前沿动态综述

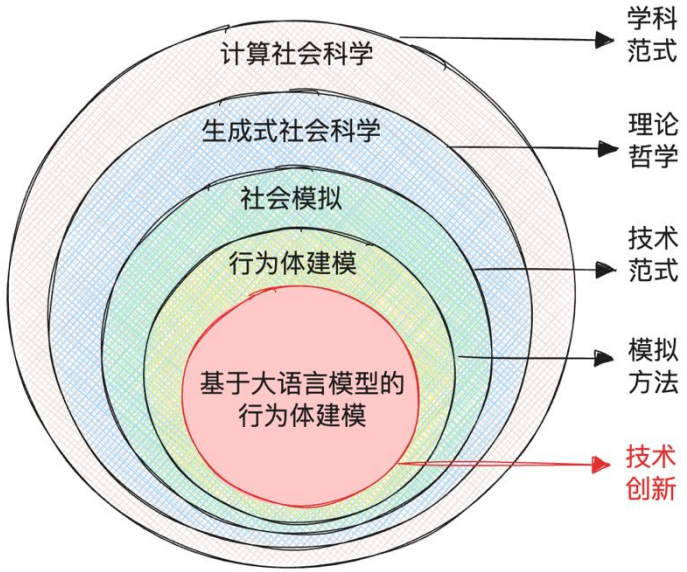
基于大语言模型的社会模拟

冯夏冲 香港大学

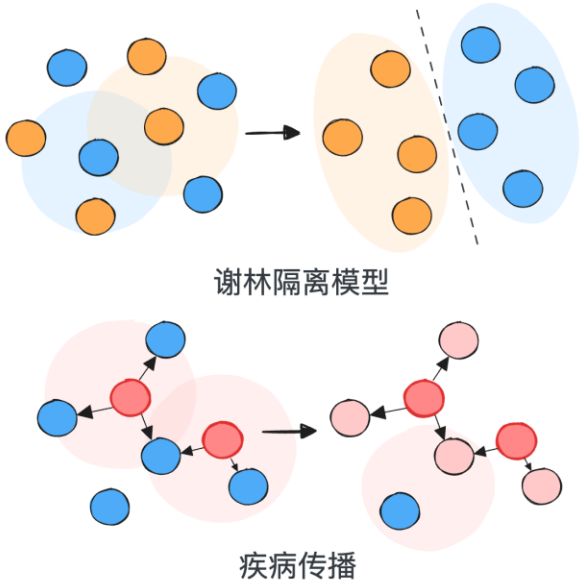
2025年8月14日

社会模拟

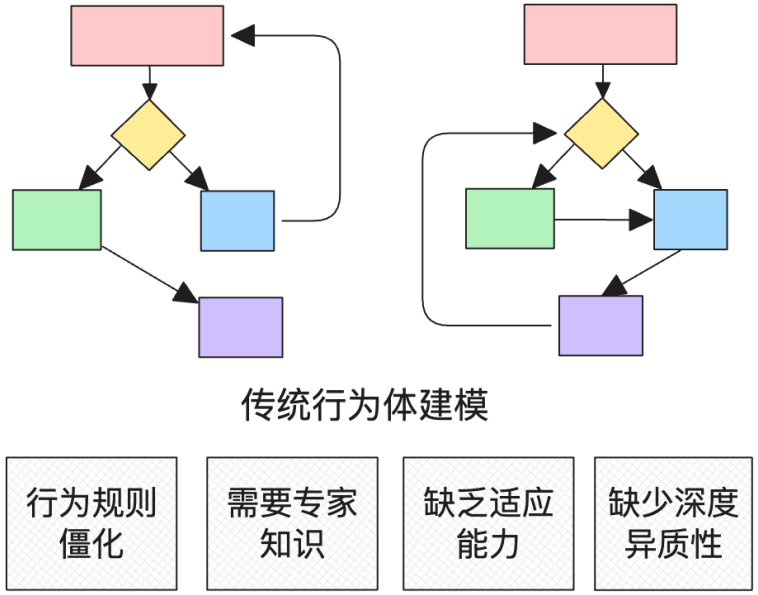
- **定义**：通过构建一个由拥有互动规则的微观主体组成的人工社会来探索宏观现象涌现机制的计算研究方法论
- **目的**：解释复杂社会现象的因果机制，推演多种可能性以预测未来情景
- **方法**：基于行为体的建模（Agent-based Modeling，ABM）
- **应用**：社会学与政治学、流行病学、交通系统、反事实验证和敏感场景模拟等



概念分类



应用领域



传统方法缺陷

类人的大语言模型

内在心智的类人性

诺尔是一家繁忙咖啡店的咖啡师。她想为一位点了燕麦奶的顾客做一杯美味的拿铁。诺尔拿了一个奶缸，在里面装满了**燕麦奶**。一位没有听到顾客要求的同事，趁诺尔在处理另一项任务时，**将奶缸里的燕麦奶换成了杏仁奶**。

情景1：错误信念

诺尔**没有看到**同事更换牛奶。

诺尔认为奶缸里装的是什么？

诺尔认为奶缸里装的是**燕麦奶**。

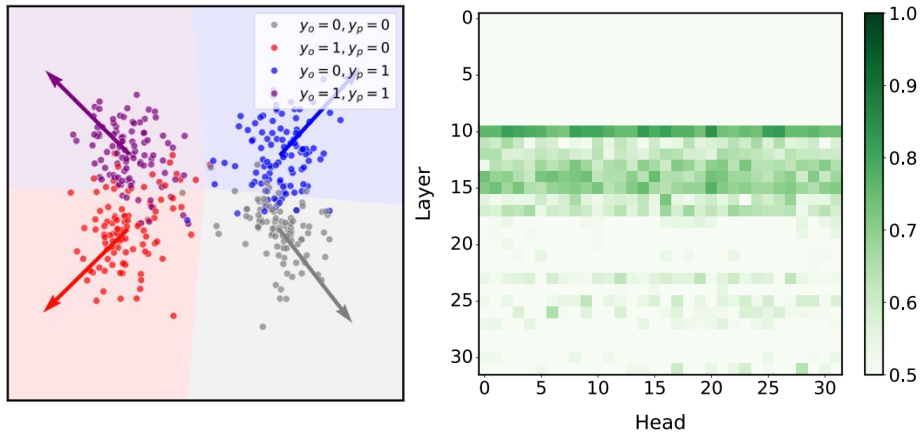
情景1：真实信念

诺尔**看到了**同事更换牛奶。

诺尔认为奶缸里装的是什么？

诺尔认为奶缸里装的是**杏仁奶**。

探针实验证明模型内部具备心智状态



外在行为的类人性

独裁者博弈

独裁者博弈是一项经济学实验，其中一名玩家独裁者单方面决定如何与另一名玩家分给定的金钱，而玩家必须接受该

Dictator's Decision	Dictator's Payoff	Recipient's Payoff
Keeps all	$10	$0

最后通牒博弈

最后通牒是一项议价实验，其中一名玩家向另一名玩家提出一个分配方案，另一名玩家可以选择接受或拒绝。

Proposer's Offer	Responder Accepts?	Proposer's Payoff	Responder's Payoff
$8 / $2	Yes	$8	$2
$8 / $2	No	$0	$0
$2 / $8	Yes	$2	$8

公共物品博弈

公共物品博弈是一项实验，玩家们向一个能使所有人受益的共享池中投资源，但有些人能会通过减少投来搭便车。

Player A's	Player B's	Player A's	Player B's

囚徒困境

囚徒困境是一个博弈论场景，其中个体在合作与背叛之间做出选择，以权衡个人和集体利益。

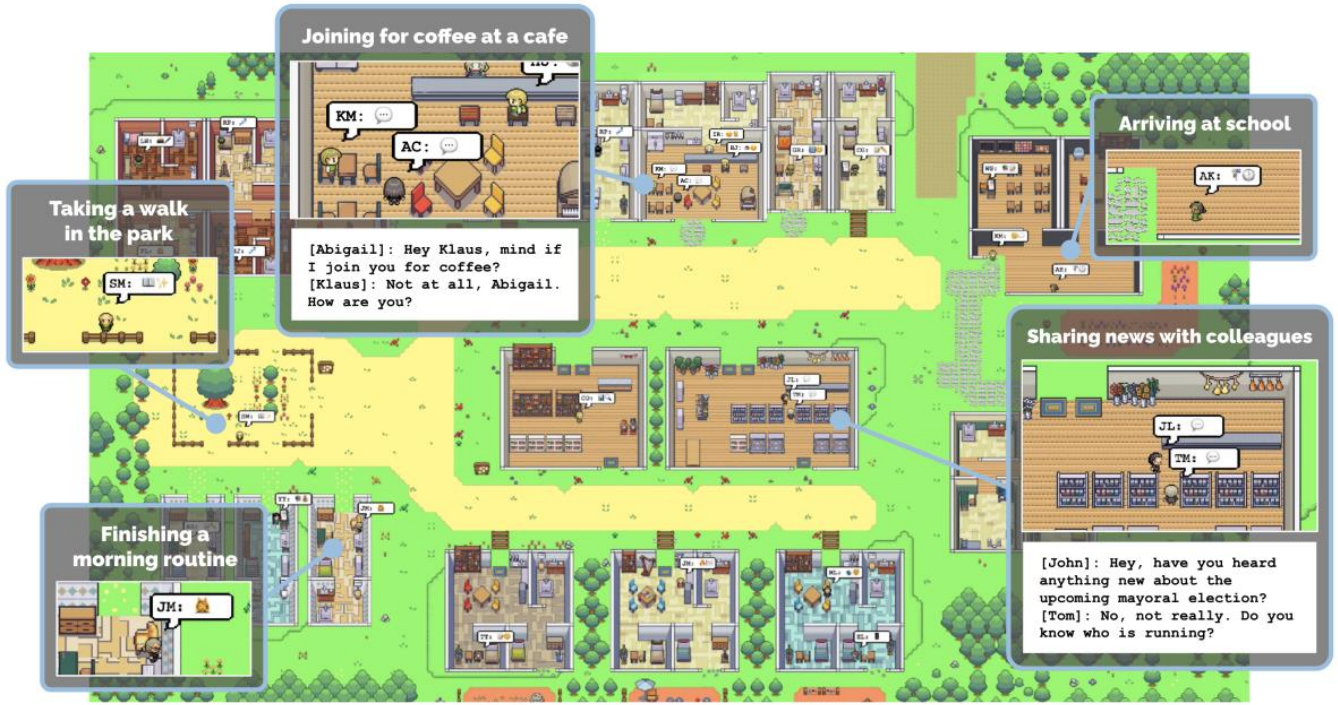
Payoff	Cooperate	Defect
Cooperate	(3, 3)	(0, 5)
Defect	(5, 0)	(1, 1)

基于大语言模型的社会模拟

- 将大语言模型作为大脑嵌入到虚拟的行为体中，来替代传统的手动编程规则，从而在人工社会里生成更真实、更复杂的动态人类行为。

传统行为体建模	基于LLM的行为体建模
行为规则过于简单且僵化	实现深度认知主义
开发成本高昂且需要专家知识	零样本行为生成
缺乏真正的适应性与学习能力	具备记忆、反思、推理能力
难以创建深度的异质性	角色扮演创建丰富异质性

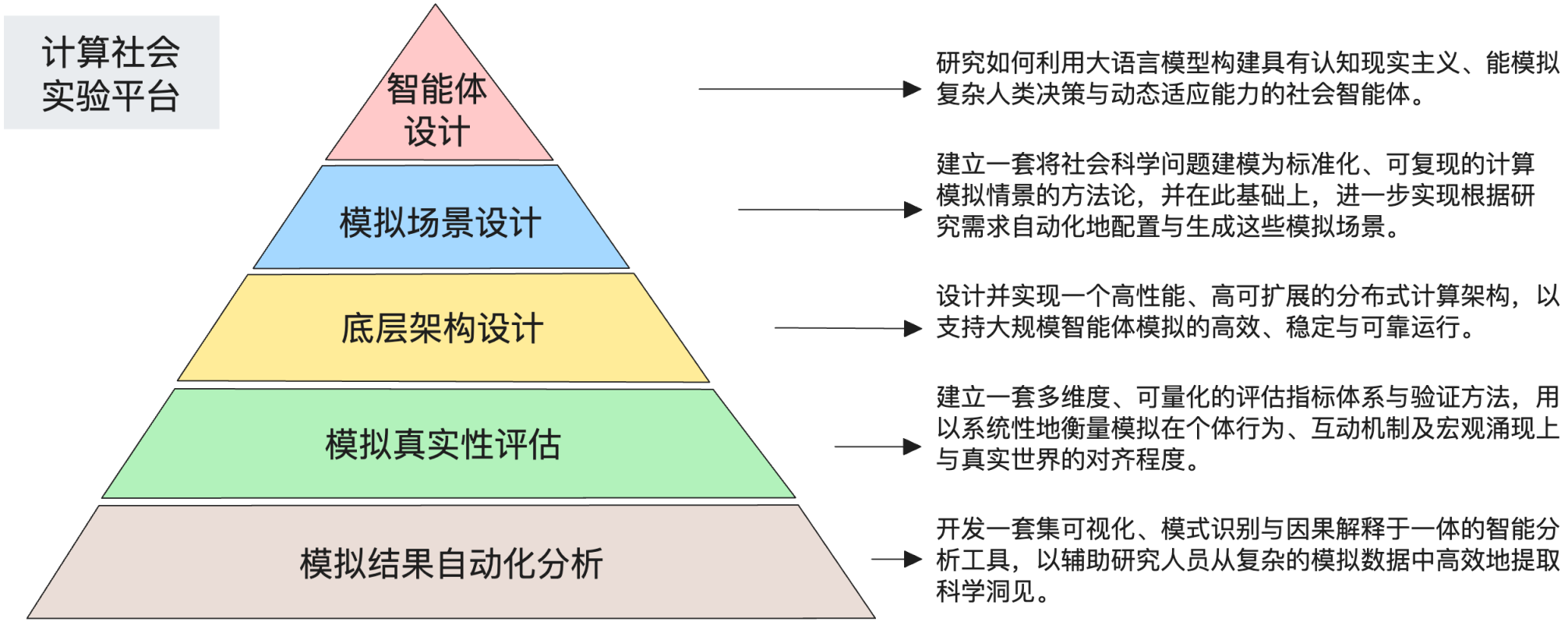
斯坦福小镇：可行可信



Generative Agents: Interactive Simulacra of Human Behavior

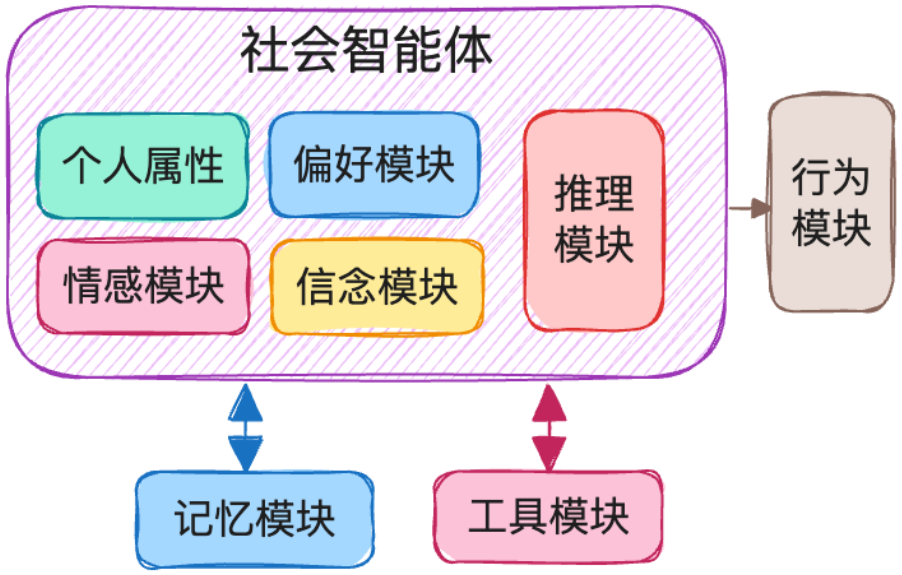
研究问题

- 如何优化智能体设计、模拟场景设计和底层架构设计，系统性地进行整合，从而构建真实性可被有效评估，研究人员高效易用的计算社会实验平台？

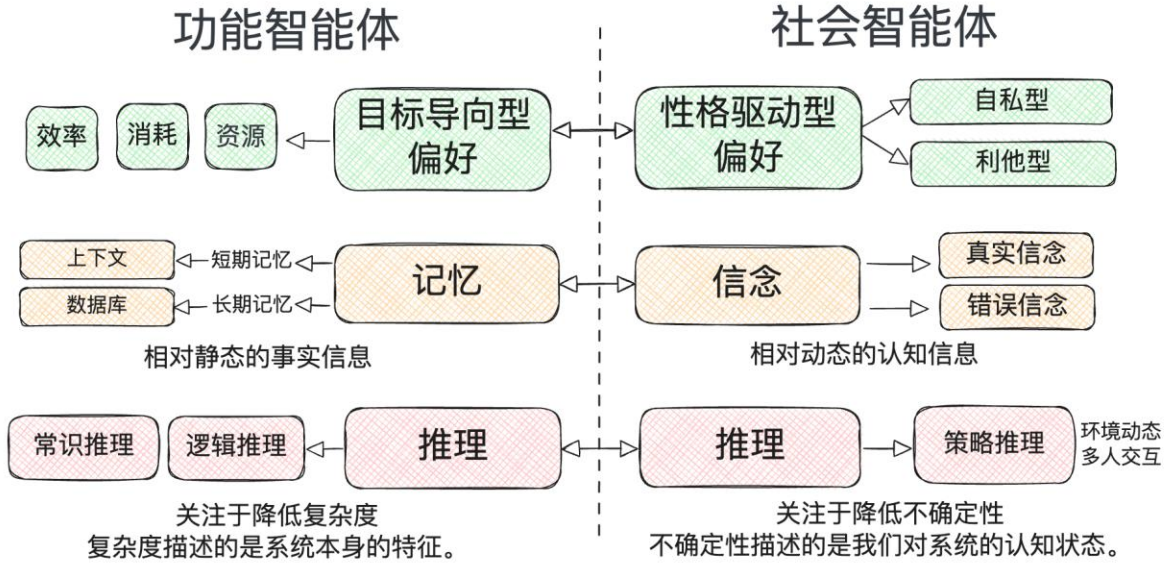


智能体设计

- 社会智能体是一种计算模型，其目标是研究如何利用大语言模型，构建一个在认知层面高度逼真、能够模拟人类决策与动态适应能力的虚拟实体。
- 通过将大语言模型置于推理模块的核心，并让其与定义好的多个模块协同工作，使其不仅能模仿行为，更能模拟思考，从而扮演一个可信、动态、富有深度的人。



常见社会智能体架构

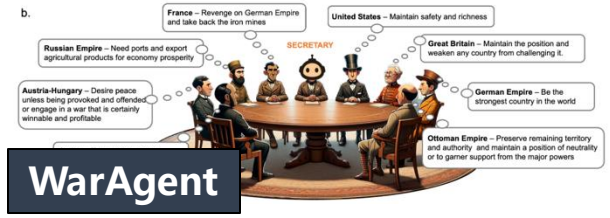


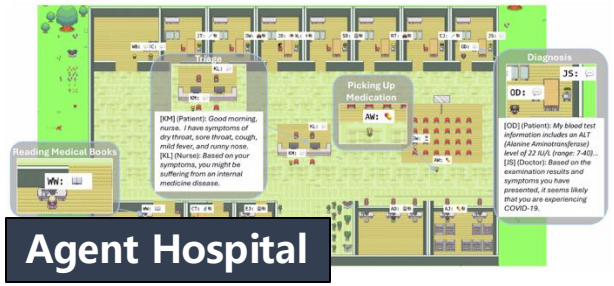
核心概念辨析

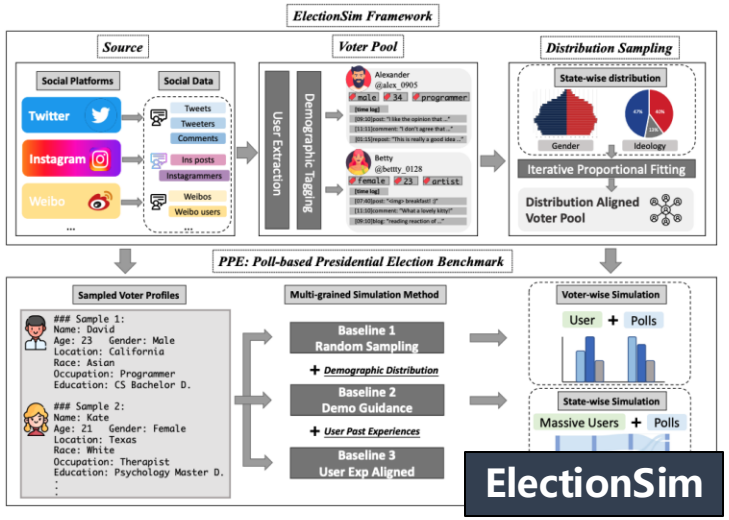
工作均在特定情境下证明了
社会模拟的可行性和可信性

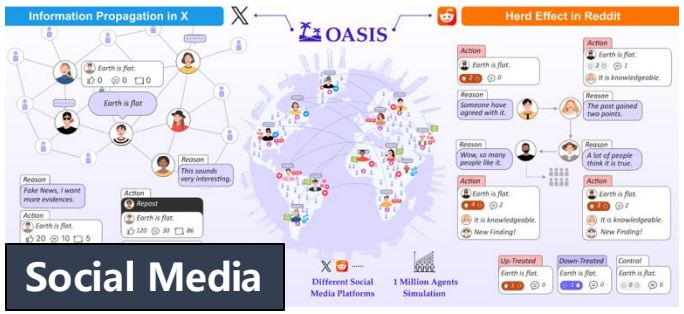
模拟场景设计：具体模拟场景

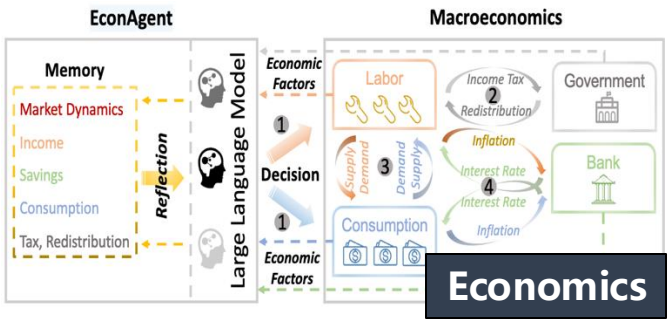
- **定义**：模拟环境构成了智能体赖以生存、互动和演化的数字化时空与规则体系。
- **分类**：物理环境，社交环境，经济环境等等。
- **作用**：提供上下文、介导交互、施加约束、记录状态等。











核心特点

场景驱动

解决挑战

智能体可行+模拟可信

引出需求

通用化规模化平台

模拟场景设计：通用社会模拟器

- 通用社会模拟器是一个与特定研究场景解耦的、高度可配置的框架，它为研究人员提供了一套标准化的工具和接口，用以高效地构建、运行和分析多样化的社会模拟情景。

易用性

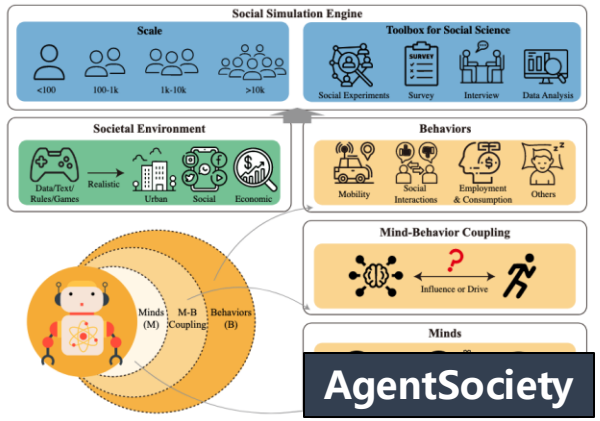
高度可配置+低代码构建

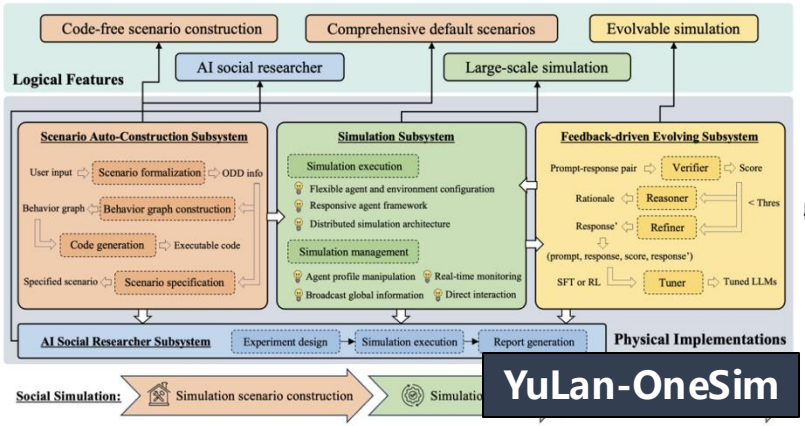
可扩展性

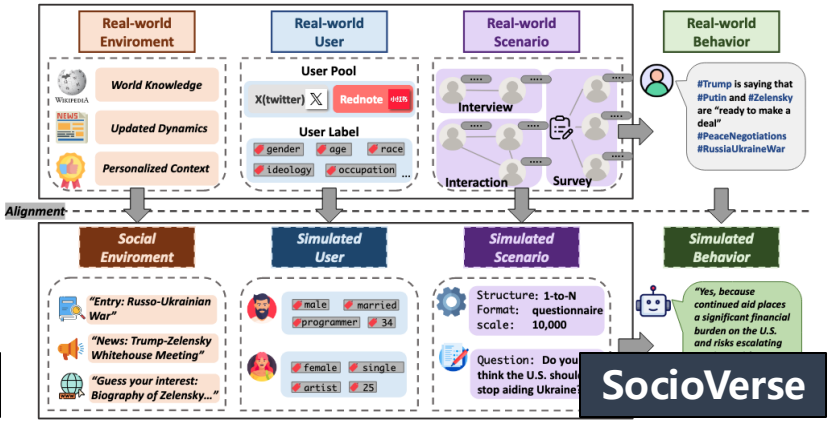
标准化组件+模块化

鲁棒性

大规模能力+分布式架构







模拟场景设计：无/低代码场景构建

- 无/低代码的场景构建旨在打破社会科学理论与计算模拟实践之间的技术壁垒，让社会科学家无需深厚的编程背景，即可将研究思想高效地转化为可执行的模拟实验。

场景形式化

将使用者模糊、抽象的自然语言需求，转化为一个结构化、详尽的标准描述，以便大型语言模型能够准确理解与处理。

我想模拟一个求职市场

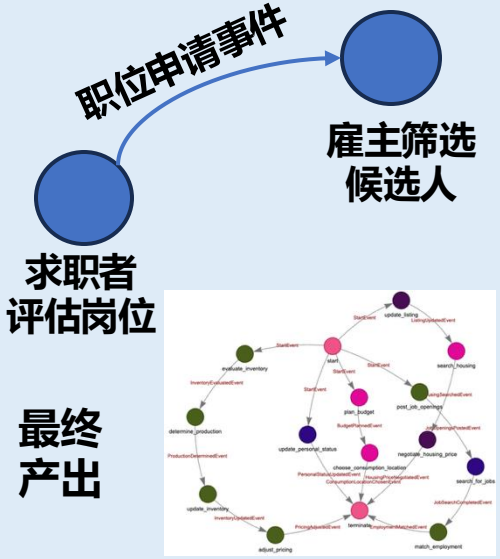
ODD智能体

市场中有哪些角色？他们的目标是什么？招聘流程是怎样的

最终产出 一份详细的ODD文档，清晰定义了求职者和雇主的目标、行为、互动规则

行为图构建

在ODD协议的基础上，以图形化的方式定义不同智能体之间的具体交互逻辑和行为流程。



最终产出

代码生成

将逻辑清晰但抽象的行为图，翻译成可直接运行的程序代码。

图引导生成

模板化引导

迭代验证与修复

最终产出


求职者类


雇主类

场景详述

为模拟场景填充具体的数据，这一步将代码与数据完全解耦，使得场景配置更加灵活。

逻辑层 — 求职者

数据结构层面 — 技能经验

具体数据层面 — Java 5年

最终产出


实例化好的智能体

采样角度：统计学
生成角度：智能化的群体合成

模拟场景设计：智能体配置

- 如何将通用的智能体模型实例化为一个符合特定研究目标的虚拟群体？其本质在于智能体的群体配置策略与实现机制，核心在于匹配研究问题+匹配真实世界。

迭代比例拟合

核心挑战

在仅掌握多维度人口宏观统计数据（即边缘分布），而缺乏个体层面交叉数据（即联合分布）时，如何生成一个与所有已知宏观统计都相符的微观虚拟人口。

应用案例

用于美国大选预测情景。结合美国人口普查局提供的各州性别、种族、年龄、党派等边缘分布数据，IPF算法成功合成了一个与各州人口结构高度吻合的、大规模的虚拟选民数据库，解决了缺乏具体交叉选民数据的问题。

相同分布采样

核心挑战

当目标群体的完整联合分布已知，且拥有一个足够庞大、多样化的基础用户池时，如何高效、无偏地抽取一个在多维特征上与目标群体完全一致的代表性样本。

应用案例

用于突发新闻反馈情景。由于研究者掌握了讨论ChatGPT的小红书用户的完整画像（即联合分布已知），因此采用IDS方法，从千万级用户池中直接抽取了一个特征分布完全一致的模拟群体，保证了样本的高度代表性。

混合分布建模

核心挑战

如何精确建模现实世界中如收入、财富等典型的高度右偏态数据？此类数据通常包含一个集中的主体和一条极长的长尾，少数极端高值，单一标准分布难以有效拟合。

应用案例

用于中国全国经济调查情景。通过该方法，研究者为31个地区生成了符合统计规律的虚拟居民收入数据，为后续的消费行为模拟提供了高质量的先验分布。

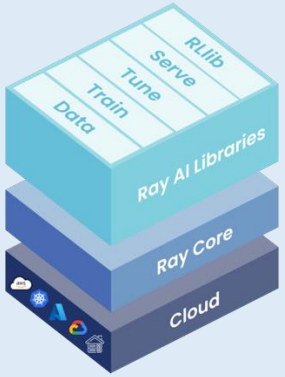
底层架构设计

- 底层架构旨在解决大规模社会模拟中的两大核心挑战：
  - 有效管理海量智能体带来的高并发问题
  - 为模拟真实社会行为所需的异步消息交互提供一个稳健、可扩展的通信机制

基于分组的分布式异步执行

在模拟上万个智能体时，如果每个智能体都作为一个独立的进程运行，会因创建过多的网络连接而耗尽系统的TCP端口资源（上限为65535个），导致模拟失败。

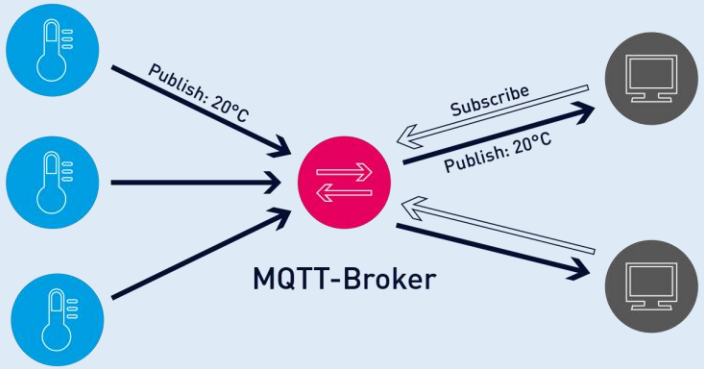
智能体分组



分布式与异步并行

基于MQTT的智能体消息系统

真实社会中，个体间的交流是异步且独立的，模拟器需要一个能支持海量智能体之间高并发、高吞吐量、可靠传递消息的系统。

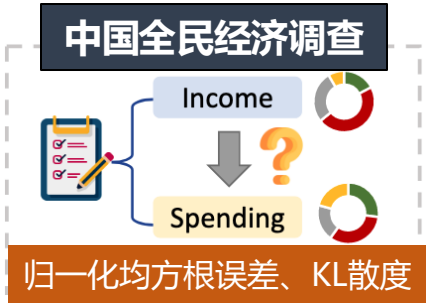


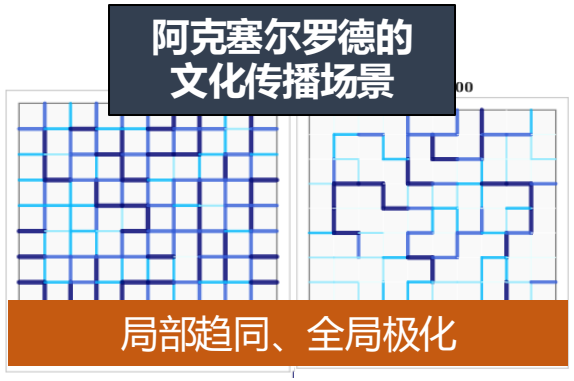
消息队列遥测传输协议

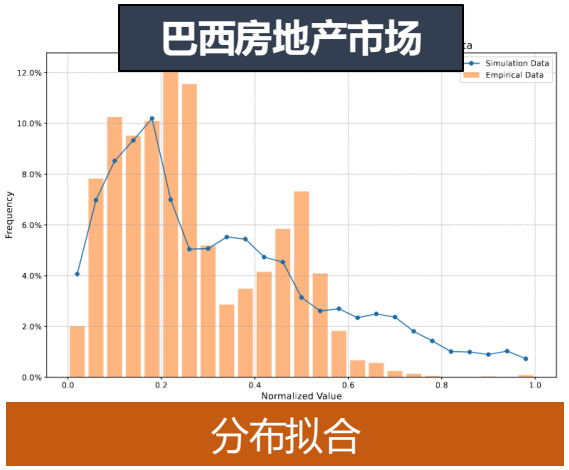
智能体的目标是像人而非是人
模拟只需把握研究问题的核心动力学

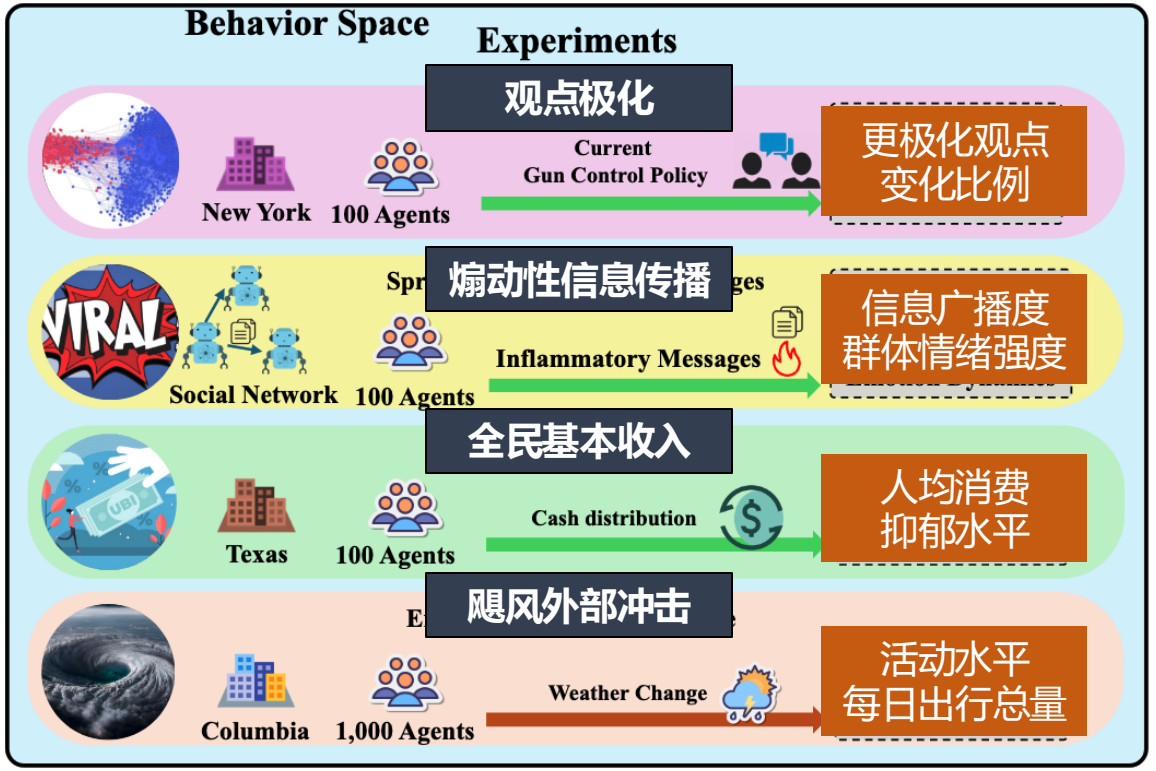
模拟评估

- 模拟的可信度评估，本质上是其结果与经典理论及物理现实的对齐验证。这一过程构建了从理论抽象到现实应用的桥梁，是确保模拟模型能够作为有效认知工具和决策依据的根本前提。









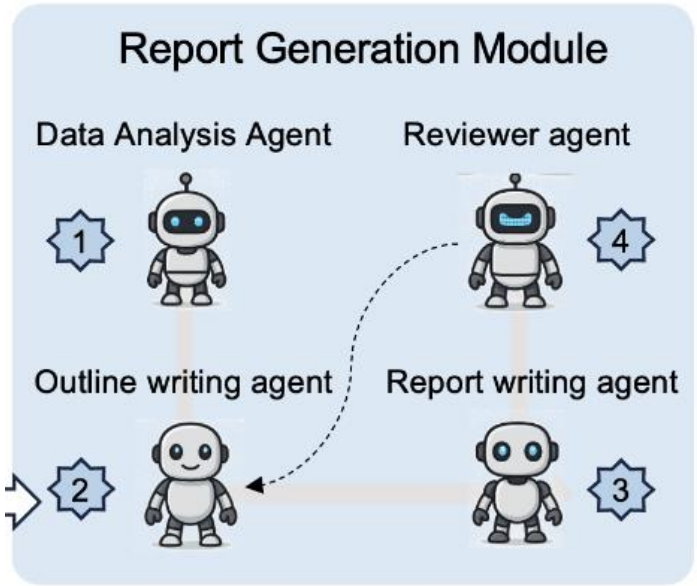
SocioVerse, YuLan-OneSim, AgentSociety

结果自动化分析

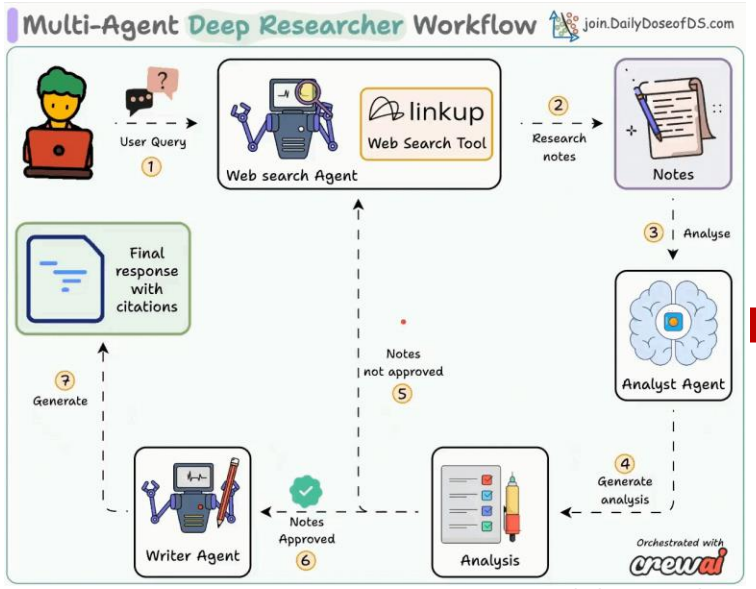
- 人工智能社会研究员是一个专攻模拟分析的自主智能体，其核心任务是通过深度挖掘数据中的因果机制与涌现模式，将复杂的模拟过程自动化地提炼为富有洞察力的技术分析报告，从而加速科学洞见的产出。

模拟结果

AI社会研究员



DeepResearch



@akshay_pachaar

DeepSocialResearch

宏观模式检测器	微观归因引擎
因果链条挖掘	交互式时空可视化
关键时刻识别	文本数据科学

文本分析+数据科学

未来方向

- 制约模拟真实性的三大技术瓶颈，源于模型固有的内在偏见（导致行为刻板化）、输出幻觉（导致决策失真），以及智能体异质性不足（导致群体行为均质化）。

消除内在偏见

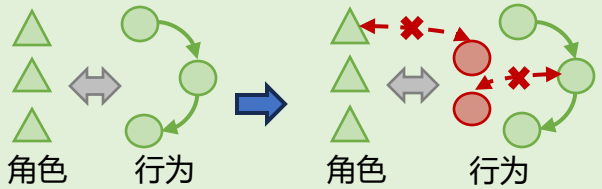
- 社会理论指导数据合成
- 无偏数据自动化构建
- 底层模型去偏微调
- 基于无偏模型的角色构建

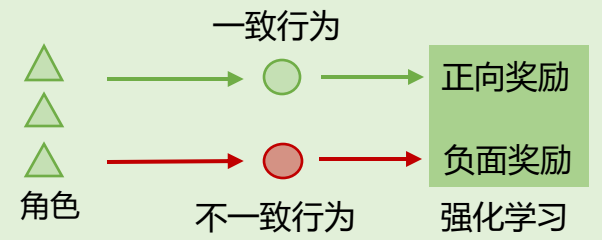
角色： 李经理与王工程师。
情景： 两人作为联合负责人，在一个有高层支持的紧急项目中被迫紧密合作。
要求： 接触理论，该理论认为不同群体在平等、合作的条件下接触能减少偏见。
要求： 生成一段20轮的详细对话。

基于社会学理论的合成数据干预

降低幻觉输出

时序知识图谱的构建

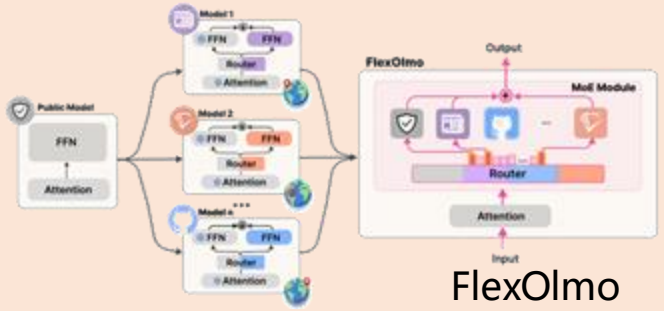




基于历史一致性的幻觉惩罚

塑造群体多样性

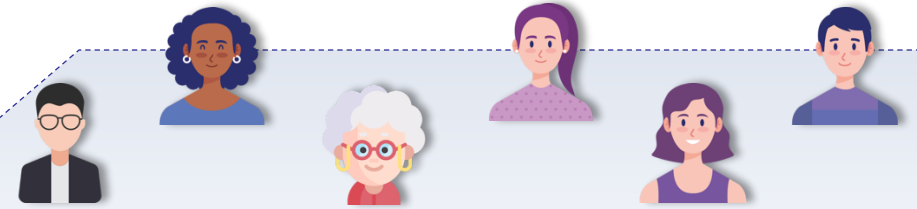
- 混合专家架构
- 角色特定模块
- 基于角色的动态路由
- 推理时动态组合
- 复杂角色构建



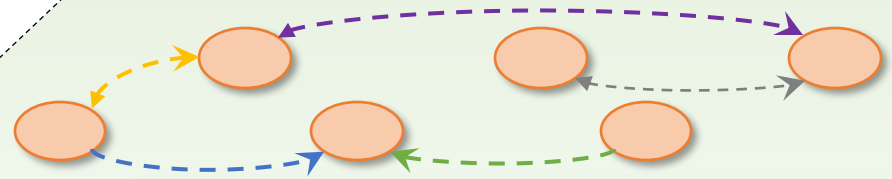
FlexOlmo

基于混合专家的角色配置

总结



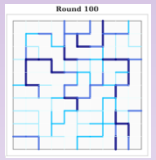
社会智能体设计：设计具备认知现实主义的智能体



模拟情景设计：低代码复杂模拟场景的自动化构建

底层架构设计：支持高并发、高吞吐，可靠的模拟

模拟评估



与经典理论及物理现实的对齐验证

AI社科研究员

宏观模式检测器	微观归因引擎
因果链条挖掘	……

自动化数据分析、因果提取、报告生成

引用

- Generative Agent Simulations of 1,000 People
- A Survey on Large Language Model-Based Social Agents in Game-Theoretic Scenarios Language Models Represent Beliefs of Self and Others
- SocioVerse: A World Model for Social Simulation Powered by LLM Agents and A Pool of 10 Million Real-World Users
- EconAgent: Large Language Model-Empowered Agents for Simulating Macroeconomic Activities
- AgentSociety: Large-Scale Simulation of LLM-Driven Generative Agents Advances Understanding of Human Behaviors and Society
- S^3: Social-network Simulation System with Large Language Model-Empowered Agents
- YuLan-OneSim: Towards the Next Generation of Social Simulator with Large Language Models
- OASIS: Open Agent Social Interaction Simulations with One Million Agents
- What Limits LLM-based Human Simulation: LLMs or Our Design?
- LLM Social Simulations Are a Promising Research Method
- War and Peace (WarAgent): Large Language Model-based Multi-Agent Simulation of World Wars
- Agent Hospital: A Simulacrum of Hospital with Evolvable Medical Agents
- ElectionSim: Massive Population Election Simulation Powered by Large Language Model Driven Agents
- EconAgent: Large Language Model-Empowered Agents for Simulating Macroeconomic Activities
- LLM-Augmented Agent-Based Modelling for Social Simulations: Challenges and Opportunities
- From Individual to Society: A Survey on Social Simulation Driven by Large Language Model-based Agents

人员

冯夏冲
香港大学 博士后

孔令鹏
香港大学 助理教授

谢谢！

冯夏冲 香港大学
fengxc@hku.hk