
Personal Traits	Gender	Male Female
	Age	Youth (18-35 years old) Middle-aged (36-65 years old) Elderly (over 65 years old)
	Race	White Black Asian Hispanic
Political Orientation	Ideology	Liberal Moderate Conservative
	Partisanship	Democrat Republican Independent Others

Table 3: The Demographic Label Taxonomy



ElectionSim: Massive Population Election Simulation Powered by Large Language Model Driven Agents, Arxiv 2024

PPE: 基于民调的大选模拟Benchmark

- 数据来源
  - 民调研究：ANES 2020
  - 人口普查：U.S. Census Bereau
- 数据来源
  - 话题筛选：投票行为、政治、经济、文化等
  - 选项合并与问题转化
- 不同粒度评测
  - 个体级别：单一个体的民调结果是否准确？
  - 州级别：聚合后的选举结果是否与真实情况一致？

Number of questions	Topic
≥ 5	Democratic Norms, Immigration, LGBTQ+ Rights
4	Environment, Government
2	Criminal Justice, Education, Gender Resentment, Health Care, Social Welfare, US Position in World
1	Abortion, Aid to Blacks, Aid to Poor, Defense, Economy, Election Integrity, Inequality, Infrastructure, Parental Leave, Social Security, Taxes, Unrest, Voting Behavior

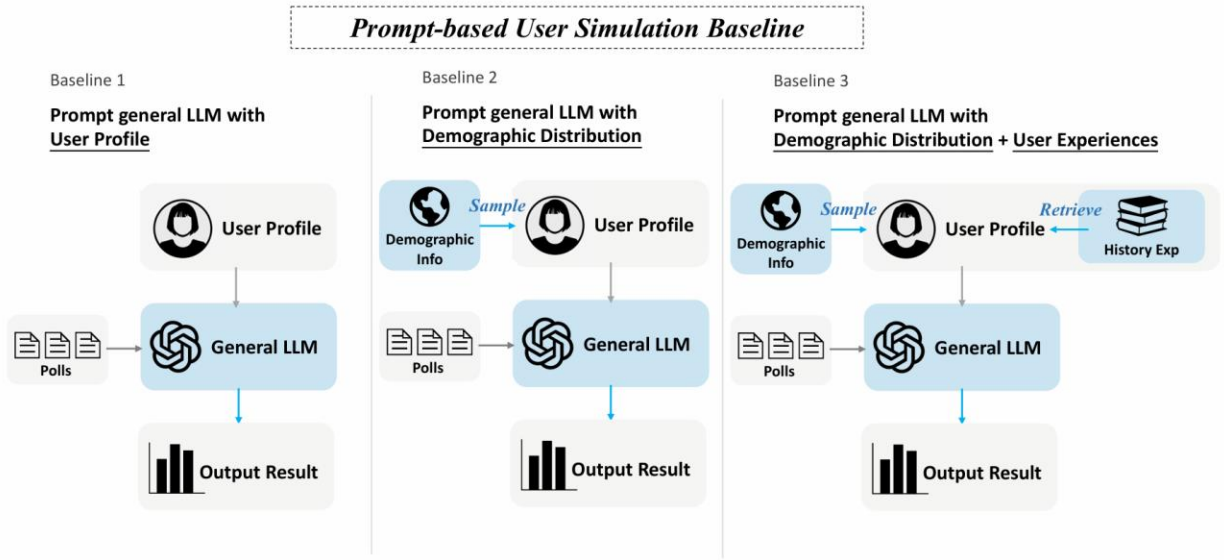
Table 7: Selected questionnaire topics

Original Question	Revised Question
Some people feel the government in Washington should see to it that every person has a job and a good standard of living. Suppose these people are at one end of a scale, at point 1. Others think the government should just let each person get ahead on their own. Suppose these people are at the other end, at point 7. And, of course, some other people have opinions somewhere in between, at points 2, 3, 4, 5, or 6. Where would you place yourself on this scale, or haven't you thought much about this? -9. Refused -8. Don't know 1. Government should see to jobs and standard of living 2. 3. 4. 5. 6. 7. Government should let each person get ahead on own 99. Haven't thought much about this	Some people feel the government in Washington should see to it that every person has a job and a good standard of living. Others think the government should just let each person get ahead on their own. And, of course, some people have a neutral position. Which of the following best describes your view? -2. DK/RF 1. Government should see to jobs and standard of living 2. Neutral 3. Government should let each person get ahead on own
Which party do you think would do a better job of handling the nation's economy? -9. Refused -8. Don't know 1. Democrats would do a much better job 2. Democrats would do a somewhat better job 3. Not much difference between them 4. Republicans would do a somewhat better job 5. Republicans would do a much better job	Which party do you think would do a better job of handling the nation's economy? -2. DK/RF 1. Democrats would do a better job 2. Not much difference between them 3. Republicans would do a better job

ElectionSim: Massive Population Election Simulation Powered by Large Language Model Driven Agents, Arxiv 2024

评测方案介绍

- **Baseline 1:随机采样**
  - 从用户池中按既定采样容量随机采样
- **Baseline 2:融合人口统计学信息**
  - 基于各州人口统计学分布进行采样
- **Baseline 3:融合用户经验**
  - 将用户历史发言、时间信息等融入模拟



ElectionSim: Massive Population Election Simulation Powered by Large Language Model Driven Agents, Arxiv 2024

评测内容1：个体级别模拟

- 实验设定
  - 研究对象：ANES 2020的8280个受访者
  - 标签确定：基于真实回答确定受访者的人口统计学标签
  - 问卷分类：全量问卷、投票行为子集
  - 评估每个受访者在典型问题上的模拟结果与真实情况的一致性
- 评测指标
  - 排除真实情况为拒绝回答的样本后进行评测
  - 指标：Micro-F1,Macro-F1

ElectionSim: Massive Population Election Simulation Powered by Large Language Model Driven Agents, Arxiv 2024

实验结果

- 整体模拟表现
  - 个体层面的**模拟准确率超过70%**
  - 70b开源大模型的性能媲美甚至超过商用大模型

- 选举强相关问题表现
  - 投票行为子集的模拟准确率**超过80%**
  - 投票行为的模拟在类别间更平衡（Macro-F1更高）

Type	Model	Overall		Voting Subset	
		Micro-F1	Macro-F1	Micro-F1	Macro-F1
Commercial	GPT-4o	**76.16**	55.97	**81.20**	61.03
	GPT-4o-mini	<u>75.39</u>	58.18	80.26	74.72
	Claude-3.5-Sonnet	73.65	58.70	77.52	71.95
Open-source	Qwen2-7b-Instruct	67.53	43.31	76.39	65.65
	Qwen2-72b-Instruct	74.77	<u>58.71</u>	78.39	77.95
	Qwen2.5-72b-Instruct	74.97	57.81	<u>80.41</u>	**79.27**
	Llama3-70b-Instruct	74.86	**59.96**	80.16	<u>79.17</u>

Table 9: Model performance on voter-wise simulation. We compared the performance of commercial and open-source LLMs on both the overall test set (**Overall**) and the voting-related subset (**Voting Subset**). The best results are **bolded**, and the second-best results are <u>underlined</u>.

ElectionSim: Massive Population Election Simulation Powered by Large Language Model Driven Agents, Arxiv 2024

评测内容2：州级别模拟

- 实验设定
  - 场景：2020年美国大选
  - 采样比例：1/1000，1/10000
  - 按各州采样容量从用户池进行采样，对每个个体单独进行民调问答后，将回答聚合成各州模拟结果
- 评测指标
  - 投票结果一致性（CER）：粗粒度指标，州模拟结果与真实投票结果的一致性
  - 相对得票率一致性（CVS）：细粒度指标，州模拟结果的相对得票率与真实得票率的均方误差

ElectionSim: Massive Population Election Simulation Powered by Large Language Model Driven Agents, Arxiv 2024

实验结果

- Qwen2.5-72b-instruct成功预测47个州结果，并成功预测12个摇摆州结果（遵循CNN2020共15个摇摆州的设定）
- 在摇摆州的得票率平均误差为4%，且在代表州上的平均相对误差小于ABM预测结果

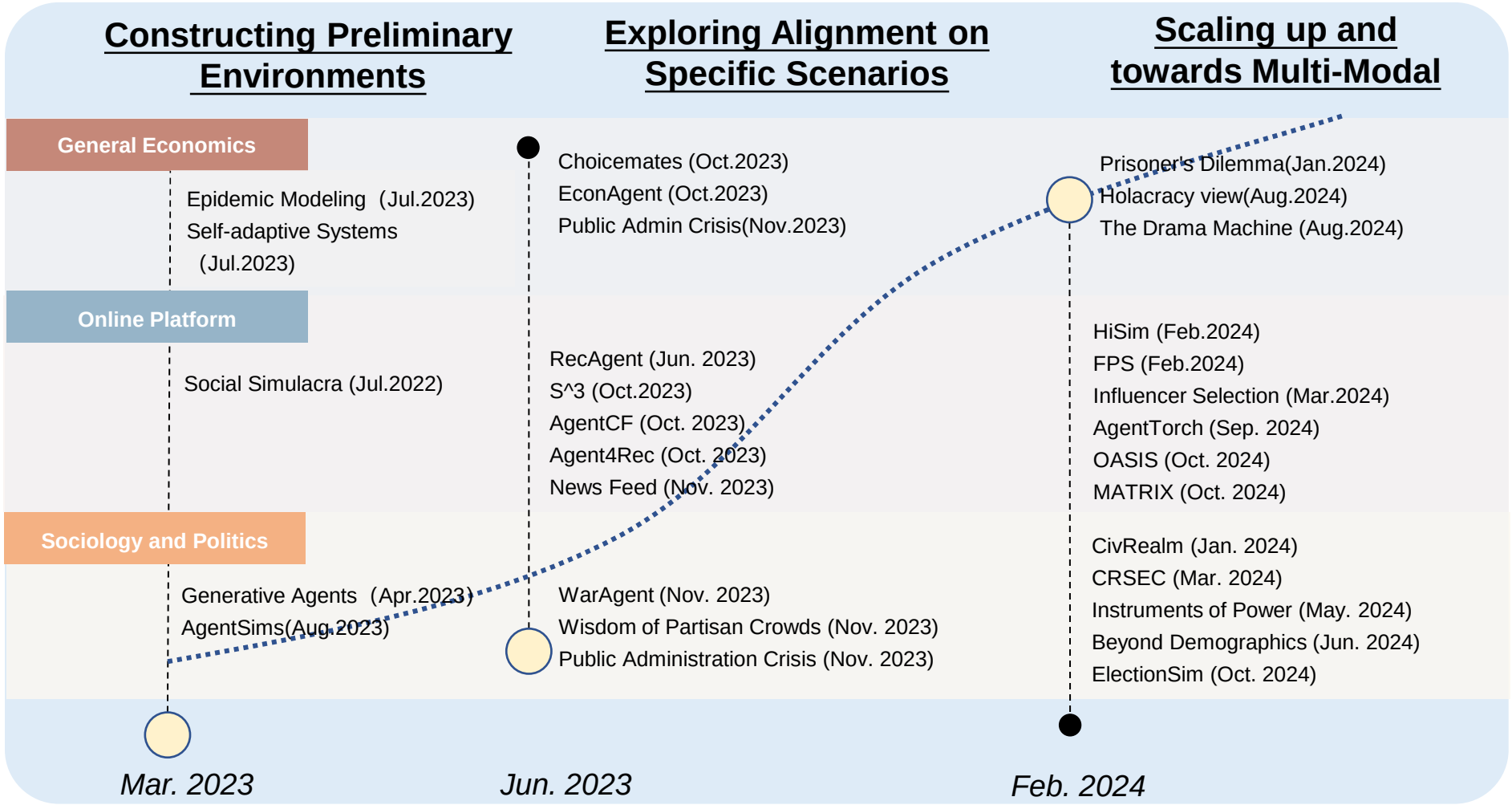
Model	Method	Overall		Battleground	
		CER ↑	CVS ↓	CER ↑	CVS ↓
Llama3-70b-Instruct	Baseline1	0.510	0.399	0.600	0.386
	Baseline2	0.745	0.118	0.533	0.093
	Baseline3	0.843	0.094	0.733	0.065
Qwen2.5-72b-Instruct	Baseline1	0.510	0.383	0.600	0.370
	Baseline2	0.843	0.078	0.733	0.054
	Baseline3	0.902	0.071	0.733	0.045
	Baseline3*	**0.922**	**0.070**	**0.800**	**0.042**
GPT-4o-mini	Baseline1	/	/	0.667	0.323
	Baseline2	/	/	**0.800**	0.052
	Baseline3	/	/	**0.800**	0.056

Table 10: Model performance on state-wise simulation. We evaluate different methods for their accuracy in forecasting the 2020 U.S. Presidential Election across all 51 states (**Overall**) and 15 battleground states (**Battleground**). The CER measures state-level prediction accuracy, while CVS denotes the RMSE of simulated versus actual vote shares. *: Building on Qwen2.5-72b's strong performance on Baseline3, we extend its use to a 1/1000 population sample (around 300,000 agents). This approach effectively predicts outcomes in **47 states** and **12 battleground states**, with reduced RMSE in vote share predictions.

State	Candidates	Ours		ABM		Actual Result	
		Relative Vote Share	Winner	Relative Vote Share	Winner	Relative Vote Share	Winner
MI	Biden-Harris	0.5412	⋆	0.5454	⋆	0.5142	⋆
	Trump-Pence	0.4588		0.4546		0.4858	
OH	Biden-Harris	0.4371		0.4925		0.4589	
	Trump-Pence	0.5629	⋆	0.5075	⋆	0.5411	⋆
PA	Biden-Harris	0.5280	⋆	0.5204	⋆	0.5061	⋆
	Trump-Pence	0.4720		0.4796		0.4939	
IN	Biden-Harris	0.4652		0.4835		0.4184	
	Trump-Pence	0.5348	⋆	0.5165	⋆	0.5816	⋆
WV	Biden-Harris	0.3811		0.3831		0.3022	
	Trump-Pence	0.6189	⋆	0.6169	⋆	0.6978	⋆
MO	Biden-Harris	0.4603		0.4440		0.4216	
	Trump-Pence	0.5397	⋆	0.5560	⋆	0.5784	⋆
RMSE		**0.0439**		**0.0476**			

Table 13: Comparison of GPT-4o-mini simulation results in 6 states with the ABM method and actual results. The reporting states are Michigan (MI), Ohio (OH), Pennsylvania (PA), Indiana (IN), West Virginia (WV), and Missouri (MO).

社会模拟发展趋势


Constructing Preliminary Environments
Exploring Alignment on Specific Scenarios
Scaling up and towards Multi-Modal
General Economics
Epidemic Modeling (Jul.2023)
Self-adaptive Systems (Jul.2023)
Choicemates (Oct.2023)
EconAgent (Oct.2023)
Public Admin Crisis(Nov.2023)
Prisoner's Dilemma(Jan.2024)
Holacracy view(Aug.2024)
The Drama Machine (Aug.2024)
Online Platform
Social Simulacra (Jul.2022)
RecAgent (Jun. 2023)
S^3 (Oct.2023)
AgentCF (Oct. 2023)
Agent4Rec (Oct. 2023)
News Feed (Nov. 2023)
HiSim (Feb.2024)
FPS (Feb.2024)
Influencer Selection (Mar.2024)
AgentTorch (Sep. 2024)
OASIS (Oct. 2024)
MATRIX (Oct. 2024)
Sociology and Politics
Generative Agents (Apr.2023)
AgentSims(Aug.2023)
WarAgent (Nov. 2023)
Wisdom of Partisan Crowds (Nov. 2023)
Public Administration Crisis (Nov. 2023)
CivRealm (Jan. 2024)
CRSEC (Mar. 2024)
Instruments of Power (May. 2024)
Beyond Demographics (Jun. 2024)
ElectionSim (Oct. 2024)
Mar. 2023
Jun. 2023
Feb. 2024

从个体到社会：
大模型智能体驱动的社会模拟

魏忠钰（Wei, Zhongyu）

复旦大学

数据智能与社会计算实验室 (Fudan DISC)

自然语言处理组 (Fudan-NLP)

2024年12月01日

第三届全国大模型智能生成大会（LMG 2024）讲习班