Simulating Human Society with Large Language Model Agents: City, Social Media, and Economic System

Chen Gao[1], Fengli Xu[1], Xu Chen[2], Xiang Wang[3], Xiangnan He[3] and Yong Li[1]

[1]Tsinghua University

[2]Renmin University of China

[3]University of Science and Technology of China

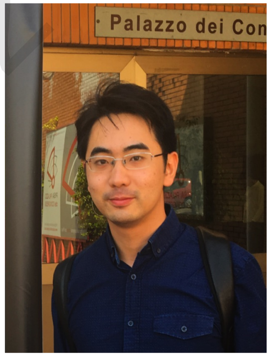
About us

Chen Gao
Research Assistant Professor
Tsinghua University

Fengli Xu
Assistant Professor
Tsinghua University

Xu Chen
Associate Professor
Renmin University of China

Xiang Wang
Professor
University of Science and Technology of China



Xiangnan He
Professor
University of Science and Technology of China

Yong Li
Tenured Associate Professor
Tsinghua University

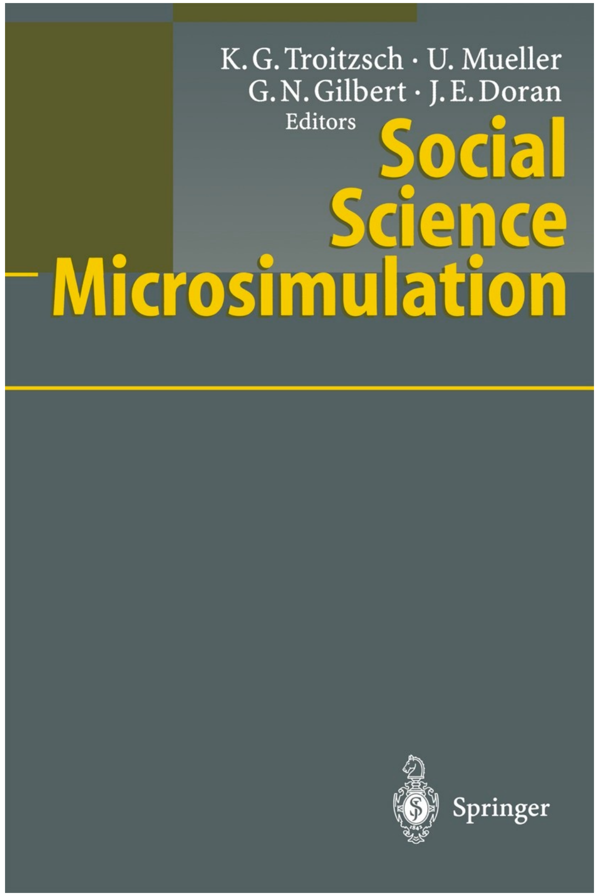
Outline

- Background of LLM Agent-based Simulation (25minutes) 13:30-13:55
- Online behavior simulation with LLM Agents (65 minutes) 13:55-15:00
- Break (15 minutes) 15:00-15:15
- Social and economic simulation with LLM agents (50 minutes) 15:15-16:05
- City system simulation with LLM agents (45minutes) 16:05-16:50
- Open discussions (10minutes) 16:50-17:00

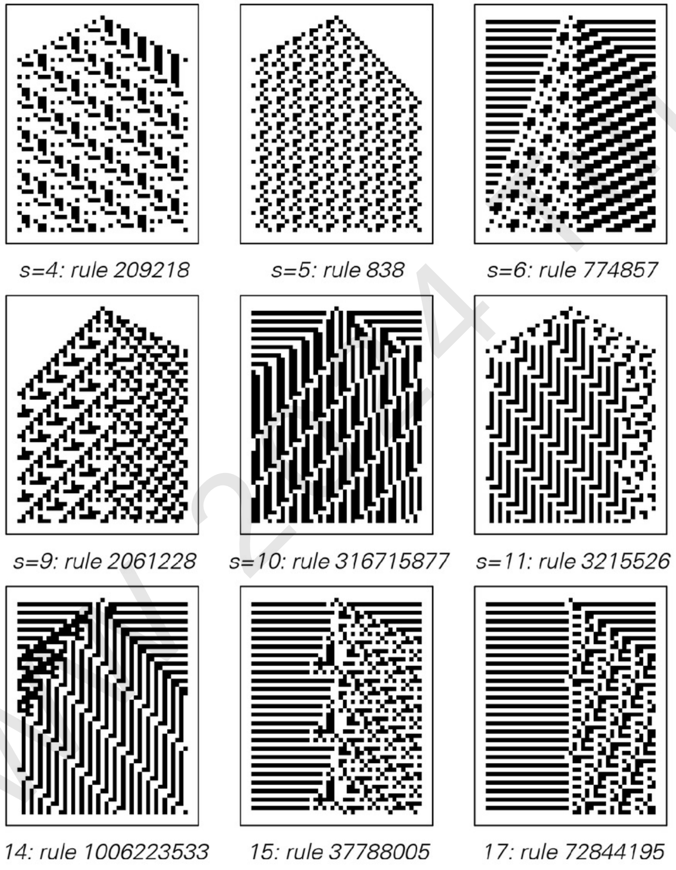
Outline

- Background of LLM Agent-based Simulation
- Online behavior simulation with LLM Agents
- Social and economic simulation with LLM agents
- City system simulation with LLM agents
- Open discussions

Background: simulation in social science



Troitzsch, Klaus G., et al., eds. *Social science microsimulation*. Springer Science & Business Media, 1996.



Cellular Automata


Generative Agents: Interactive Simulacra of Human Behavior

Joon Sung Park, Stanford University, Stanford, USA, joonspk@stanford.edu
Joseph C. O'Brien, Stanford University, Stanford, USA, jobrien3@stanford.edu
Carrie J. Cai, Google Research, Mountain View, CA, USA, cjcai@google.com
Meredith Ringel Morris, Google Research, Seattle, WA, USA, merrie@google.com
Percy Liang, Stanford University, Stanford, USA, pliang@cs.stanford.edu
Michael S. Bernstein, Stanford University, Stanford, USA, msb@cs.stanford.edu

Figure 1: Generative agents create believable simulacra of human behavior for interactive applications. In this work, we demonstrate generative agents by populating a sandbox environment, reminiscent of The Sims, with twenty-five agents. Users can observe and intervene as agents they plan their days, share news, form relationships, and coordinate group activities.

AI-based Simulation

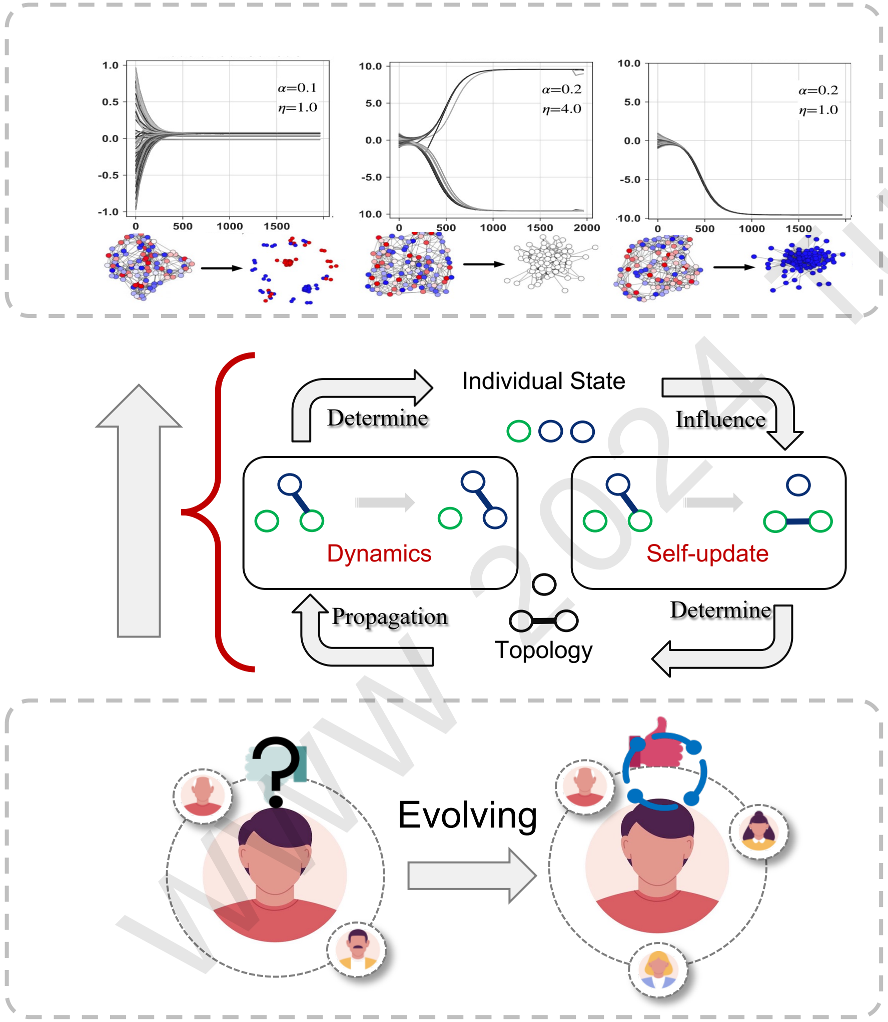
Simulation is fundamental tool of Social Science

Background: simulation in social science

- Agent-based modeling
  - Agents
  - Environment
  - Interaction
- Well-known concepts in multiple areas

Background: simulation in social science

Population-level Phenomena



System-based Macrosimulation

Individual-level State Evolving

Agent-based Microsimulation

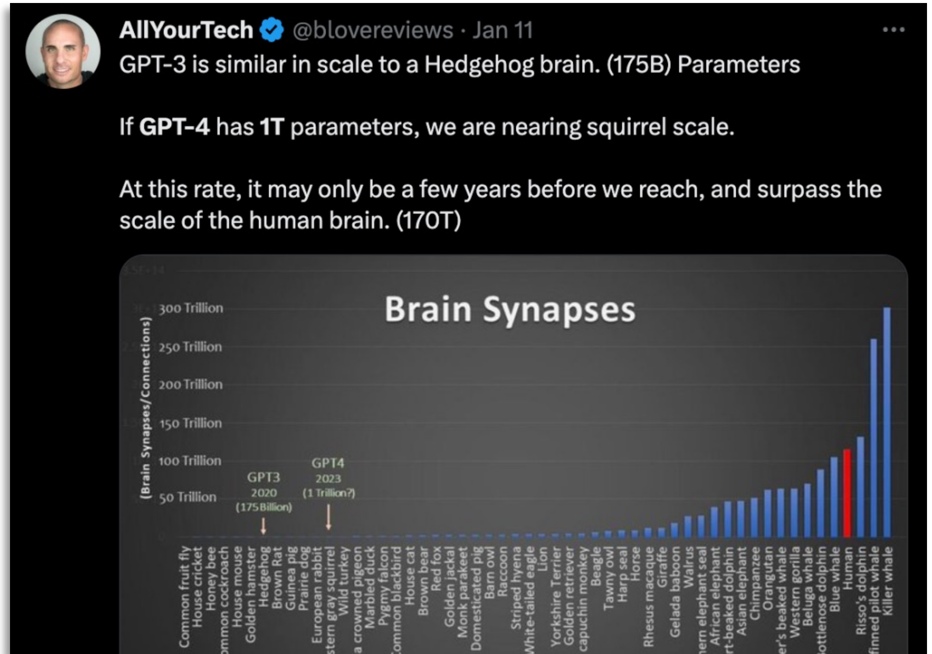
Agent-based simulation is ubiquitous but requires accurate modeling and description for the agent

Why can LLM support simulation

GPT-3 has 175B parameters, close to hedgehog's brain

GPT-4 has 1T parameters, close to squirrel's brain

Human brain has 170TB parameters, GPT-N?



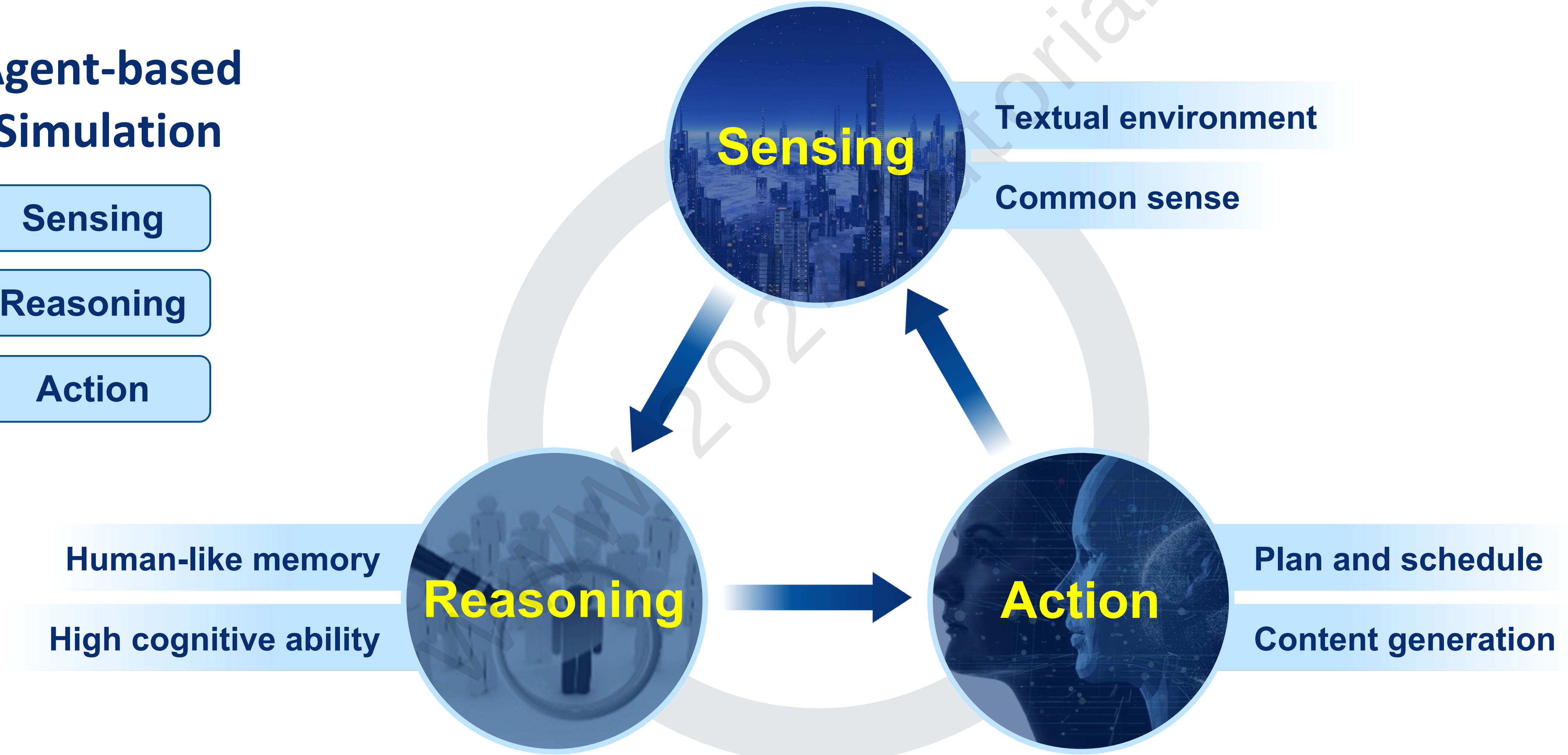
"At this rate, it may only be a few years before we reach, and surpass the scale of the human brain. (170T)"

How about use LLM as agent?

Why can LLM support simulation

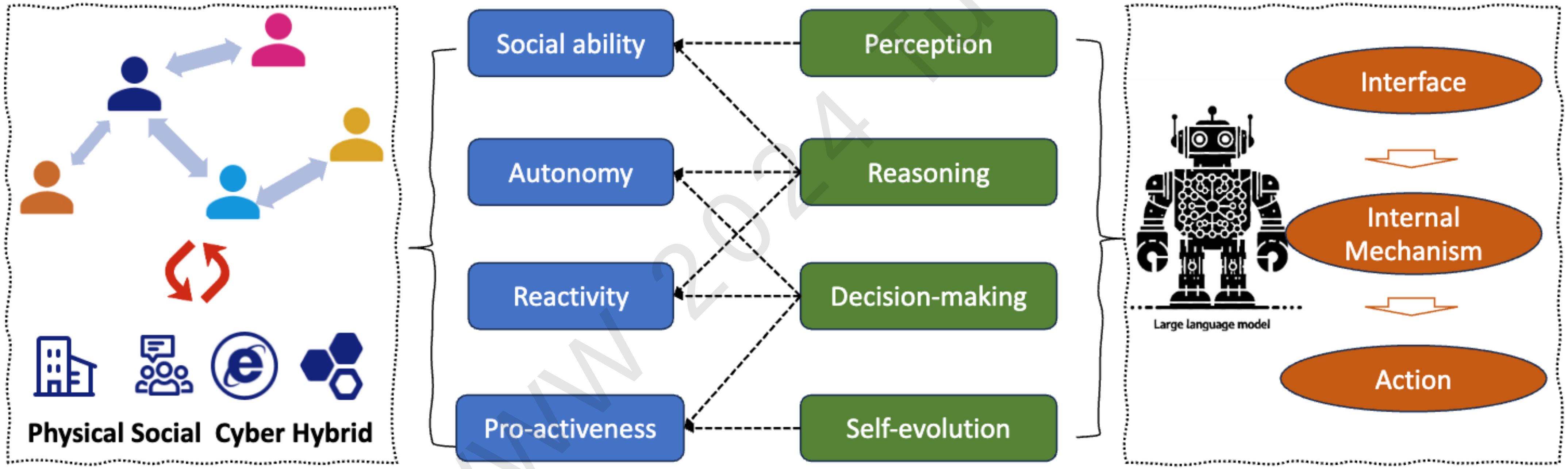
Agent-based Simulation

- Sensing
- Reasoning
- Action



Large language models (LLMs) well fits the paradigm of agent-based simulation

Why can LLM support simulation



Requirements for Agent-based Modeling and Simulation

Large Language Model Agents' Abilities

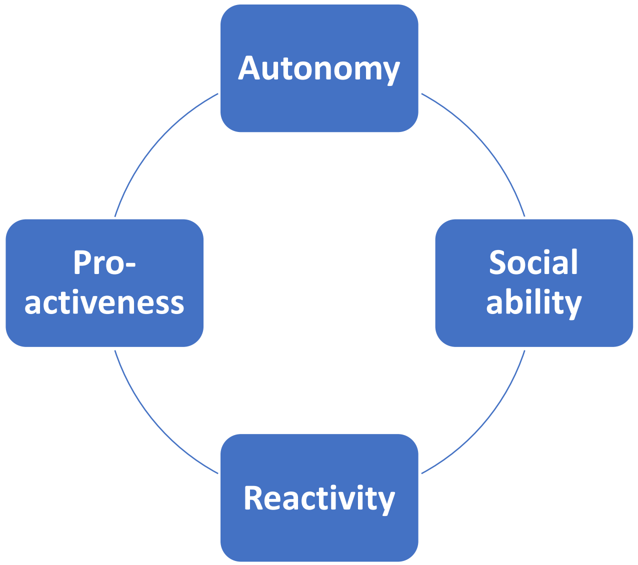
Why can LLM support simulation

- **Autonomy**. Agents should be able to operate without the direct intervention of humans or others, which is important in real-world applications such as microscopic traffic flow simulation and pedestrian movement simulation.

- **Social ability**. Agents should be able to interact with other agents (and possibly humans) to complete the assigned goals.

- **Reactivity**. Agents should be able to perceive their environment and respond quickly to changes in the environment.

- **Pro-activeness**. Agents should be able to exhibit goal-directed behavior by taking the initiative instead of just responding to their environment.



Why can LLM support simulation：Existing methodologies

- **Predefined rules.** Define explicit rules that govern agent behaviors. These rules are typically based on logical or conditional statements that dictate how agents react to specific situations or inputs.

- **Symbolic equations.** Algebraic equations, differential equations, or other mathematical formulations.

- **Stochastic modeling**. Introduces randomness and probability into agent decision-making, which is useful for capturing the uncertainty and variability inherent in many real-world systems.

- **Machine learning models.** Allow agents to learn from data or through interaction with their environment.

1
- **Predefined Rule**

2
- **Symbolic equations**

3
- **Stochastic modeling**

4
- **Machine learning models**

Why can LLM support simulation：Existing methodologies

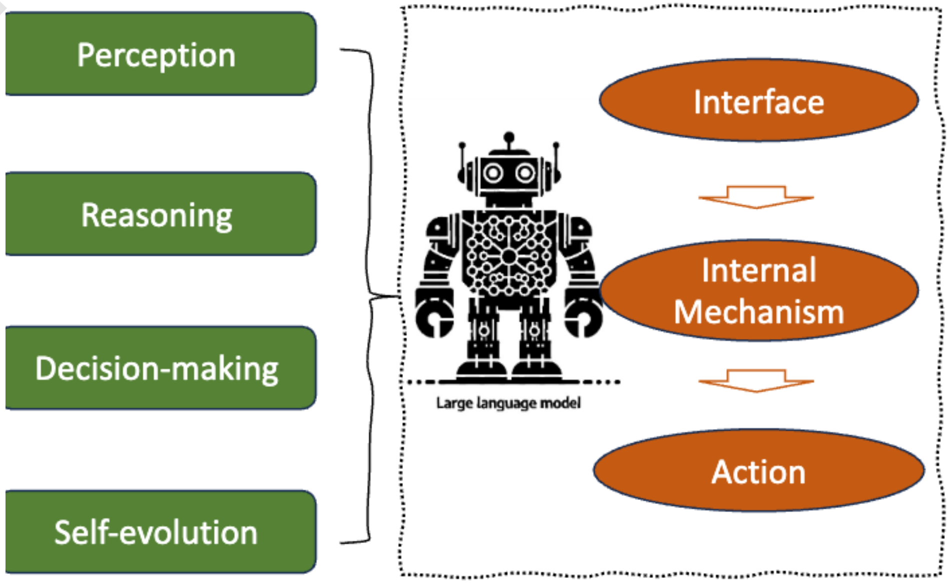
- **Simple agent architecture is not enough to cope with complex tasks.**

- **It is difficult to develop a general agent that can support simulations across different environments.**

- **Existing methods cannot support integrative simulation in real-world problems.**

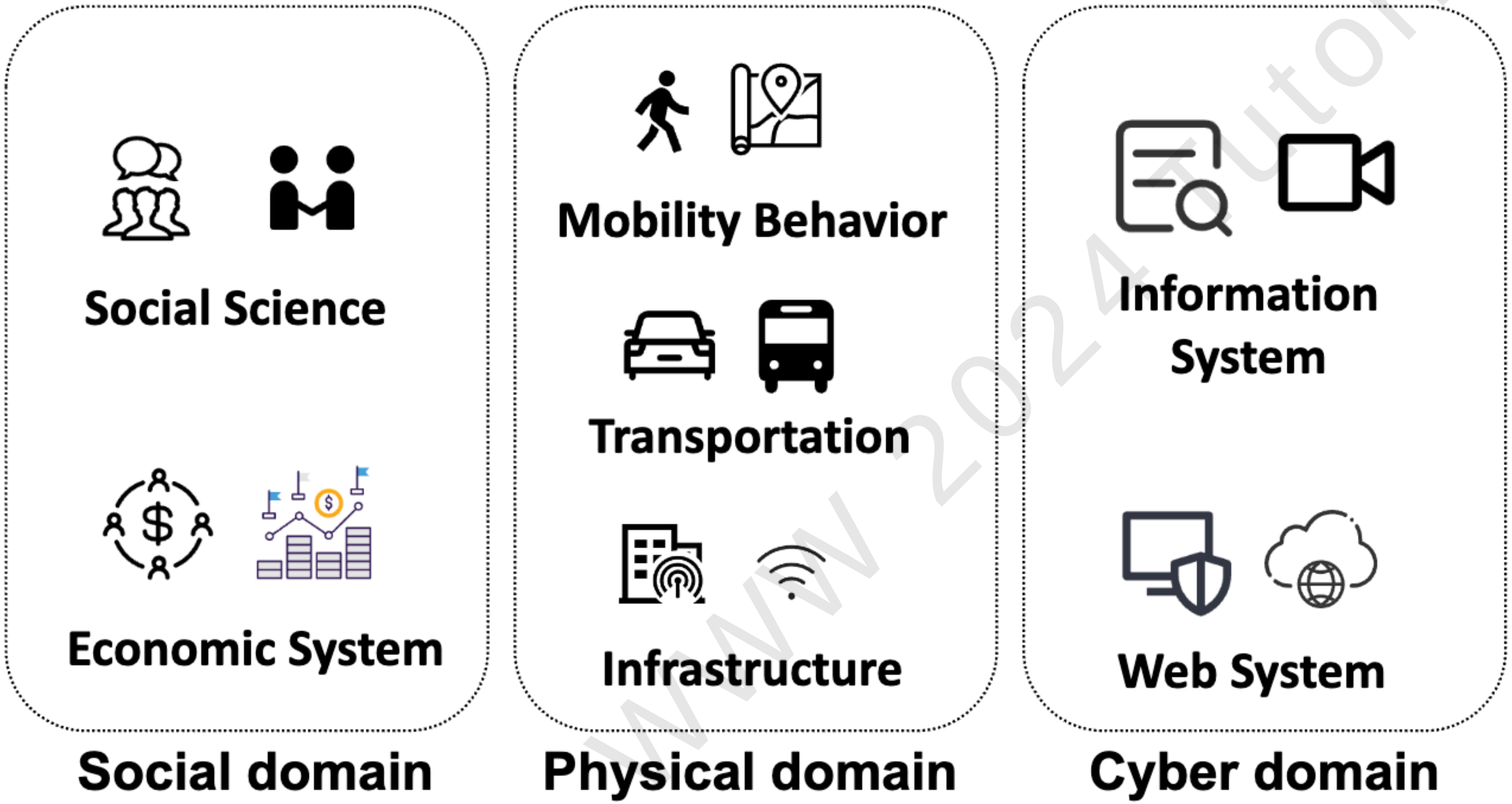
Why can LLM support simulation：LLM agents' abilities

- **Perception**
  - Be able to comprehend, perceive, and respond to diverse needs, emotions, and attitudes within different contexts, from the "first-view sight.
- **Reasoning and Decision-making**
  - With only limited guidance, regulations, and goals, agents equipped with large language models can autonomously take actions, make plans for the given goal, or even achieve new goals without the need for explicit programming or predefined rules.
- **Adaptive learning and evolution**
  - LLM agents can assimilate new information, analyze emerging patterns in data, and modify their responses or actions accordingly



Large Language Model Agents' Abilities

Why can LLM support simulation



Domain	Environment	Advance	What to simulate
Social	Virtual	Schwitzgebel *et al.* [173]	Conversation and interaction
Social	Virtual	Xu et al. [222]	Werewolf game
Social	Virtual	Acerbi et al. [1]	Information Propagation
Social	Virtual	Zhang et al. [231]	Collaboration Mechanism
Social	Virtual	Suzuki et al. [191]	Cooperation and defection
Social	Virtual	Zarza et al. [54]	Social interaction
Social	Real	Mukobi et al. [146]	Welfare diplomcy game
Social	Real	S3 [73]	Online social network
Social	Virtual	SimReddit [153]	Online forum
Social	Real	COLA [113]	Cooperative task solving
Social	Virtual	MAD [124]	Cooperative task solving
Social	Virtual	CHATDEV [163]	Cooperative task solving
Social	Virtual	MetaGPT [95]	Cooperative task solving
Social	Virtual	ChatEval [36]	Cooperative task solving
Social	Virtual	CAMEL [119]	Cooperative task solving
Social	Virtual	AgentVerse [43]	Cooperative task solving
Social	Virtual	SPP [210]	Cooperative task solving
Social	Virtual	CoELA [230]	Cooperative task solving
Social	Virtual	Humanoid Agents [211]	Individual social behavior
Social	Real	SocioDojo [14]	Individual social behavior
Social	Virtual	Liu *et al.* [128]	Individual social behavior
Social	Virtual	Argyle *et al.* [10]	Individual social behavior
Social	Virtual	Hamalainen *et al.* [87]	Individual social behavior
Social	Virtual	Singh *et al.* [186]	Individual social behavior
Social	Virtual	Binz *et al.* [25]	Individual social behavior
Social	Virtual	Elyoseph *et al.* [66]	Individual social behavior
Social	Virtual	Li *et al.* [123]	Individual social behavior
Social	Virtual	Horton [96]	Economic system: individual behavior
Social	Virtual	Chen et al. [44]	Economic system: individual behavior
Social	Virtual	Geerling et al. [76]	Economic system: individual behavior
Social	Real	Xie et al. [220]	Economic system: market behavior
Social	Real	Faria et al. [68]	Economic system: market behavior
Social	Real	Bybee et al. [32]	Economic system: market behavior
Social	Virtual	Phelps et al. [159]	Economic system: game theory
Social	Virtual	Akata et al. [3]	Economic system: game theory
Social	Virtual	Guo et al. [82]	Economic system: game theory
Social	Virtual	Zhao et al. [234]	Economic system: consumption market
Social	Virtual	Han et al. [89]	Economic system: consumption market
Social	Virtual	Zhao et al. [147]	Economic system: consumption market
Social	Virtual	Chen et al. [41]	Economic system: auction market
Physical	Real	Shah et al. [175]	Navigation behavior
Physical	Real	NLMap [40]	Navigation behavior
Physical	Real	Zou et al. [245]	Wireless network users
Physical	Real	Cui et al. [51]	Vehicle drivers
Physical	Virtual	GITM [241]	Tool-usage simulation in sandbox game
Cyber	Real	WebAgent [83]	Human behaviors in Web
Cyber	Real	Mind2Web [57]	Human behaviors in Web
Cyber	Real	Zhou *et al.* [237]	Human behaviors in Web
Cyber	Real	Park *et al.* [151]	Human behaviors in Web
Cyber	Virtual	RecAgent [209]	Interaction with recommender system
Cyber	Virtual	Agent4Rec [228]	Interaction with recommender system
Hybrid	Virtual	Williams et al. [215]	Epidemic spreading
Hybrid	Virtual	Generative Agents [152]	Sandbox social life
Hybrid	Real	WarAgent [100]	War simulation
Hybrid	Real	Li et al. [121]	Economic system: macroeconomics
Hybrid	Real	UGI [221]	Human behaviors in real-world city

Why can LLM support simulation

Domain	Environment	Advance	What to simulate
Social	Virtual	Schwitzgebel *et al.* [173]	Conversation and interaction
Social	Virtual	Xu et al. [222]	Werewolf game
Social	Virtual	Acerbi et al. [1]	Information Propagation
Social	Virtual	Zhang et al. [231]	Collaboration Mechanism
Social	Virtual	Suzuki et al. [191]	Cooperation and defection
Social	Virtual	Zarza et al. [54]	Social interaction
Social	Real	Mukobi et al. [146]	Welfare diplomcy game
Social	Real	S3 [73]	Online social network
Social	Virtual	SimReddit [153]	Online forum
Social	Real	COLA [113]	Cooperative task solving
Social	Virtual	MAD [124]	Cooperative task solving
Social	Virtual	CHATDEV [163]	Cooperative task solving
Social	Virtual	MetaGPT [95]	Cooperative task solving
Social	Virtual	ChatEval [36]	Cooperative task solving
Social	Virtual	CAMEL [119]	Cooperative task solving
Social	Virtual	AgentVerse [43]	Cooperative task solving
Social	Virtual	SPP [210]	Cooperative task solving
Social	Virtual	CoELA [230]	Cooperative task solving
Social	Virtual	Humanoid Agents [211]	Individual social behavior
Social	Real	SocioDojo [14]	Individual social behavior

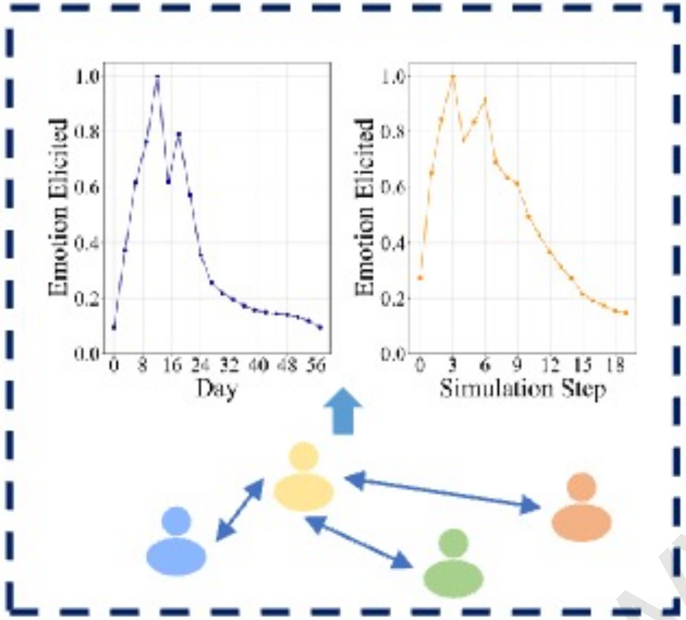
Why can LLM support simulation

Physical	Real	Shah et al. [175]	Navigation behavior
Physical	Real	NLMap [40]	Navigation behavior
Physical	Real	Zou et al. [245]	Wireless network users
Physical	Real	Cui et al. [51]	Vehicle drivers
Physical	Virtual	GITM [241]	Tool-usage simulation in sandbox game
Cyber	Real	WebAgent [83]	Human behaviors in Web
Cyber	Real	Mind2Web [57]	Human behaviors in Web
Cyber	Real	Zhou *et al.* [237]	Human behaviors in Web
Cyber	Real	Park *et al.* [151]	Human behaviors in Web
Cyber	Virtual	RecAgent [209]	Interaction with recommender system
Cyber	Virtual	Agent4Rec [228]	Interaction with recommender system
Hybrid	Virtual	Williams et al. [215]	Epidemic spreading
Hybrid	Virtual	Generative Agents [152]	Sandbox social life
Hybrid	Real	WarAgent [100]	War simulation
Hybrid	Real	Li et al. [121]	Economic system: macroeconomics
Hybrid	Real	UGI [221]	Human behaviors in real-world city

LLM simulation: take social simulation as an example

Social network dynamics

Trends, Message propagation

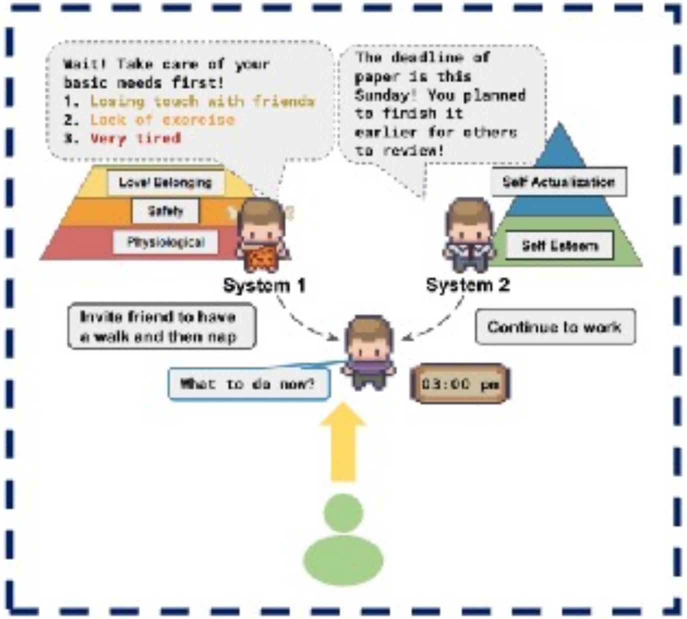


Cooperation

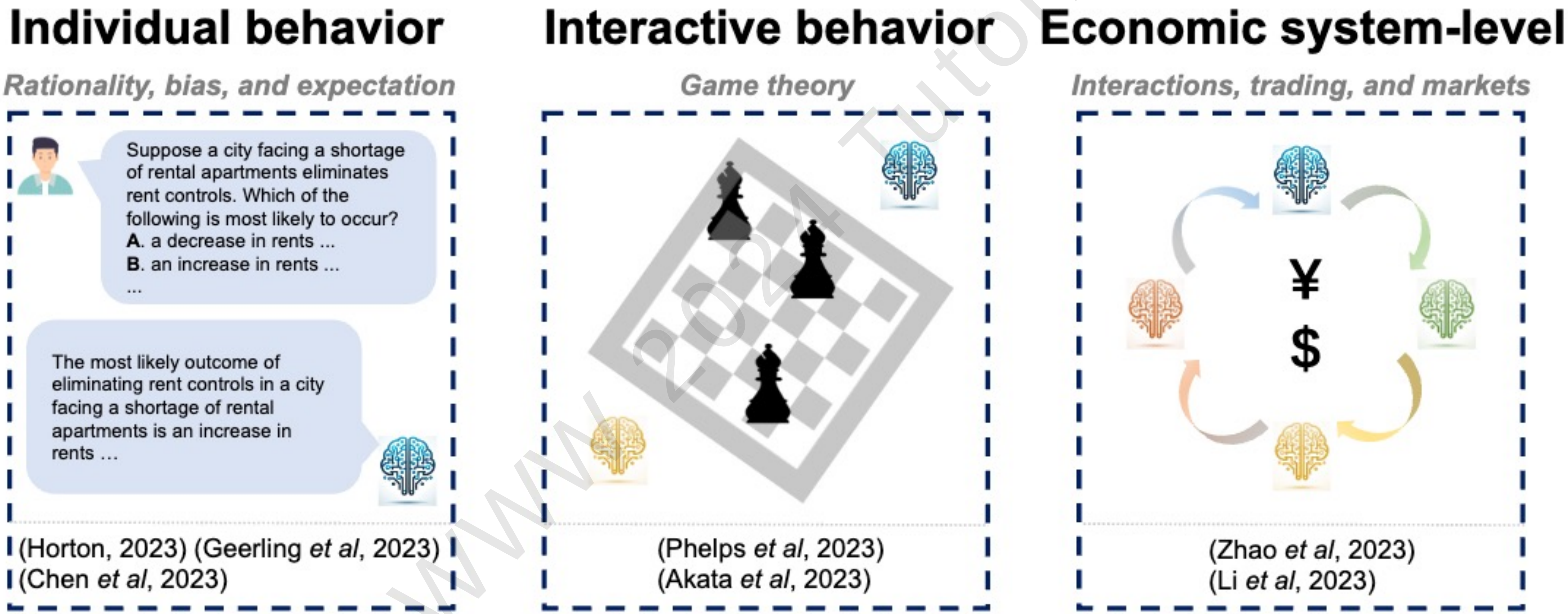
Cooperative task solving

Individual

Individual social behavior

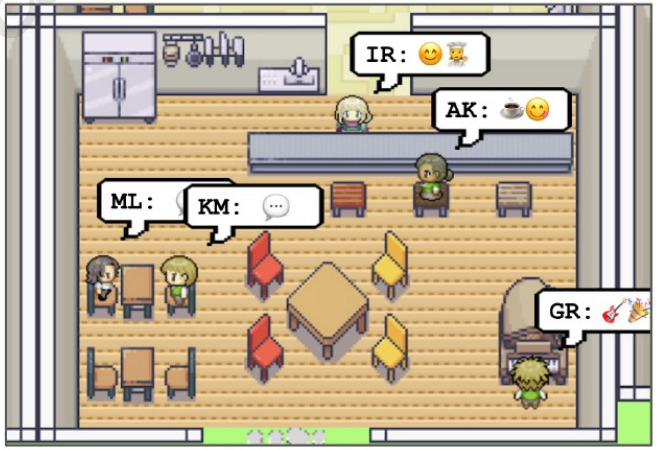


LLM simulation: take economic simulation as an example



LLM for simulation

- **Environment construction and interface**
  - Environment: define the world and rules
  - Interface
- **Human alignment and personalization**
  - Human alignment
  - Personalization
- **How to simulate actions**
  - Planning
  - Memory
  - Reflection
- **Evaluation**
  - Realness validation with real human data
  - Provide explanations for simulated behaviors
  - Ethics evaluation



About today's following presentations

- Online behavior simulation with LLM Agents **@Xu Chen @ Xiang Wang**
- Social and economic simulation with LLM agents **@Chen Gao**
- City system simulation with LLM agents **@Fengli Xu**
- Open discussions **@Fengli Xu**

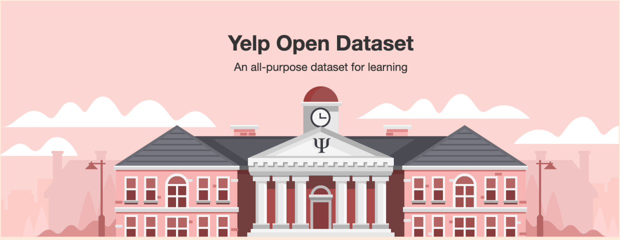
Outline

- Background of LLM Agent-based Simulation (25minutes) 13:30-13:55
- Online behavior simulation with LLM Agents (65 minutes) 13:55-15:00
- Break (15 minutes) 15:00-15:15
- Social and economic simulation with LLM agents (50 minutes) 15:15-16:05
- City system simulation with LLM agents (45minutes) 16:05-16:50
- Open discussions (10minutes) 16:50-17:00

Different Study Paradigms in AI

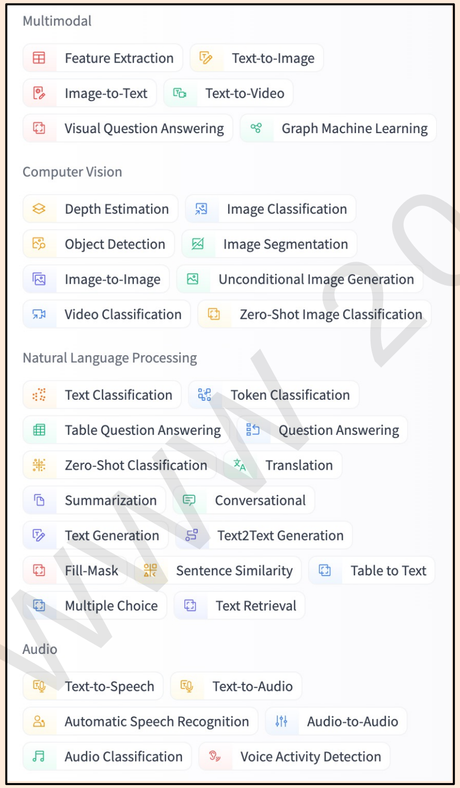
高瓴人工智能学院
Gaoling School of Artificial Intelligence

Real-world Datasets



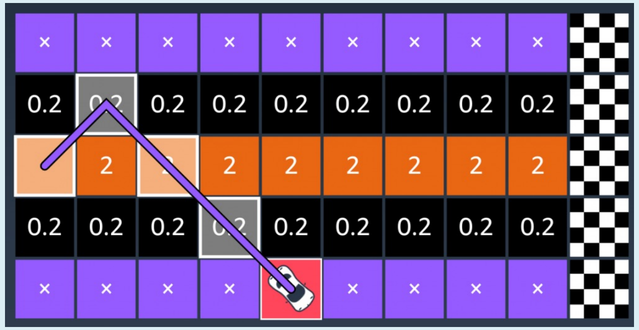
Hugging Face

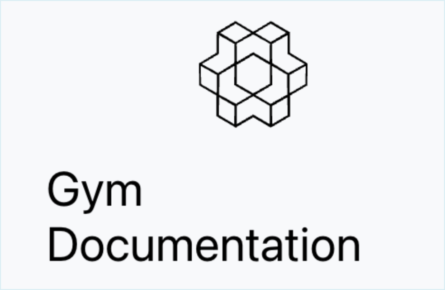
Datasets 75,262

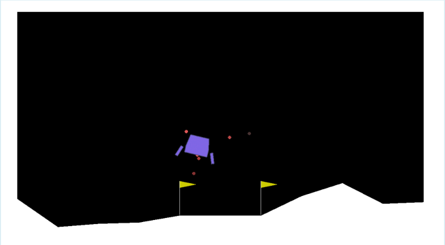


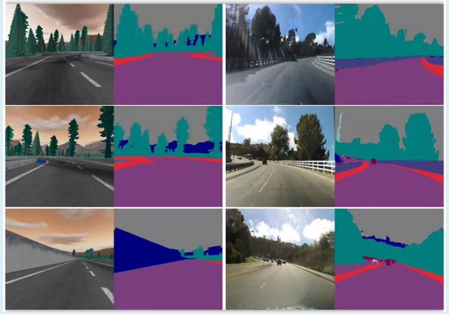
- **A large amount of public available datasets**
- **The datasets can be easily collected**

Simulation Environments





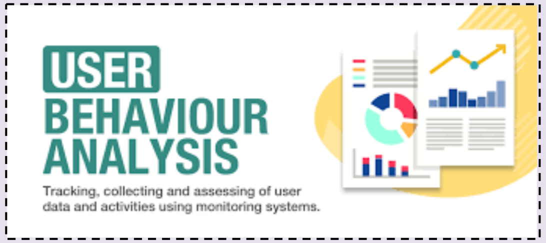




The data generation mechanisms are known or can be accurately predicted

Where is the Position of User Behavior Analysis

User Behavior Analysis



Recommendation System

Social Network

User Behavior Tracking

…

- User Privacy
- Commercial Confidentiality
- Ethical Problem

- Intricate Human Cognitive Process
- Complex Environments
- Complex Influential Factors

Real-world datasets

- ➢ **A large amount of public available datasets**
- ➢ **The datasets can be easily collected**

Simulation Environments

The data generation mechanisms are known or can be accurately predicted

Simulation based User Behavior Analysis

International Conferences

International Workshops

./ User Simulation

Evaluating interactive intelligent systems

Book draft | CIKM'23 tutorial

Tutorial on User Simulation for Evaluating Information Access Systems at the 31st ACM International Conference on Information and Knowledge Management (CIKM '23)

Books

arXiv:2306.08550v1 [cs.HC] 14 Jun 2023

User Simulation for Evaluating Information Access Systems

PREPRINT MANUSCRIPT (version 1.0, 2023-06-14)
This is an unreviewed preprint of a monograph under review for Foundations and Trends in Information Retrieval. Feedback, suggestions, and comments from the community are greatly appreciated and are invited to be shared with the authors via email.

Krisztian Balog
University of Stavanger
krisztian.balog@uis.no

ChengXiang Zhai
University of Illinois at Urbana-Champaign
czhai@illinois.edu

Simulation based User Behavior Analysis

Simplified Environments

Different Environments May Mutually Influence Each Other

Watching Movies

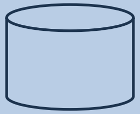
Opinion Sharing

Simplified User Decision Processes

- Clicking Behaviors
- Browsing Behaviors
- Purchasing Behaviors
- Watching Behaviors
- …

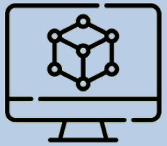
- Inner Product
- Multilayer Perceptron
- Deep Neural Network
- …

Rely on Real-world Datasets



Real-world Datasets

Training



Simulator

User Behaviors

Zero-Shot Simulation is Important

Simple environments

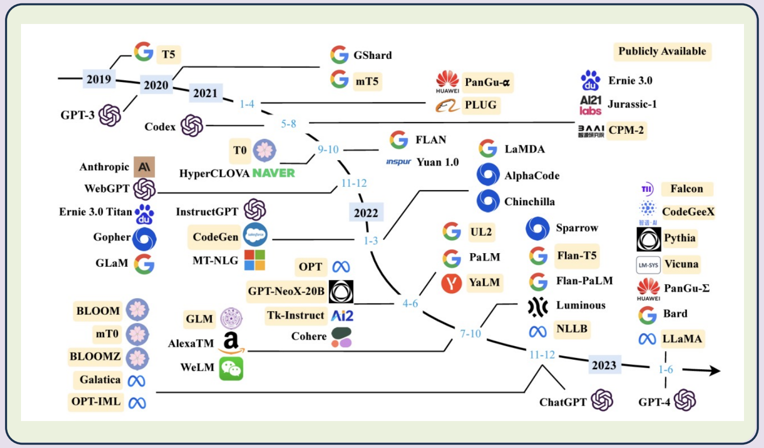
Simple functions

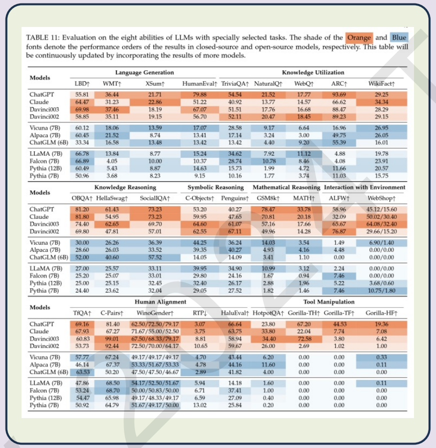
Sparse datasets

Unreliable Simulation

A Novel Paradigm for User Behavior Simulation

The Fast Growing of Large Language Models





- ✓ Human-level Intelligence
- ✓ Surprisingly Strong Generalization Capability

Human-level Intelligence

Zero-shot Inference
Unified NLP interface

Human-like decisions

Generalization Capability

Zero-shot Inference

Simplified Environments | **Simplified User Decision Processes** | **Rely on Real-world Datasets**

A Novel Paradigm for User Behavior Simulation

Potential Challenges

[Local] How to make LLMs act like real users ?

- How to profile the users ?
- How to make LLMs dynamically evolve in the environments ?
- What user behaviors should be simulated ?

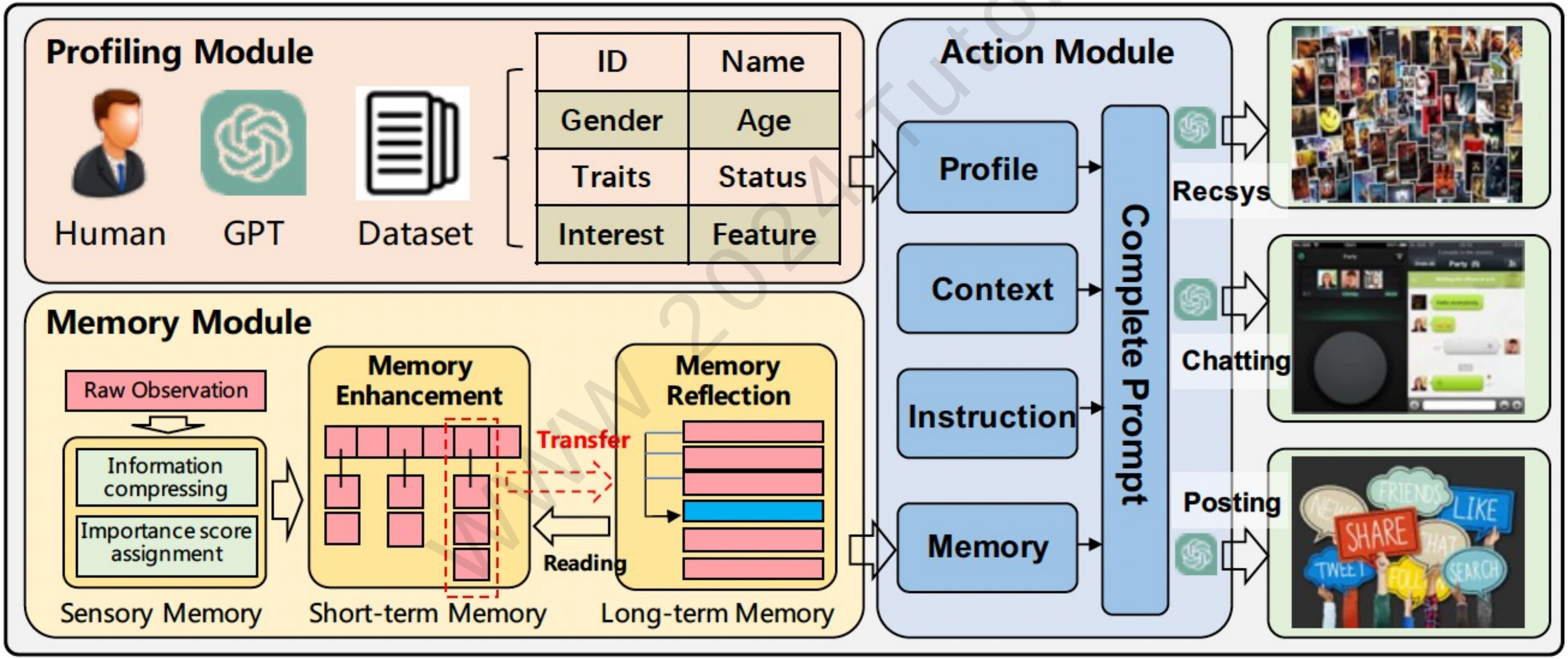
[Global] How to build a system to simulate user behaviors?

- How to organize different users in the system ?
- How to design the simulation process ?
- What auxiliary functions can be designed and how to realize?

User Behavior Simulation with on LLM-based Agents

Agent-level Design

Agent = LLM + Profiling Module + Memory Module + Action Module



Agent-level Design

Profiling Module

ID	Name	Gender	Age	Traits	Career	Interest	Feature
0	David Smith	male	25	compassionate, caring, ambiti ous, optimistic	photographer	sci-fi movies, comedy movies	Watcher;Critic;Poster
1	David Miller	female	39	Funloving, creative, practical, energetic, patient	writer	action movies, scifi movies, classic movies	Watcher;Explorer;Poster
2	James Brown	male	70	independent, creative, patient , empathetic	engineer	comedy movies, familyfriendly movi es, documentaries, thriller movies	Watcher;Critic;Poster
3	Sarah Miller	female	33	independent, compassionate	farmer	romantic movies, comedy movies, c lassic movies, family-friendly movies	Watcher;Critic;Poster
4	John Taylor	male	68	optimistic	doctor	action movies, thriller movies	Watcher;Poster
5	Sarah Williams	female	51	meticulous	musician	action movies, documentaries, scifi movies, familyfriendly movies	Watcher;Explorer;Chatter
6	James Jones	male	59	practical, funloving, creative, ambitious, caring	farmer	documentaries	Watcher;Poster
7	Jane Brown	female	30	patient, adventurous, fun-loving, optimistic	doctor	documentaries	Watcher;Explorer;Poster
8	David Jones	male	23	analytical, energetic, introspe ctive, independent	scientist	familyfriendly movies, thriller movie s, action movies, sci-fi movies	Poster
9	James Brown	female	20	ambitious, analytical, optimist ic, energetic, meticulous	designer	familyfriendly movies, romantic mov ies	Critic; Chatter
10	James Garcia	male	20	practical, energetic, introspect ive, patient	engineer	documentaries, thriller movies, com edy movies, classic movies, romanti c movie	Watcher; Explorer; Poster

Agent-level Design

Profiling Module

Handcrafting Method

- ✅ More flexible
- ❌ Labor intensive
- ❌ Hard to scale up

GPT-generation Method

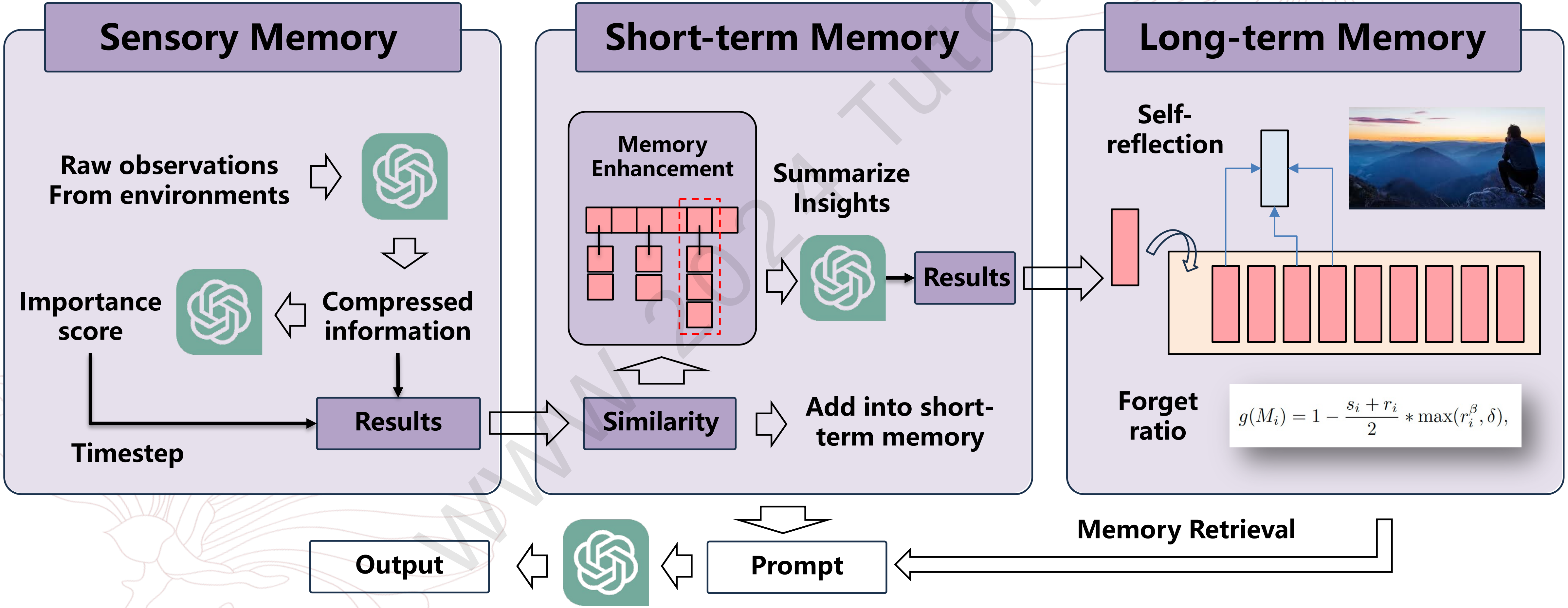
- ❌ Less flexible
- ✅ Lower expenses
- ✅ Easy to scale up

Dataset Alignment Method

- ❌ Less flexible
- ✅ Lower expenses
- ✅ More real

Agent-level Design

Memory Module

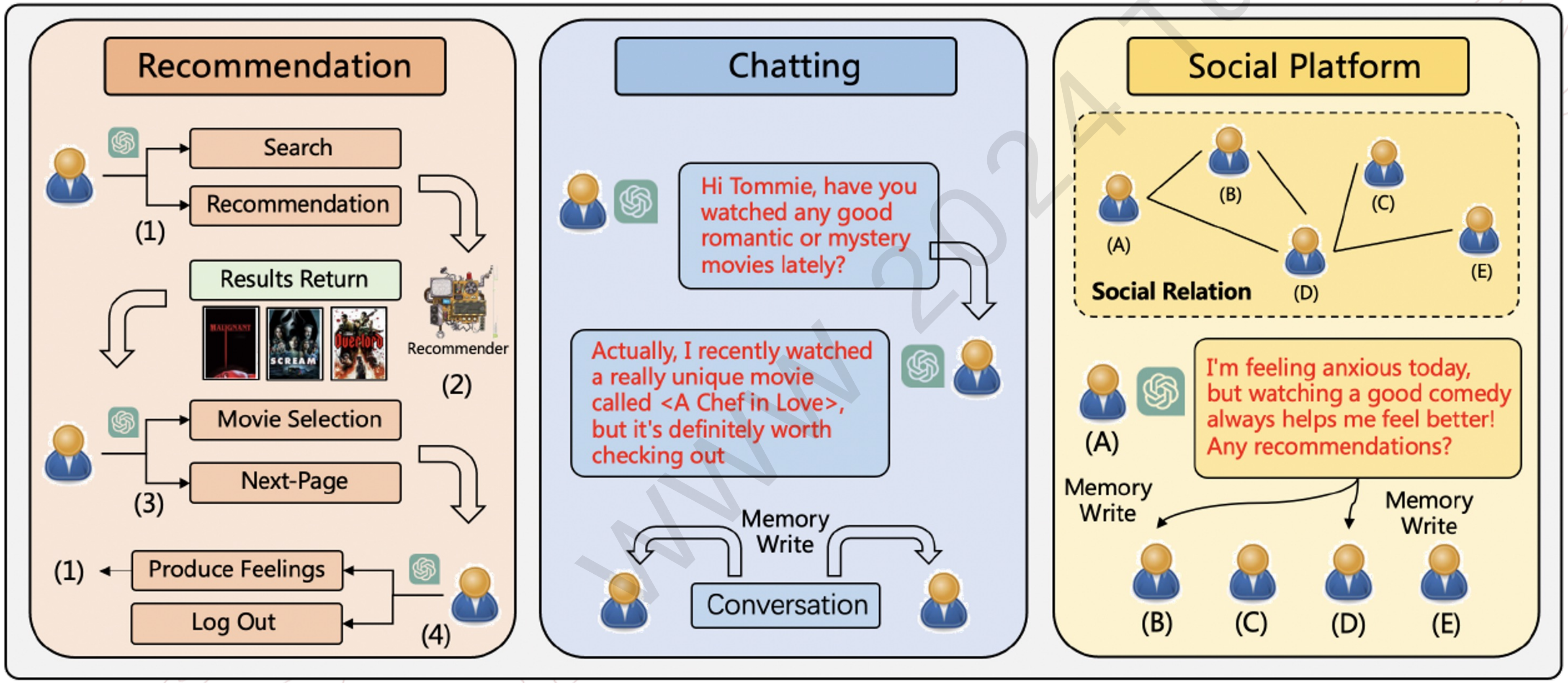


Richard C Atkinson and Richard M Shiffrin. **Human memory: A proposed system and its control processes**. In Psychology of learning and motivation, volume 2, pages 89–195. Elsevier, 1968.

Agent-level Design

Action Module

Simulate more complete recommendation ecosystem



Inside Recsys

- Searching behavior
- Browsing behavior
- Watching behavior
- Next-page behavior

Outside Recsys

- One to one chatting
- One to many social broadcasting

Agent-level Design

Behavior Adaptive Prompt Generation

Name: David Smith (age: 25), David Smith, a 25-year-old male photographer, is compassionate, caring, ambitious, and optimistic. He enjoys watching sci-fi and comedy movies and provides feedback and ratings to the recommendation system. He demands high standards for movies and the recommendation system and may criticize both. The observation about David watching "The Neon Bible" aligns with his interest in drama films and explores themes of faith, family, and coming-of-age. **Profile**

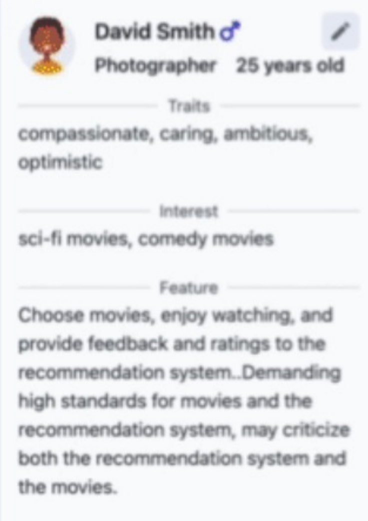
It is August 18, 2023, 12:00 AM. **Context**

Most recent observations: David Smith enjoys and finds captivating films that have captivating plots, humorous elements, thought-provoking themes, delve into complexities of human nature and sexual desire, uplift viewers, and have vibrant and engaging performances by the cast.

Observation: David Smith has just finished watching Neon Bible, The (1995):"The Neon Bible" is a drama film set in the 1940s in a small southern town in the United States. It follows the story of a young boy named David who is struggling to understand the complexities of the world around him. David\'s mother is mentally unstable and his father is absent, leaving him to navigate the challenges of adolescence on his own. As he tries to make sense of his surroundings, he turns to religion and finds solace in the teachings of his local preacher. However, his faith is tested when he discovers the secrets and hypocrisies of those around him. The film explores themes of faith, family, and coming-of-age in a poignant and powerful way. **Memory**

All occurrences of movie names should be enclosed with <>. David Smith has not seen this movie before. Imagine you are David Smith, how will you feel about this movie just watched? Please share your personal feelings about the movie in one line. Please act as David Smith well. **Instruction**

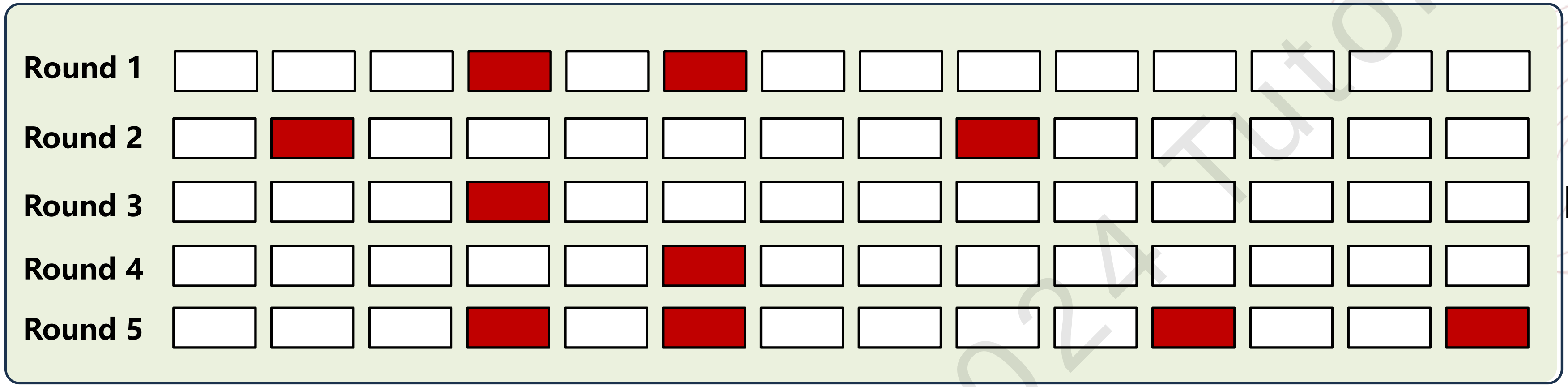
- **Simplified profile according to the current behavior**



- **Adaptive Memory based on the current behavior**

System-level Design

Execution Protocol



Pareto distribution

$$p(x) = \frac{\alpha x_{min}^{\alpha}}{x^{\alpha+1}},$$

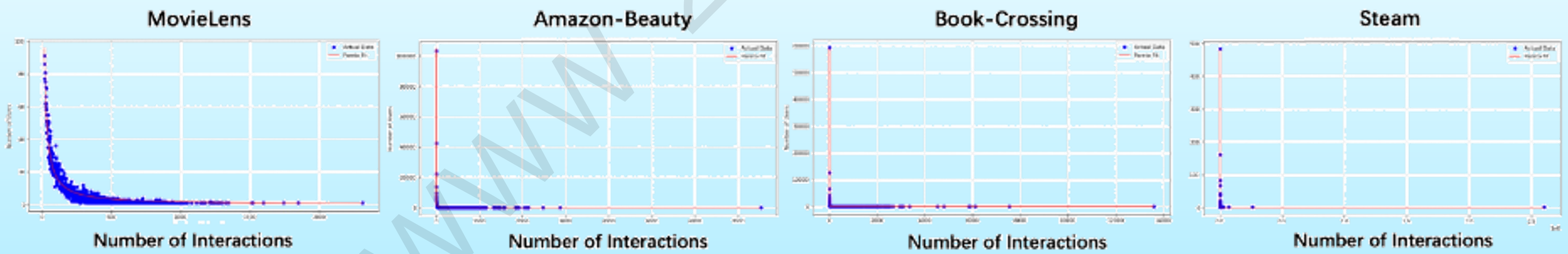
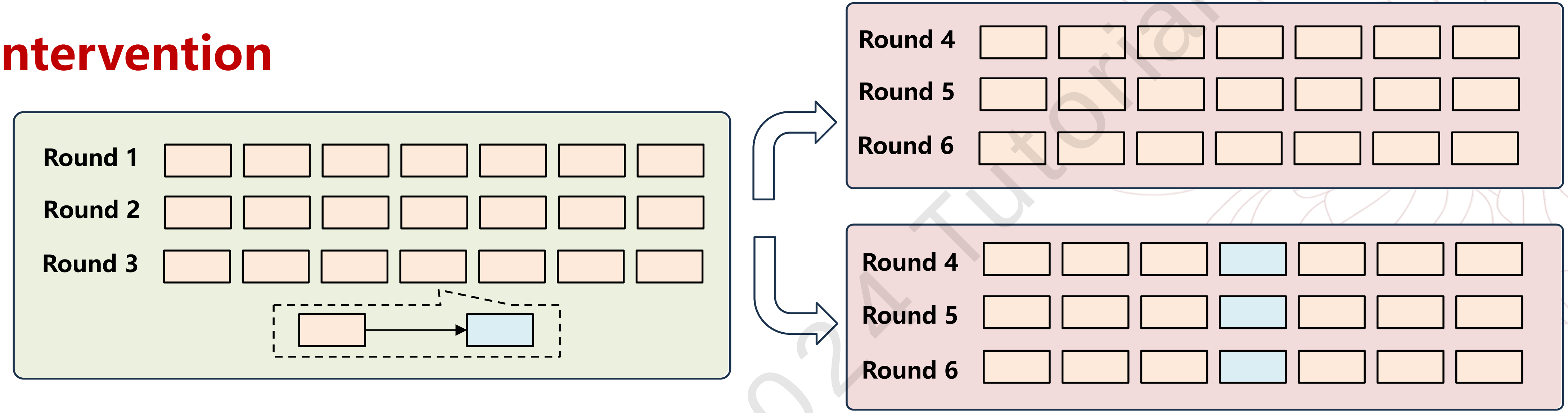
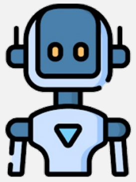


Figure 5: The results of using $p(x)$ to fit real-world datasets. The blue points are the real-world data, and the red lines are the fitted distributions.

System-level Design

Intervention





David Smith

Before Intervention

Traits: adventurous, energetic, ambitious, optimistic
Interest: sci-fi movies, thriller movies, suspense movies

After Intervention

Traits: introverted, cautious, quick-tempered
Interest: family-friendly movies, romantic movies, comedy movies

[David Smith]: I haven't come across any classics lately, but I did watch this amazing sci-fi thriller called <Inception>. It's mind-blowing! You should definitely check it out. ...
[David Smith]: I'll definitely keep an ear out for any exciting sci-fi movies and let you know. We both know how much we love that genre!

Original Branch

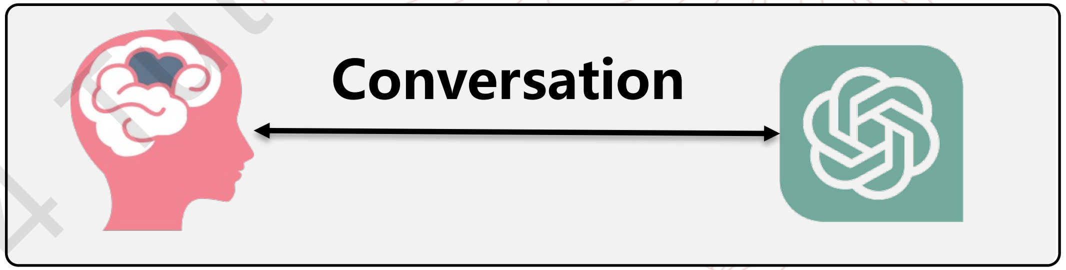
[David Smith]: That's great! I'm more into sci-fi, thriller, and suspense movies. They always keep me on the edge of my seat. Have you watched any good movies lately?
[David Smith]: Wow, that's quite a list! I'm glad you enjoyed them. Based on your interest in "The Matrix" and "Inception," I would recommend "Blade Runner" for its mind-bending concept and suspenseful elements.

Intervention Branch

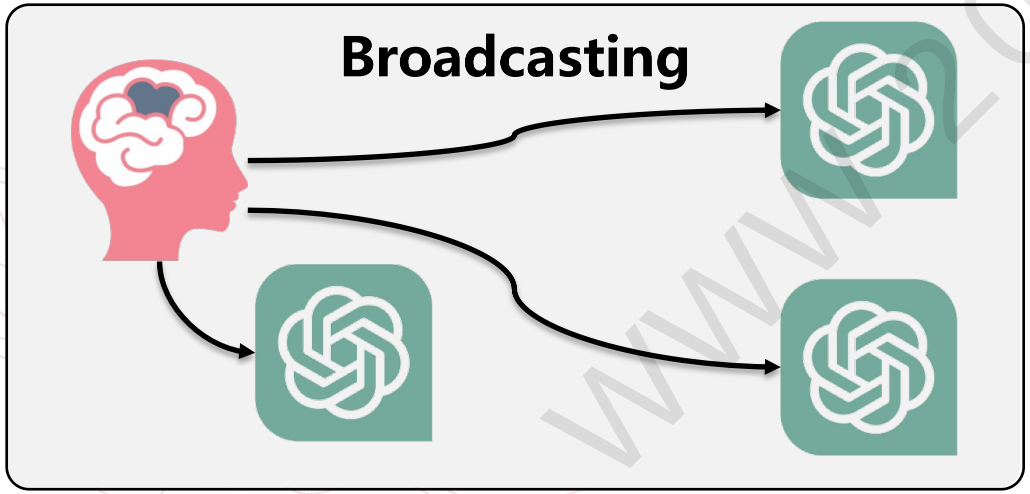
[David Smith]: I love movies that really make you think. I'm definitely going to check them out. By the way, have you come across any good family-friendly or romantic movies? I'm in the mood for something heartwarming.
[David Miller]: Absolutely! If you're looking for a heartwarming movie, I recently watched <Miracle on 34th Street> on the recommender system, and it was delightful.

System-level Design

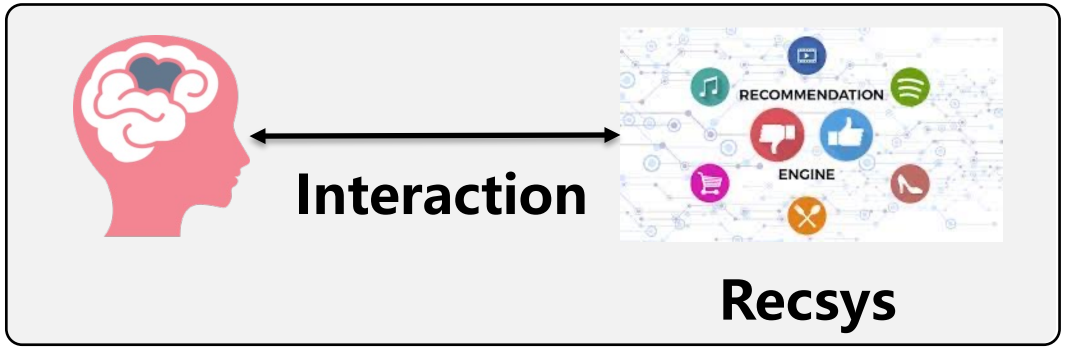
Human-Agent Collaborative Simulation



Human-agent Conversation

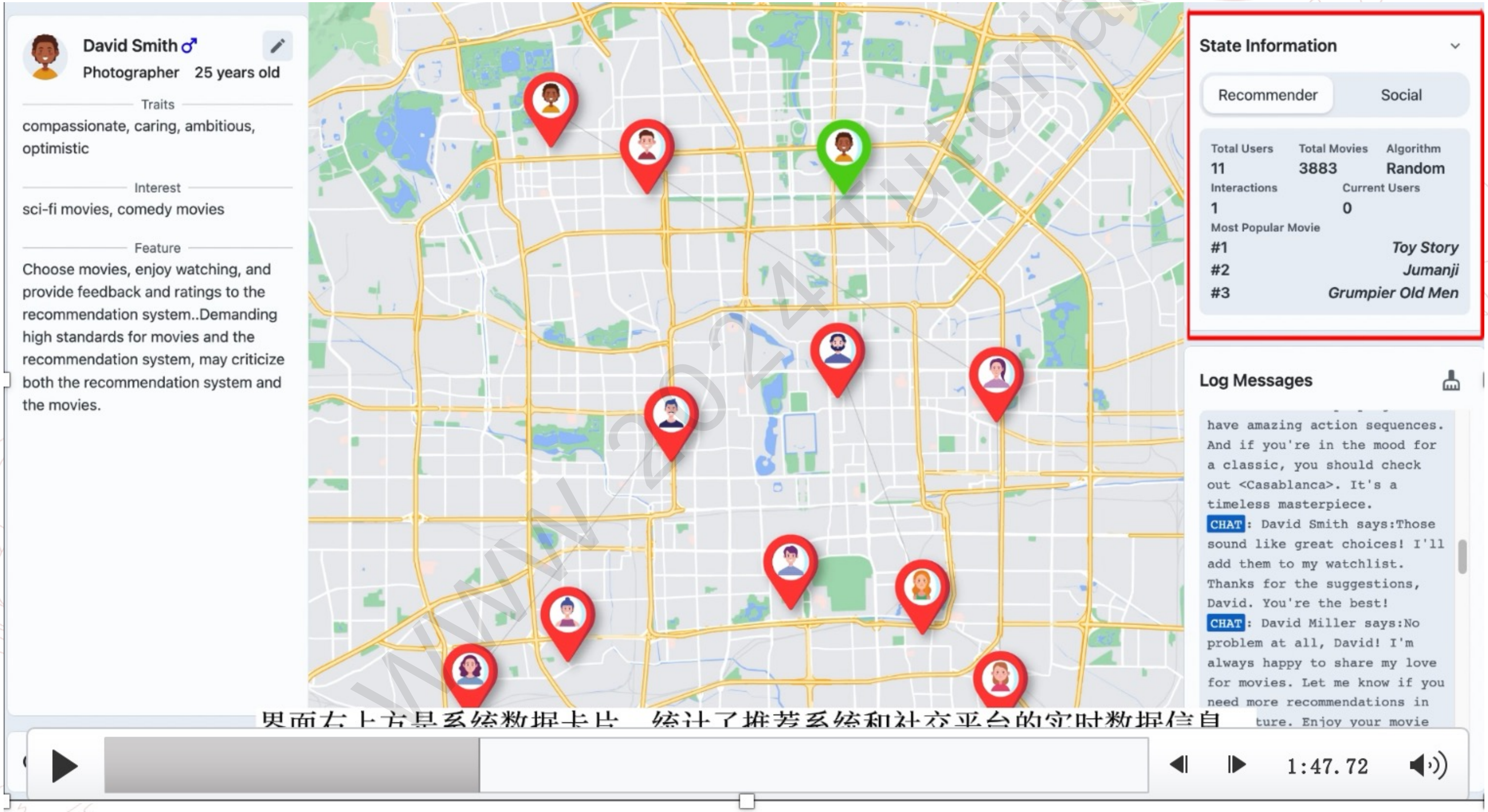


Human-agent social broadcasting



Human-system Interaction

Video Demo


David Smith
Photographer 25 years old
Traits
compassionate, caring, ambitious, optimistic
Interest
sci-fi movies, comedy movies
Feature
Choose movies, enjoy watching, and provide feedback and ratings to the recommendation system..Demanding high standards for movies and the recommendation system, may criticize both the recommendation system and the movies.
State Information
Recommender
Social
Total Users
11
Total Movies
3883
Algorithm
Random
Interactions
1
Current Users
0
Most Popular Movie
#1 Toy Story
#2 Jumanji
#3 Grumpier Old Men
Log Messages
have amazing action sequences. And if you're in the mood for a classic, you should check out <Casablanca>. It's a timeless masterpiece.
CHAT: David Smith says:Those sound like great choices! I'll add them to my watchlist. Thanks for the suggestions, David. You're the best!
CHAT: David Miller says:No problem at all, David! I'm always happy to share my love for movies. Let me know if you need more recommendations in
1:47.72

Experiments - Agent-level Evaluation

Experiment Setting

Goal: whether the agent memory can produce reasonable results

- Let the agents and humans finish the same memory-related tasks
- Recruit another group of humans to judge which one is more reasonable

Results

Table 1: The results of evaluating sensory memory (T1), short-term memory (T2), and long-term memory (T3). A and B indicate the results generated by the agent and real human, respectively. "$>>$", "$>$", and "$\approx$" mean significantly better, slight better and comparable, respectively.

	$A >> B$	$A > B$	$A \approx B$	$B > A$	$B >> A$
T1	0.6833	0.2500	0.0333	0.0333	0.0000
T2	0.3000	0.3000	0.1000	0.2500	0.0500
T3	0.2500	0.1167	0.2000	0.2500	0.1667

Experiments - Agent-level Evaluation

Experiment Setting

Goal: whether the extracted memory are informative and relevant

- Randomly sample 15 agent behaviors
- Recruit three human annotators to evaluate the extracted information
- Consider both informativeness and relevance

Results

Table 2: The results of evaluating the memory module. We use bold fonts to label the best results.

Model	Informativeness	Relevance
Memory module (w/o short)	4.09	4.02
Memory module (w/o long)	**4.55**	3.75
Memory module (w/o reflection)	4.40	3.63
Memory module	4.42	**4.09**

Experiments - System-level Evaluation

Experiment Setting

Goal: whether the agents can separate real items from irrelevant ones

- 20 Users from Movielens-1M
- Combine the **a** ground truths with **b** negative items
- Comparing the selection accuracy

Results

Table 3: The results of evaluating different models based on different (a, b)'s.

Model	$(a, b) = (1, 5)$	$(a, b) = (3, 3)$	$(a, b) = (3, 6)$	$(a, b) = (1, 9)$
Embedding	0.2500	0.5500	0.4500	0.3000
RecSim	0.2500	0.5333	0.3667	0.1000
RecAgent	0.5500	0.7833	0.6833	0.5000
Real Human	0.6000	0.8056	0.7222	0.5833

Experiments - System-level Evaluation

Experiment Setting

Goal: whether the agents can generate reliable user behavior sequences

Results

Table 4: The results of evaluating the reliability of the generated user behavior sequences (N=5).

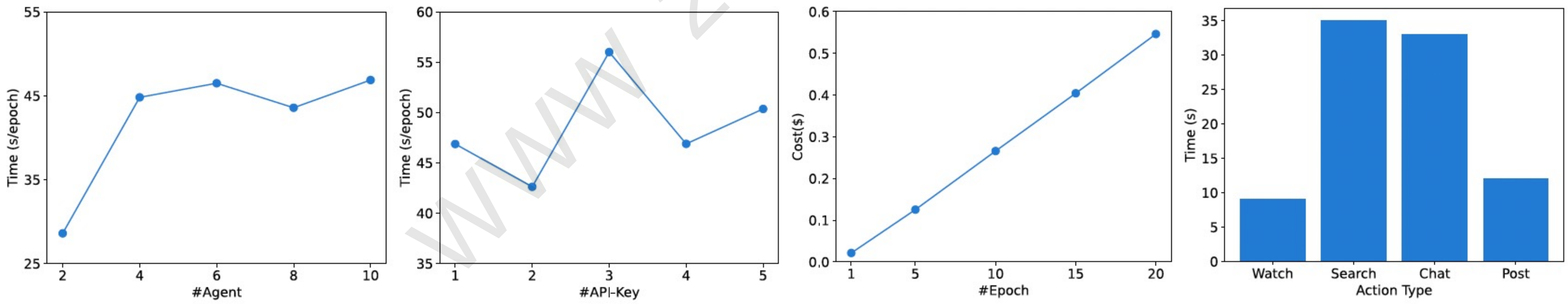
A v.s. B	$A >> B$	$A > B$	$A \approx B$	$B > A$	$B >> A$
RecAgent v.s. RecSim	0.1500	0.3167	0.1833	0.2667	0.0833
RecAgent v.s. GT	0.1333	0.2833	0.1667	0.2667	0.1500
RecSim v.s. GT	0.1167	0.2667	0.2667	0.2167	0.1333

Table 5: The results of evaluating the reliability of the generated user behavior sequences (N=10).

A v.s. B	$A >> B$	$A > B$	$A \approx B$	$B > A$	$B >> A$
RecAgent v.s. RecSim	0.1833	0.4333	0.0667	0.2000	0.1167
RecAgent v.s. GT	0.2000	0.4333	0.0000	0.2000	0.1667
RecSim v.s. GT	0.1333	0.3500	0.1500	0.3000	0.0667

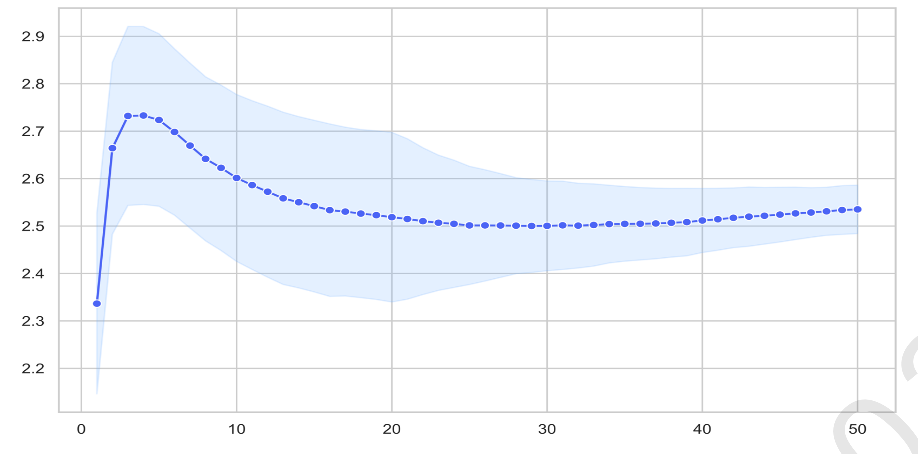
Experiments -Efficiency Analysis

- How does the time cost increase as the number of agents become larger in each epoch?
- How does the time cost increase as the number of API keys become larger in each epoch?
- How does the time cost increase as the number the epochs become larger?
- What are the time costs of different agent behaviors?

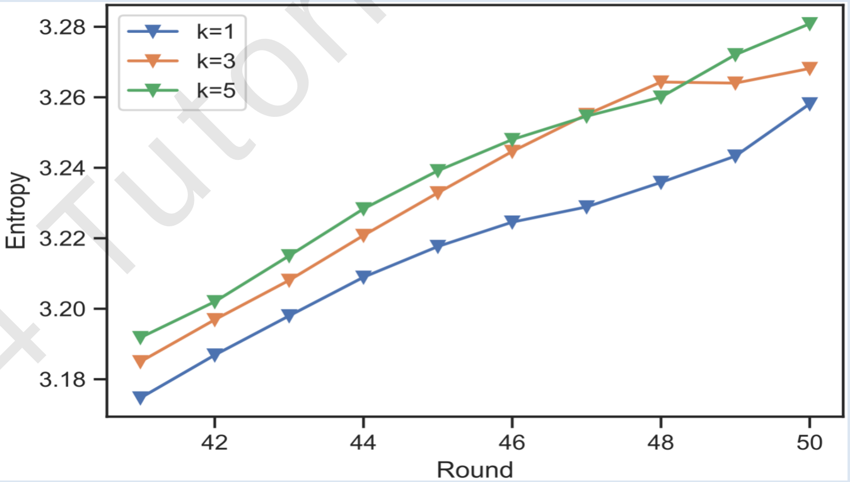


Experiments -Case Studies

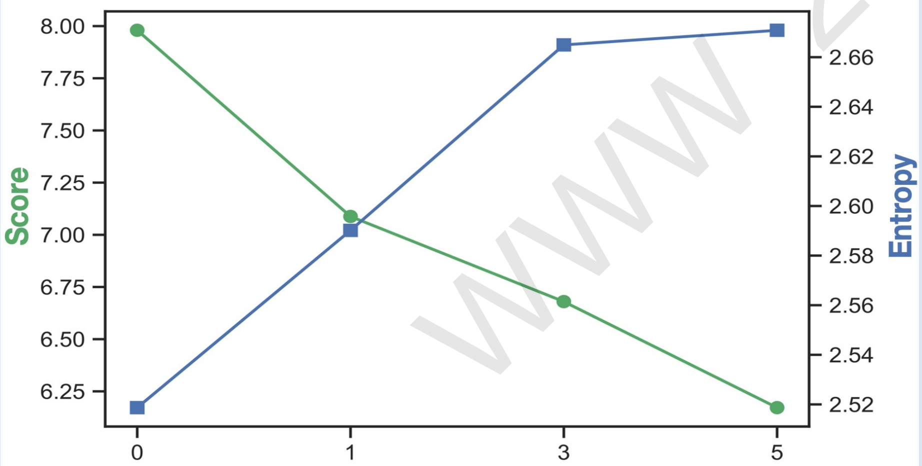
➢ Information Cocoon Room



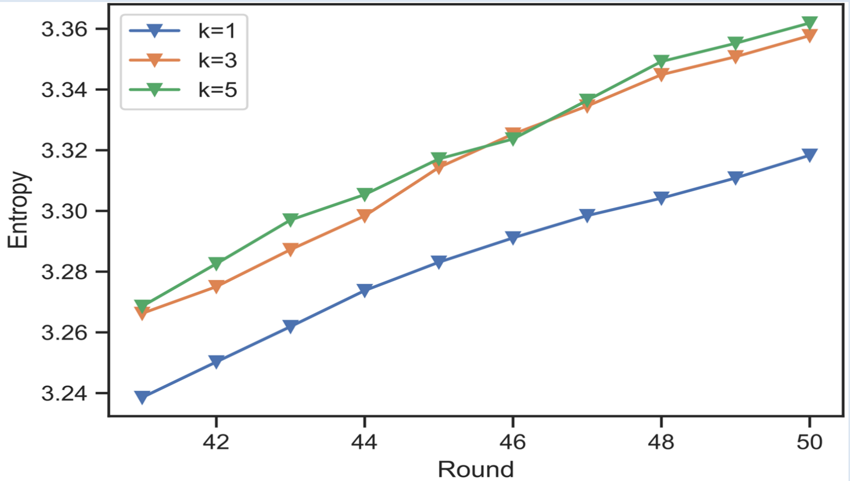
User Entropy



Random Recommendation



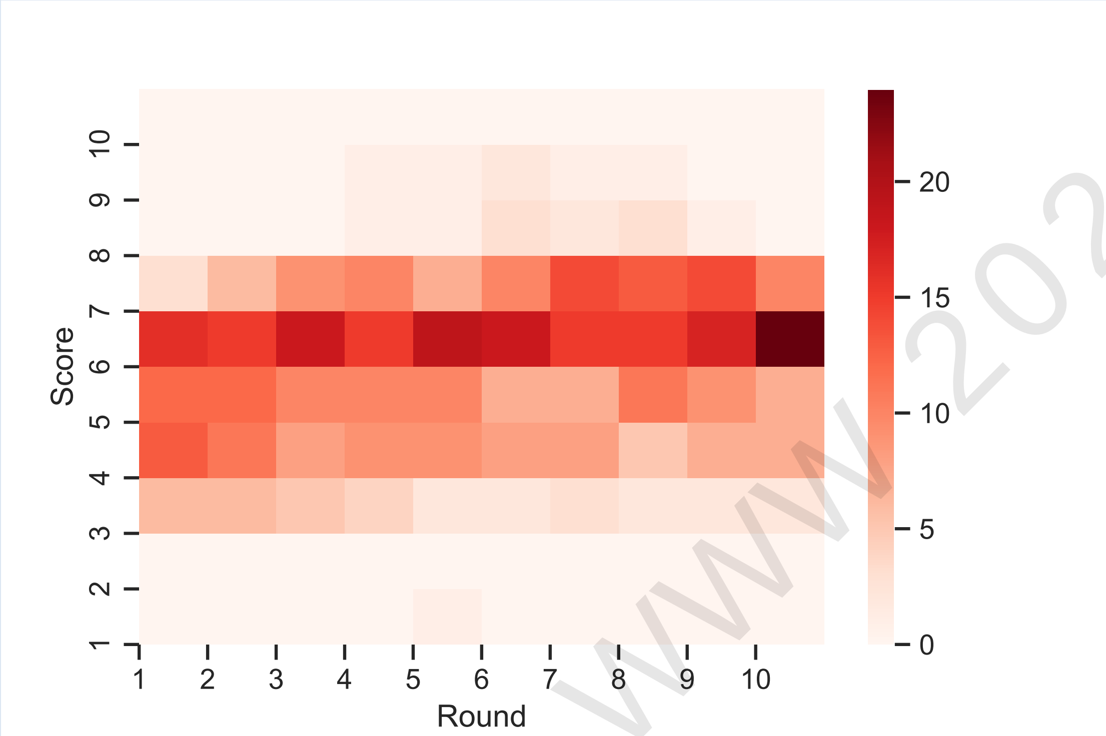
Rec Quality vs Entropy

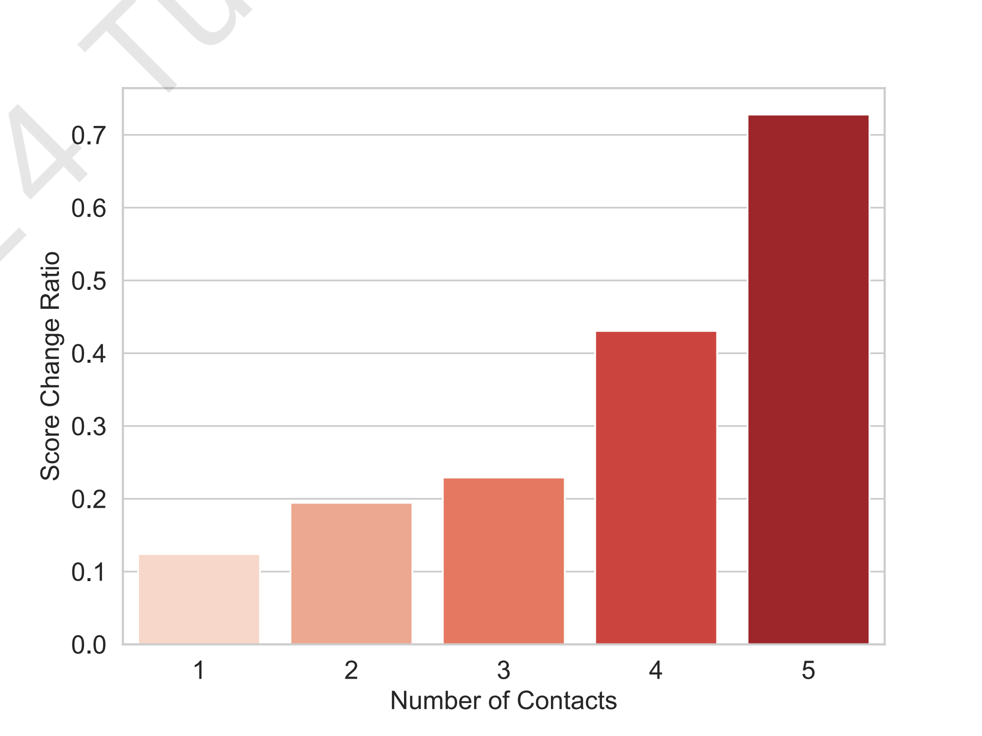


Heterogeneous Ffriends

Experiments -Case Studies

➢ User Conformity Behaviors





Development Process

2023-4 | 2023-5-31 | 2023-9-15 | ?

Beginning of RecAgent

The first version of RecAgent

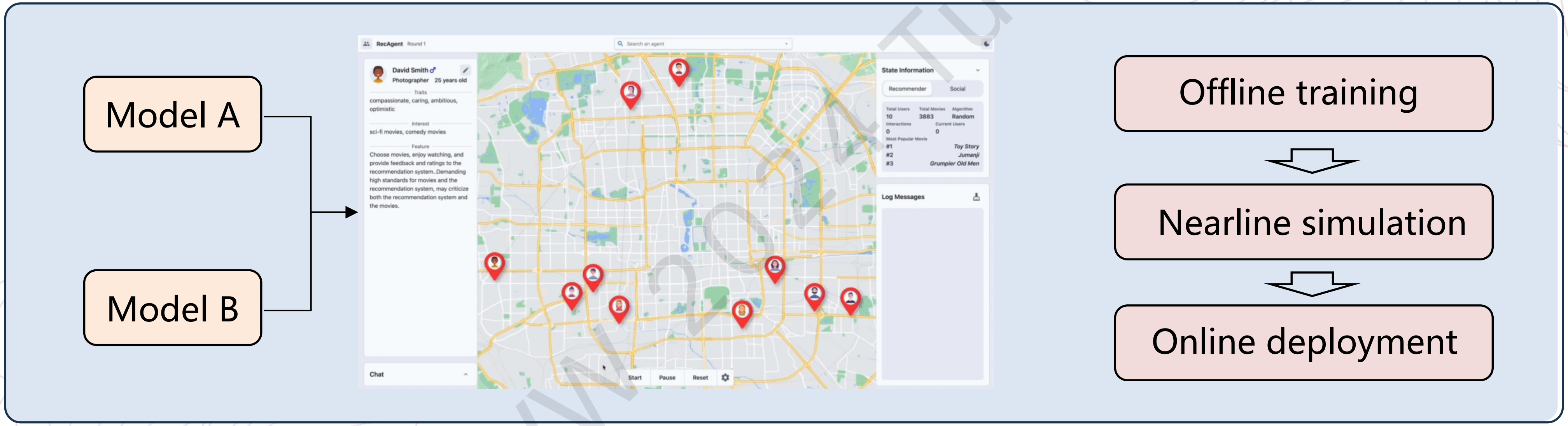
- 👥 25 Agents
- 🧠 Profile, Memory, Action Modules
- 📖 System Environments
- 🕹️ Case Studies

The second version of RecAgent

- 👥 More Agents: from 25 to at most 1000
- 🧠 Human-like Memory Mechanism
- 📖 Comprehensive Experiments
- 🕹️ System-level and Agent-level Intervention
- 🙋 Human Involved Simulation

Potential Impacts

Towards more comprehensive, explainable , controllable, efficient and less expensive recommender model evaluation before deployment



- Study are the reasons of the performance
- Study the performance evolution process
- Save much online cost
- Flexible environment settings

Future Direction

General recommendation

- Cold-start recommendation
- Dataset augmentation
- Data sparsity

RL-based recommendation

- Act as a simulator
- More comprehensive feedback
- Human-like user simulation

Explainable recommendation

- Explanation ground truth
- Multi-type explanations
- Interactive explanations

Causal recommendation

- Counterfactual world simulation
- Flexible intervention experiments
- Counterfactual ground truth

Future Direction

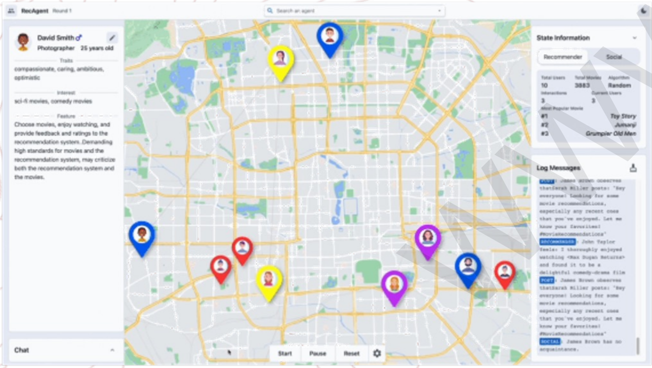
Studying the effects of emergence events (promotion, advertisement)

Deploying an adv-agent

⇩

Sending promotion information on the social network

⇩



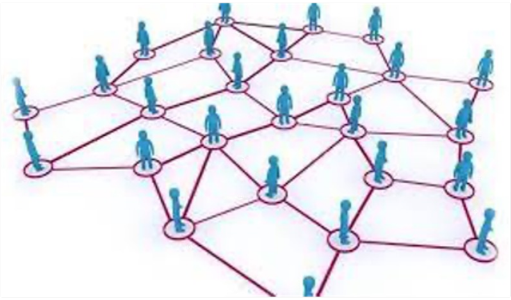
⇨

- How many users does an advertisement can influence?
- How do the advertisements diffuse in different user groups?
- How do the advertisements influence the item recommendation performance? And what are the reasons?
- …

- How to design optimal advertisement policies to enhance user CTR?
- How to design promotion words to enhance the pervasiveness?
- How to jointly design the advertisement and recommendation policies?
- …

Future Direction

Studying the influence/propagation of social information



- How fast does the social information diffuse?
- Which user groups are more suitable for social information diffusion?
- The process of friend relation building between two users?

Providing inspirations for human-centered AI applications

Economic behaviors

Court simulation

Policy simulation

Related Resource

RECAGENT

Project Page：https://github.com/RUC-GSAI/YuLan-Rec
Paper Link：https://arxiv.org/pdf/2306.02552.pdf
Chinese Introduction: https://mp.weixin.qq.com/s/bfES1ieY5pTtmVfdEgX6WQ

Related Resource

A Survey on Large Language Model based Autonomous Agents

Lei Wang, Chen Ma*, Xueyang Feng*, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, Wayne Xin Zhao, Zhewei Wei, Ji-Rong Wen
Gaoling School of Artificial Intelligence, Renmin University of China, Beijing, China

Unwatch 44 | Fork 84 | Starred 1.4k

A survey on large language model based autonomous agents 39 2023
L Wang, C Ma, X Feng, Z Zhang, H Yang, J Zhang, Z Chen, J Tang, ...
arXiv preprint arXiv:2308.11432

- The first survey paper in the field of LLM-based Agents
- Summarize 200+ papers： https://abyssinian-molybdenumf76.notion.site/237e9f7515d543c0922c74f4c3012a77
- GitHub page： https://github.com/Paitesanshi/LLM-Agent-Survey
- Paper digest：https://github.com/XueyangFeng/LLM-Agent-Paper-Digest

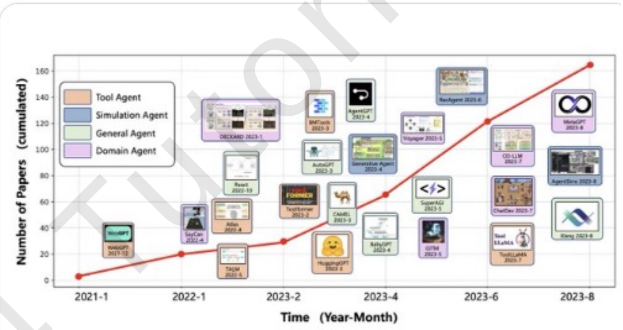
elvis @omarsar0 · Aug 31
A Survey on LLM-based Autonomous Agents

Great repository containing a collection of papers on LLM-based autonomous agents.

The survey paper for this came out a few days ago as well.

repo: github.com/Paitesanshi/LL...

paper: arxiv.org/abs/2308.11432



Sanyam Bhutani @bhutanisanyam1 · Aug 24
Top Down Overview of LLM Autonomous **Agents**!

Large Language Model backbones form a tiny part of making autonomous **agents** work

A large part is the clever engineering and architecture of the systems

This paper focusses on giving us an architectural overview and then connects...
Show more

2 | 11 | 50 | 5,920

AI自主智能体大盘点，构建、应用、评估全覆盖，人大高瓴文继荣等32页综述

机器之心 2023-08-28 13:05 Posted on 北京
20 listened

AI自主智能体大盘点，构建、应用、评估全覆盖，人大高瓴文继荣等32页综述

机器之心 2023-08-28 13:05 Posted on 北京
20 listened

模型和智能体如何结合？人大最《基于大型语言模型的自主智能》综述

2023-08-24 17:04 Posted on 北京

体的构建、潜在应用和评估，为全面了解该领域的发展以及启发未来的研究具有重要意义。

当今的 AI 时代，自主智能体被认为是通向通用工智能（AGI）的一条有前途的道路。所谓自主能体，即能够通过自主规划和指令来完成任务。早期的开发范式中，决定智能体行动的策略功能以启发式为主的，并在环境交互中逐步得到完。

过，在不受约束的开放域环境中，自主智能体的动往往很难企及人类水平的熟练程度。

Thanks & QA

Simulating Human Society with LLM Agents: City, Social Media, and Economic System

Tutorial at The Web Conference 2024 in Singapore (WWW 2024)

Xiang Wang

May 13, 2024, Singapore

Outline

- Background of LLM Agent-based Simulation (25minutes) 13:30-13:55
- Online behavior simulation with LLM Agents (65 minutes) 13:55-15:00
- Break (15 minutes) 15:00-15:15
- Social and economic simulation with LLM agents (50 minutes) 15:15-16:05
- City system simulation with LLM agents (45minutes) 16:05-16:50
- Open discussions (10minutes) 16:50-17:00

Motivation
LLMs are not AGI

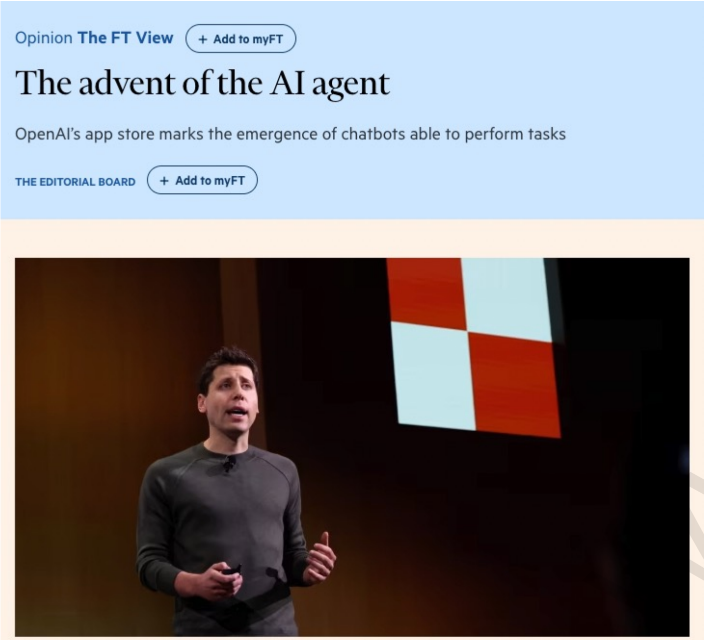
Aim of AGI

- Large LLMs exhibit characteristics of artificial general intelligence (AGI), which has cognitive abilities similar to that of human.
- In other words, AI can now perform most functions that humans are capable of doing.

Autonomous AI Agents
Importance

AI Agents

- LLM-powered Agents are artificial entities that enhance LLMs with essential capabilities, enabling them to sense their environment, make decisions, and take actions.



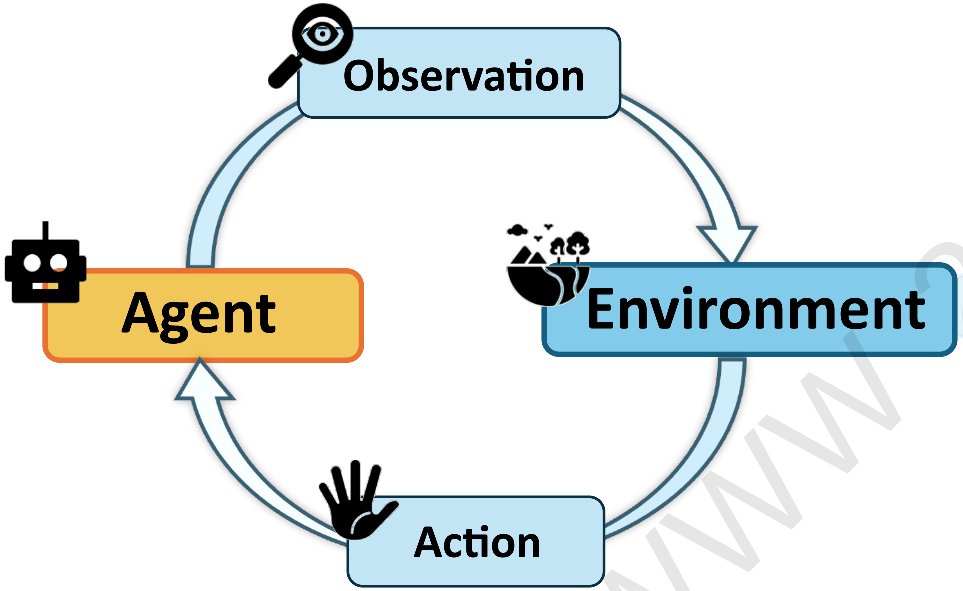
OpenAI CEO Sam Altman speaks during the OpenAI DevDay event this week in San Francisco, California © Justin Sullivan/Getty Images

- Sam Altman (Former CEO of OpenAI) himself said in his keynote: "GPTs and Assistants are precursors to agents. They will gradually be able to plan and to perform more complex actions on your behalf. These are our first step toward AI Agents."
- Bill Gates said in his BLOG: "Agents are not only going to change how everyone interacts with computers. They're also going to upend the software industry, bringing about the biggest revolution in computing since we went from typing commands to tapping on icons."
- Application: be my eyes AI-powered visual assistance.

News in Financial Times. " The advent of the AI agent" .
GatesNotes. "The Future of Agents: AI is about to completely change how you use computers".

LLM-powered Agents
Environment

- This paves the way for the use of AI agents to simulate users and other entities, as well as their interactions.



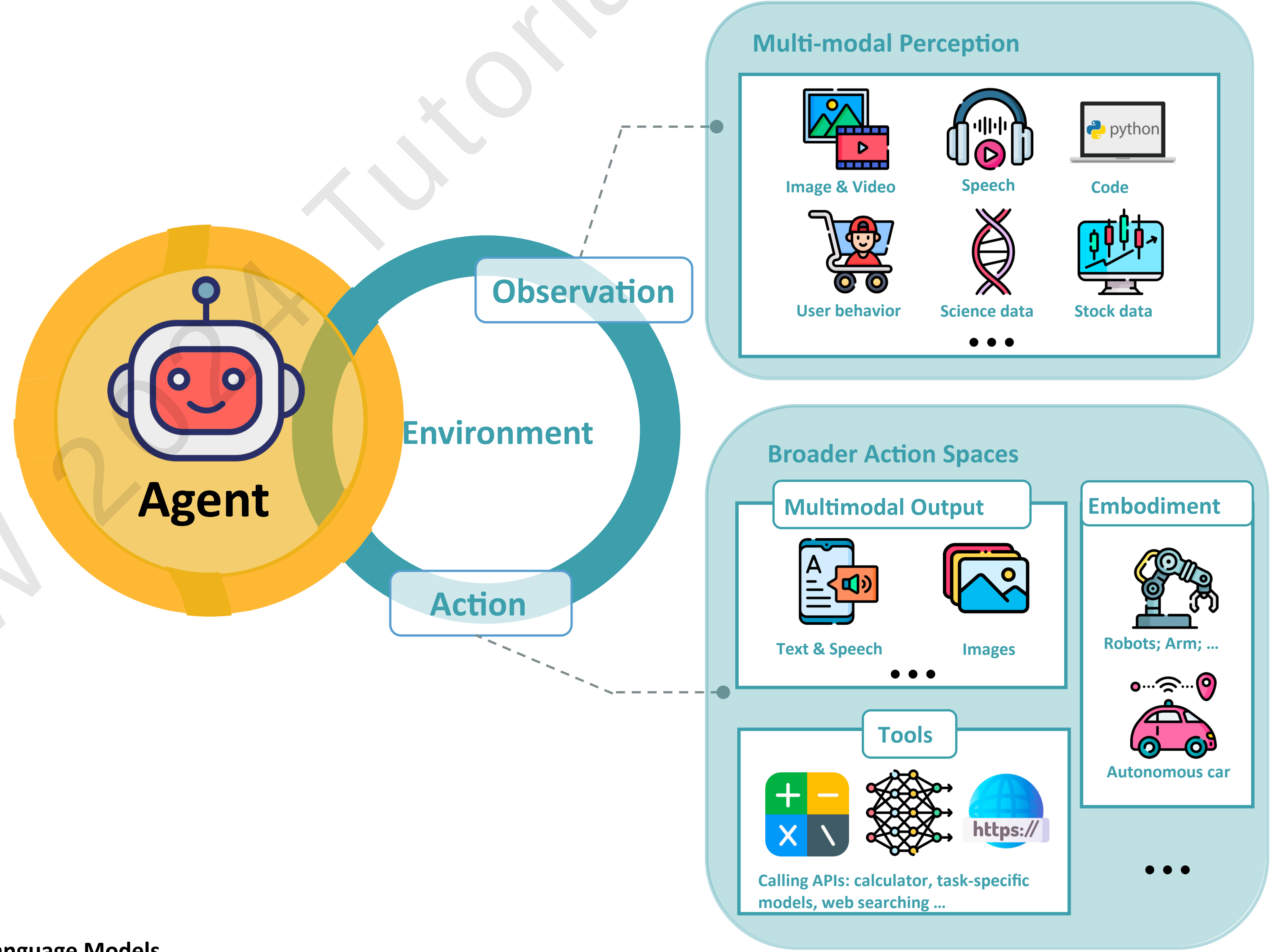
Environment

➢ The external context or surroundings in which the agent operates and makes decisions.

- Human & Agents' behaviors
- External database and knowledges

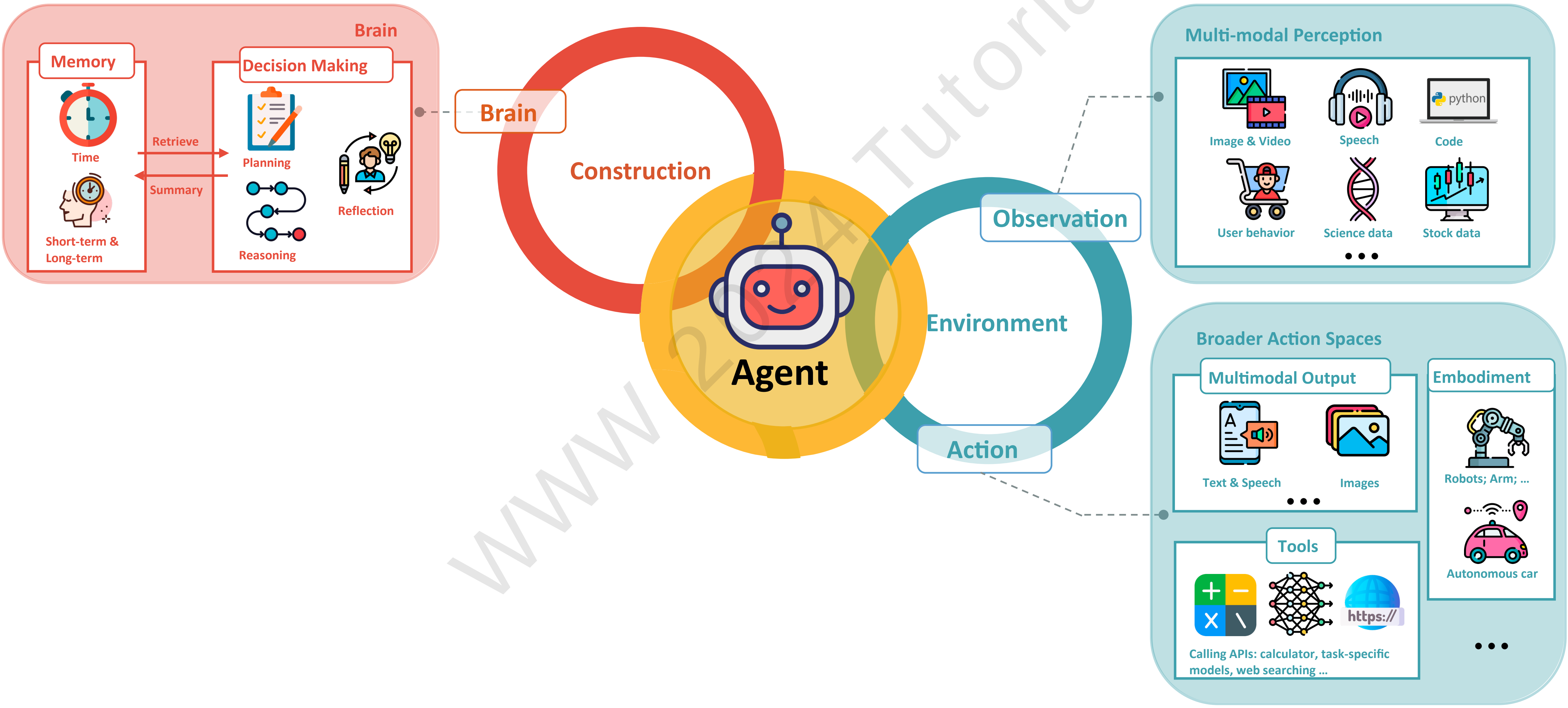
- Virtual & Physical environment

LLM-powered Agents
Observation & Action

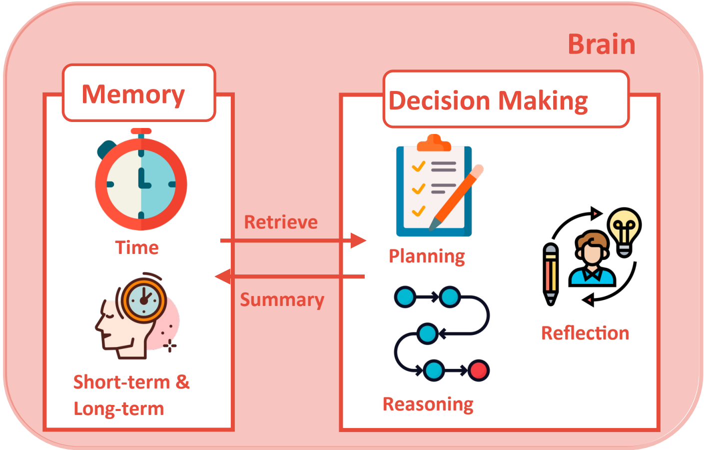


Guanzhi Wang et al., Voyager: An Open-Ended Embodied Agent with Large Language Models.

LLM-powered Agents
Brain



LLM-powered Agents
Brain



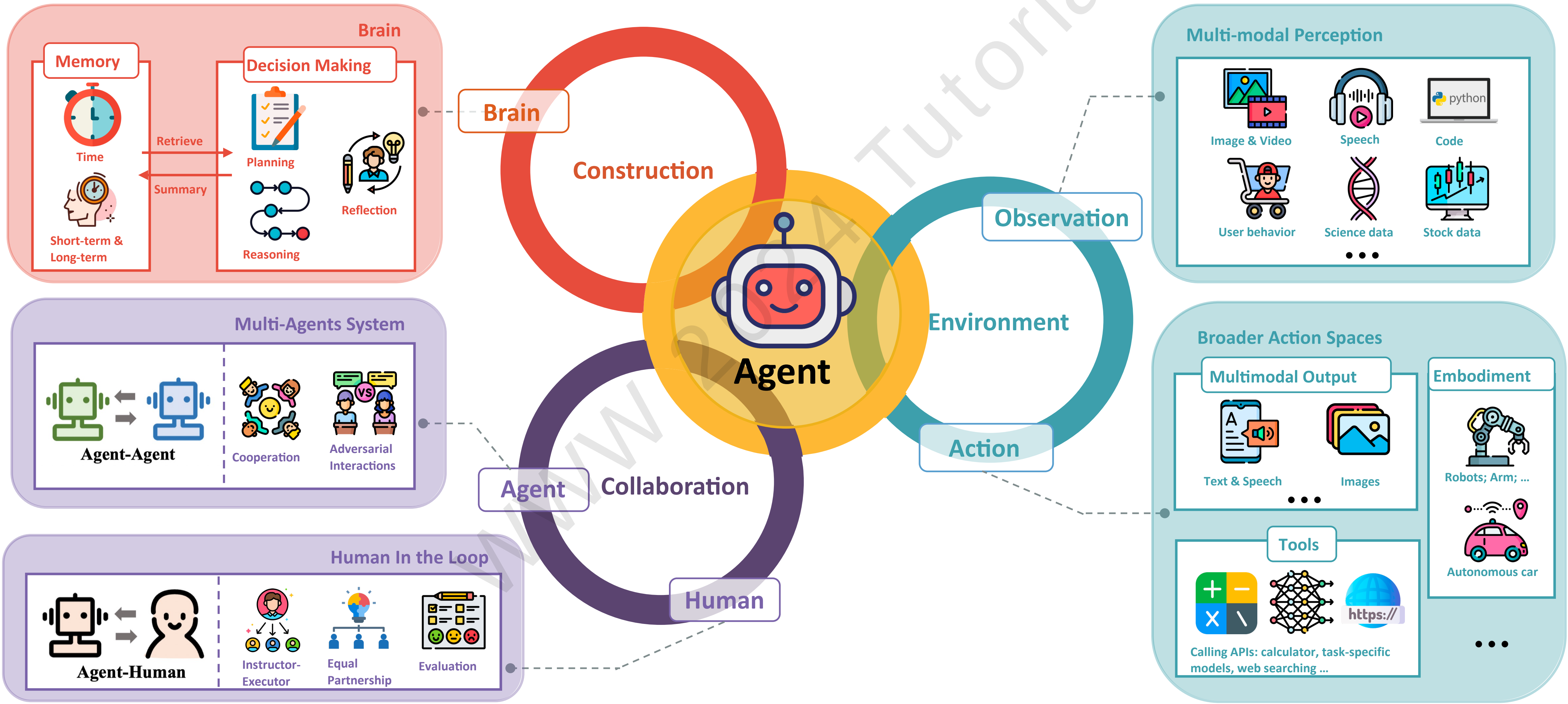
- ❑ **Memory: "memory stream" stores sequences of agent's past observations, thoughts and actions:**
  - ➢ Sufficient space for long-term and short-term memory;
  - ➢ Abstraction of long-term memory;
  - ➢ Retrieval of past relevant memory;

- ❑ **Decision Making Process:**
  - ➢ **Planning:** Subgoal and decomposition: Able to break down large tasks into smaller, manageable subgoals, enabling efficient handling of complex tasks.
  - ➢ **Reasoning:** Capable of doing self-criticism and self-reflection over past actions, learn from mistakes and refine them for future steps, thereby improving the quality of final results.

- ❑ **Personalized memory and reasoning process foster diversity and independence of AI Agents.**

LLM-powered Agents

Brain



Significant Gap
Between LLMs & RecSys

- Significant gap between large language models (LLMs) and recommender systems (RecSys).

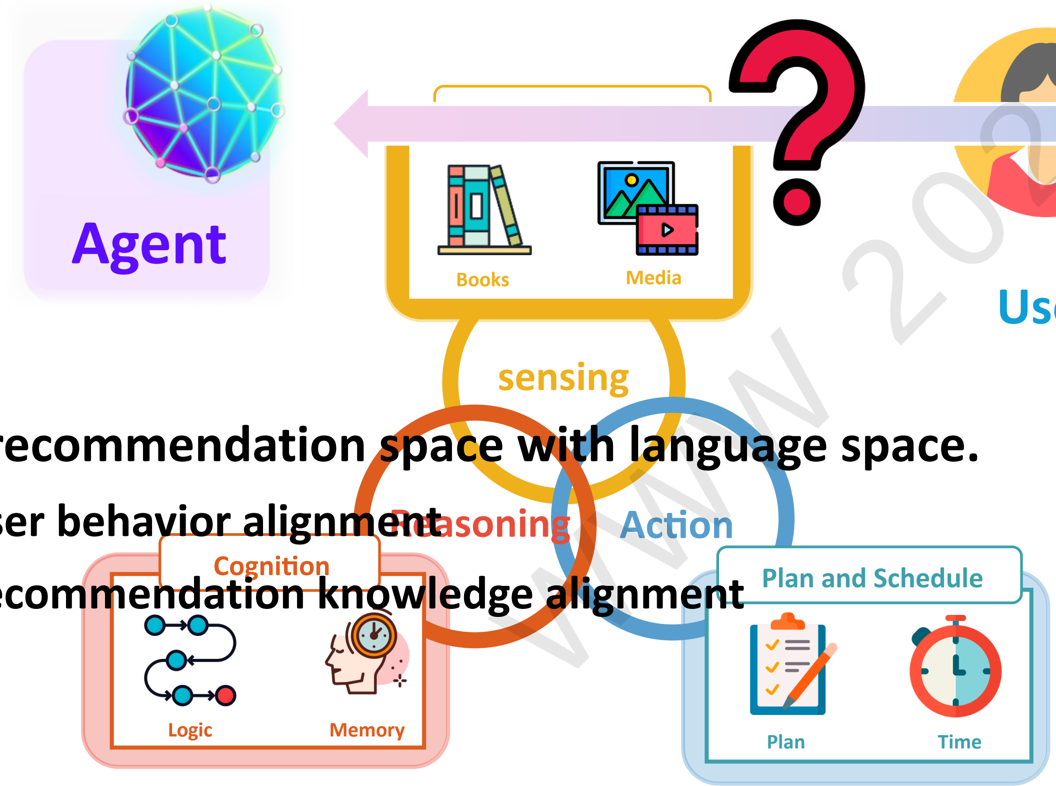
How to bridge this gap?

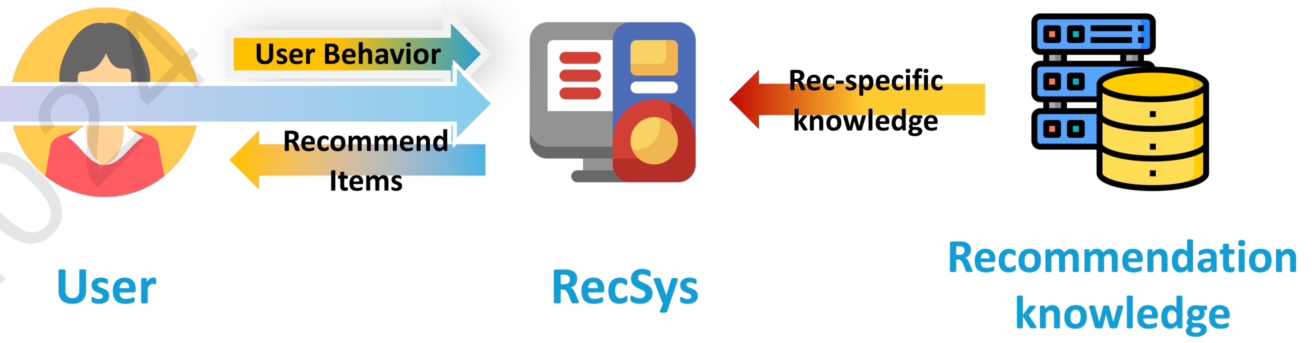
	LLMs	RecSys
Scope	Language modelling	User behaviour modelling
Data	Rich world text-based sources	Sparse user-item interactions
Tokens	A chunk of text (Ten thousand level)	Items (Billion level)
Characteristics	General model; Open-world knowledge; High complexity and long inference time;	Leveraging collaborative signals; Lack of cross-domain adaptability; Struggle with cold-start problem; Limited intention understanding;

Significant Gap Between LLMs & RecSys

➢ Significant **gap** between large language models (LLMs) and recommender systems (RecSys).

How to bridge this gap?

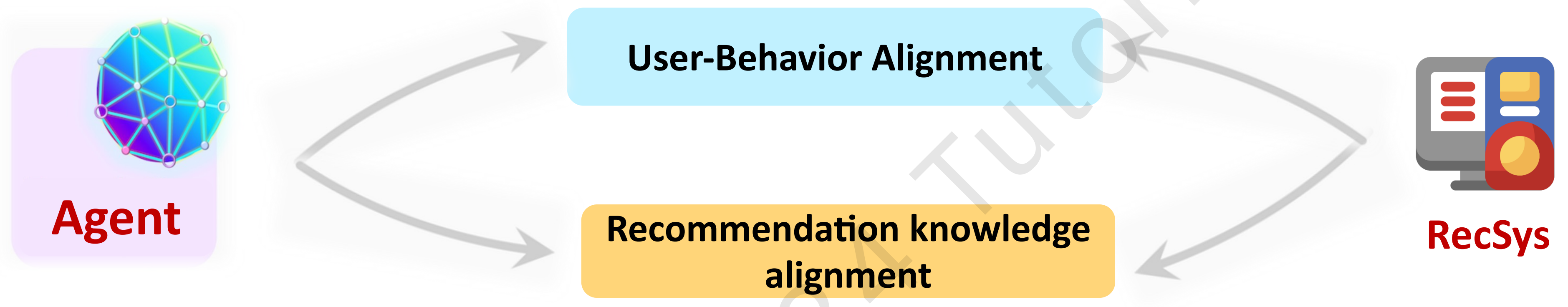




- **Align** recommendation space with language space.
  - User behavior alignment
  - Recommendation knowledge alignment

- Two critical components in RecSys:
  - Understanding user's behavior/preference
  - Acquiring recommendation-specific knowledge

LLM-powered Agents in Recommendation



- LLM-powered Agents have potentials to solve long-standing problems in recommendation
  - Can an LLM-powered Agent faithfully simulate users?
  - Can an LLM-powered Agent be a better recommender with recommendation-specific knowledge?

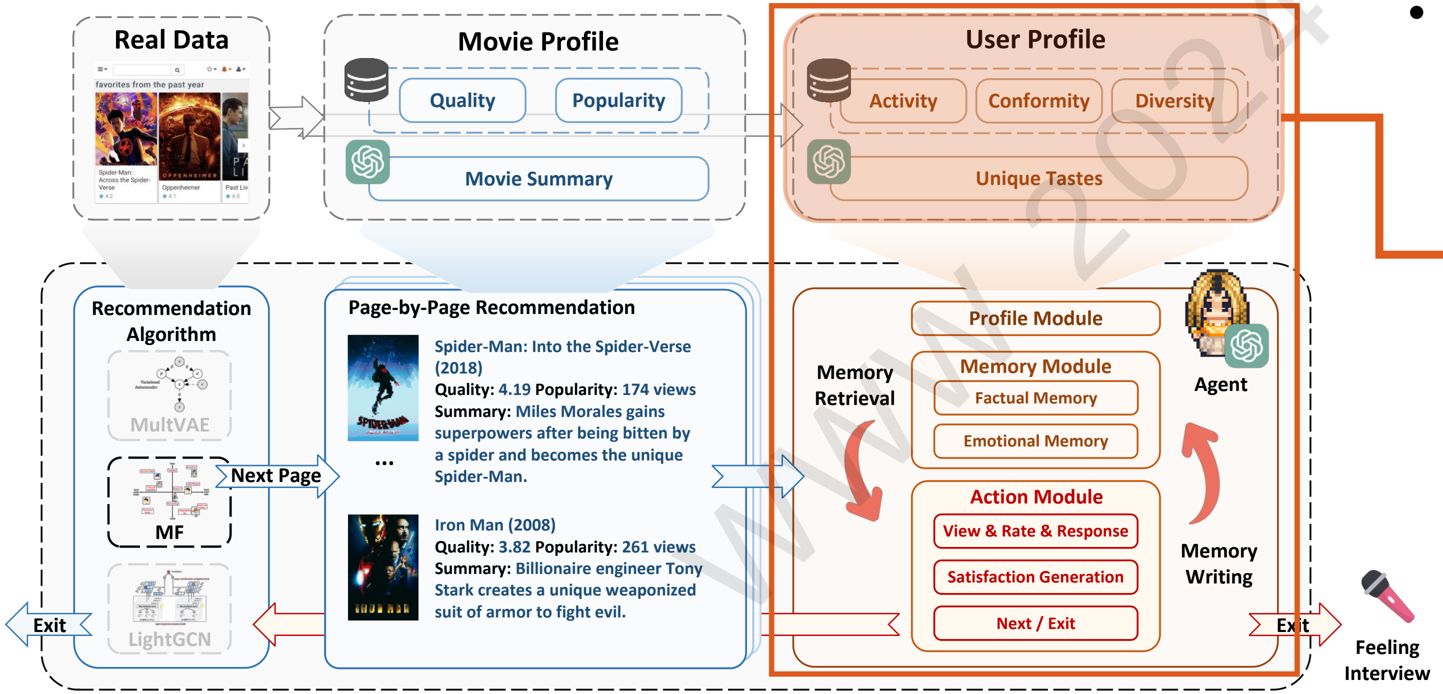
Agents as Users
Agent4Rec

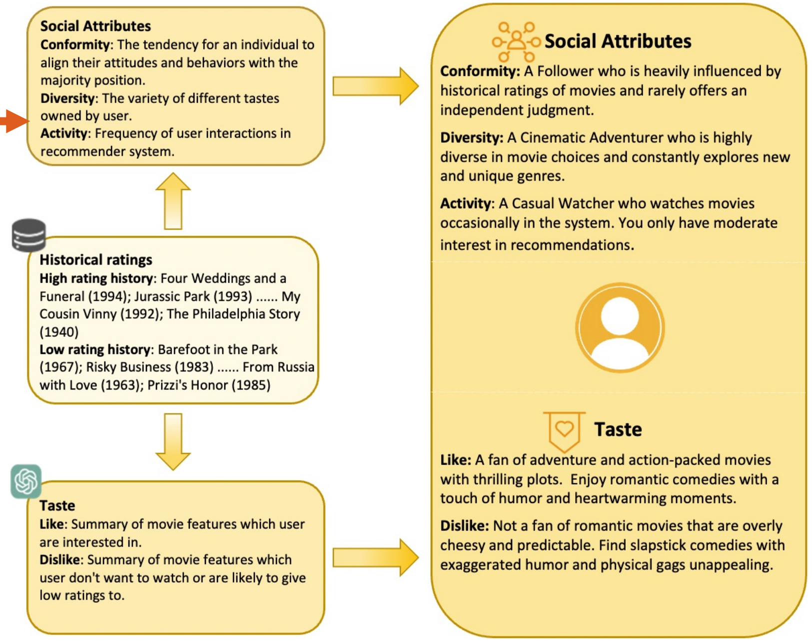
Agents as Users

❑ **Agent4Rec: Agent-driven user behavior simulation**

- **Key Points:**
  - **Can LLM-powered Agent generate faithful user behaviors?**

- **User Profile: 1,000 LLM-empowered generative agents initialized with real data in various dataset and augmented by ChatGPT.**
- **Item Profile: Statistical information in dataset and generated summary.**



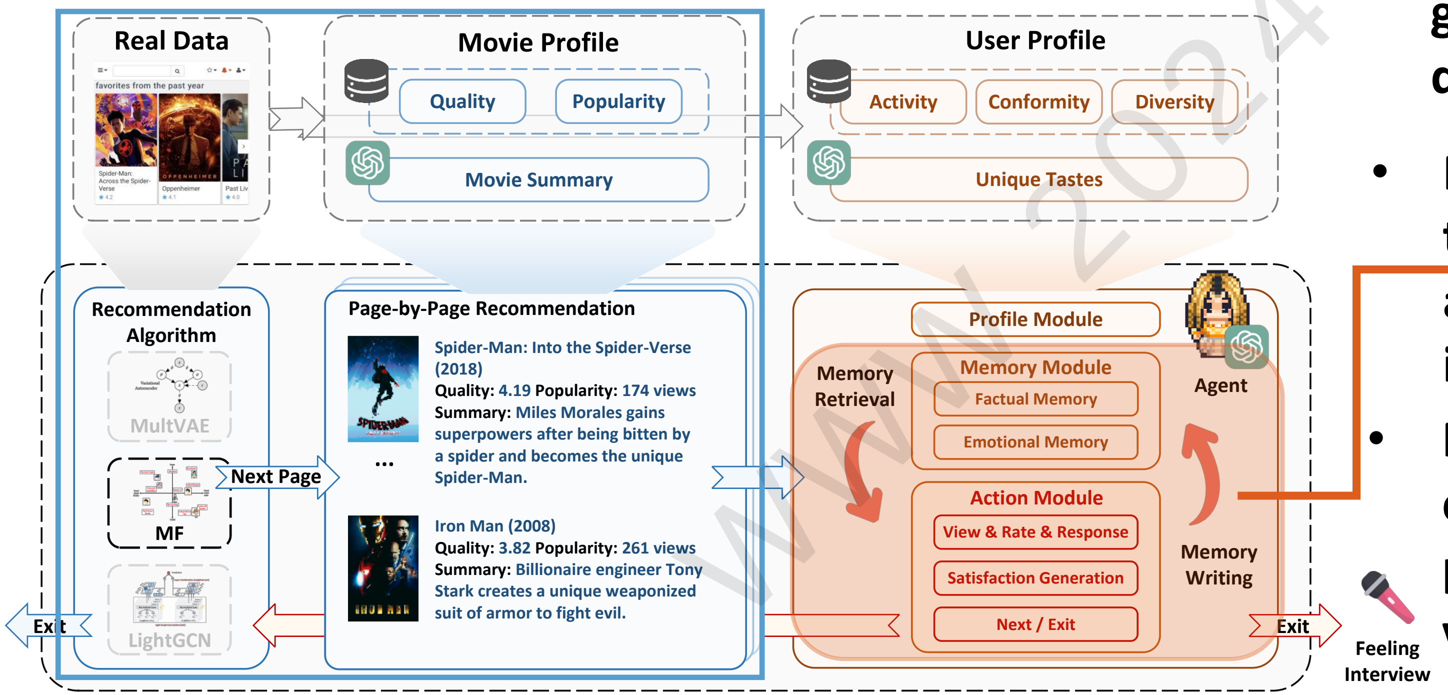


An Zhang et al. On Generative Agents in Recommendation. SIGIR 2024.

Agents as Users

❑ **Agent4Rec: Agent-driven user behavior simulation**

- **Key Points:**
  - **Can LLM-powered Agent generate faithful user behaviors?**



- **Agents as users: 1,000 LLM-empowered generative agents initialized from the real dataset.**
- **Memory and action modules enable agents to recall past interests and plan future actions (watch, rate, evaluate, exit, and interview).**
- **Recommendation environment: Agent4Rec conducts personalized recommendations in a page-by-page manner and pre-implements various recommendation algorithms.**

An Zhang et al. On Generative Agents in Recommendation. SIGIR 2024.

Agents as Users

Agent4Rec

- **Key Observations:**
  - **Agents are capable of preserving the user's social attributes and preference.**
  - **Incorporating agents' rating as augmented data can enhance the recommender's performance.**

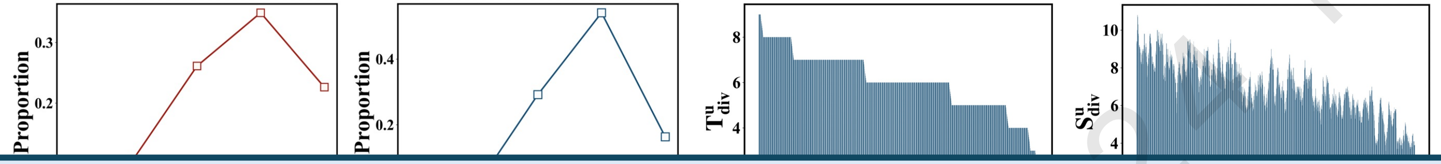
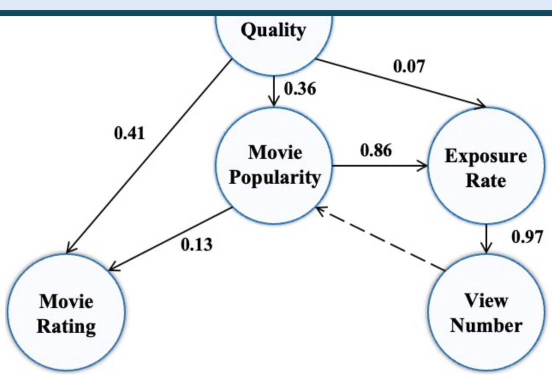


Table 3: Page-by-page recommendation enhancement results over various algorithms.

	MF		MultVAE		LightGCN	
Offline	Recall	NDCG	Recall	NDCG	Recall	NDCG
Origin	0.1506	0.3561	0.1609	0.3512	0.1757	0.3937

LLM-powered agents are able to generate faithful behaviors.

able to discover Causal Relations among movie quality, movie rating, movie popularity, exposure rate, and view number.



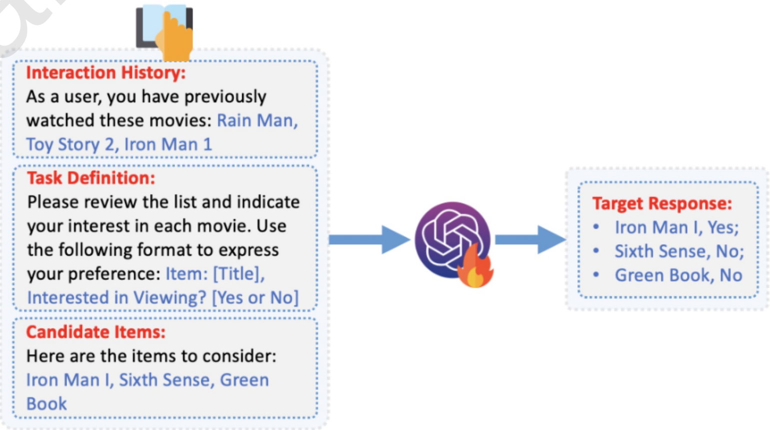
- **Offer a simulation platform to test and fine-tune recommender models.**

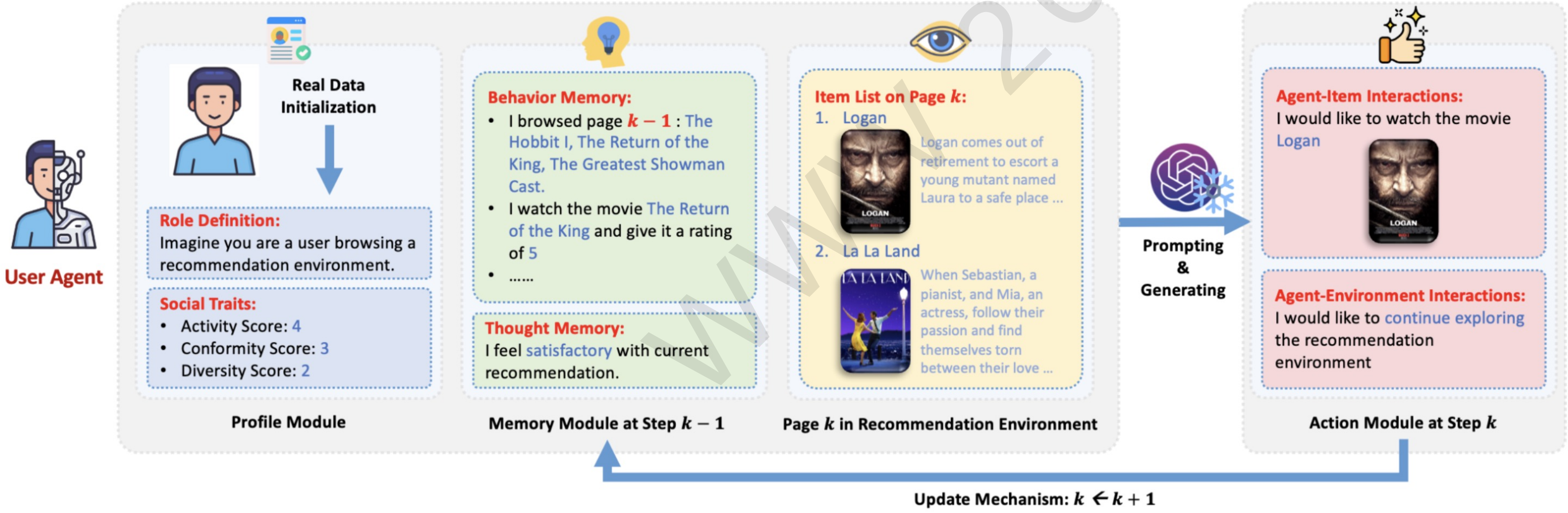
An Zhang et al. On Generative Agents in Recommendation. SIGIR 2024.

Agents as Users
UGen

Agents as Users

- **Key Points :**
  - **Can LLM-powered Agents generated behaviors benefit the recommender?**
  - **Cooperating updated Agent4Rec framework with finetuning GPT-3.5-turbo as a warmup, agents can accurately select their interested items among candidate set.**





- **Agents have potentials to replace discriminative learning with generative learning paradigms for user modeling in recommendation.**
- **Conduct extensive experiments on three dataset from different domains (movie, book, game).**

Agents as Users
UGen

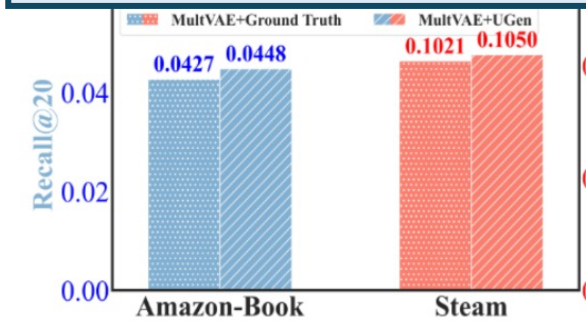
- **Key Observations:**
  - Agents are capable of providing effective behaviors, especially in scenarios with sparse data.

Table 2: Faithfulness Evaluation of Agent's Behavior Alignment with Real User Preferences. Average ground-truth positives are 7.14 (MovieLens), 6.57 (Amazon-Book), and 5.80 (Steam). UGen shows significant improvement with p-value << 0.05.

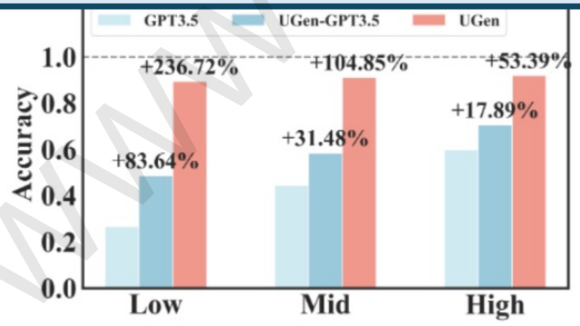
	MovieLens				Amazon-Book				Steam			
	Acc	Pre	Rec	#Select	Acc	Pre	Rec	#Select	Acc	Pre	Rec	#Select
GPT3.5	0.5295	0.4307	0.7369	11.63	0.4202	0.3855	**0.9072**	17.10	0.4350	0.3430	0.9164	16.59
GPT4	0.6930	0.5743	0.6577	7.00	0.7947	0.6500	0.6003	5.16	0.7844	0.5103	0.7072	6.22
RecAgent	0.6168	0.4519	**0.8921**	13.95	0.5411	0.3714	0.8150	14.65	0.4916	0.3485	**0.9389**	15.55
RAH	0.5758	0.4096	0.6383	9.44	0.7253	0.3355	0.3950	7.45	0.6118	0.3874	0.6262	10.37

	MovieLens		Amazon-Book		Steam	
	Recall@20	NDCG@20	Recall@20	NDCG@20	Recall@20	NDCG@20
MF	0.1529	0.3186	0.0257	0.0480	0.0694	0.0567
+ Random	0.1365	0.2913	0.0199	0.0225	0.0526	0.0432
+ GPT3.5	0.1448	0.3089	0.0253	0.0330	0.0732	0.0608
+ RecAgent	0.1400	0.2990	0.0254	0.0317	0.0696	0.0567
+ RAH	0.1363	0.2917	0.0257	0.0370	0.0731	0.0604
+ UGen	**0.1667**	**0.3396**	**0.0413**	**0.0573**	**0.0807**	**0.0659**
Imp.% over MF	*9.03%*	*6.59%*	*60.70%*	*19.38%*	*16.28%*	*16.23%*
+ Random	0.1650	0.3358	0.0257	0.0354	0.0762	0.0604
+ GPT3.5	0.1693	0.3462	0.0408	0.0536	0.0817	0.0694
+ RecAgent	0.1650	0.3393	0.0386	0.0518	0.0802	0.0668
+ RAH	0.1597	0.3340	0.0391	0.0542	0.0867	0.0719
+ UGen	**0.1899**	**0.3722**	**0.0555**	**0.0752**	**0.1140**	**0.0952**
Imp.% over LightGCN	*2.82%*	*2.59%*	*32.14%*	*12.24%*	*28.67%*	*25.76%*

Behaviors generated by LLM-powered agents can benefit recommenders.



(a) Augmented MultVAE



(b) Accuracy on Amazon-Book

Table 4: Human Evaluation on Steam

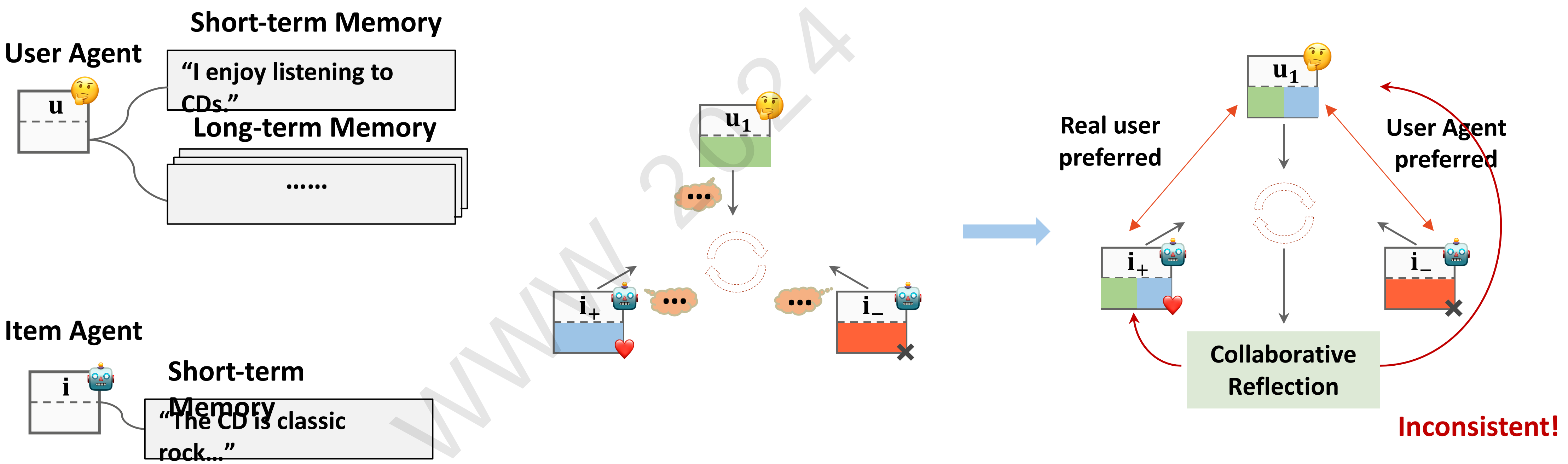
	Random	Pop	MF	MF+Full	MF+Human
Average Rank	4.72	3.22	2.61	2.50	**1.94**

Agents as Users & Items
AgentCF

Agents as Users & Items

❑ **AgentCF: text-based collaborative learning**

- **Key Points:**
  - **Can LLM-powered Agent simulate collaborative signals/user-item interactions?**

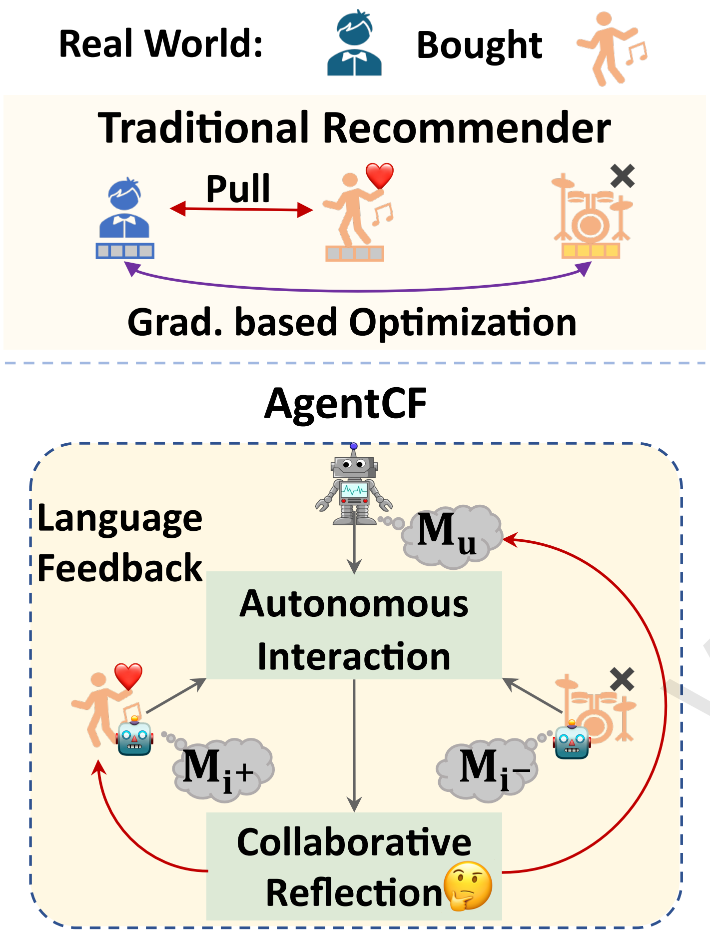


Junjie Zhang et al. AgentCF: Collaborative Learning with Autonomous Language Agents for Recommender Systems. WWW 2024.

AgentCF

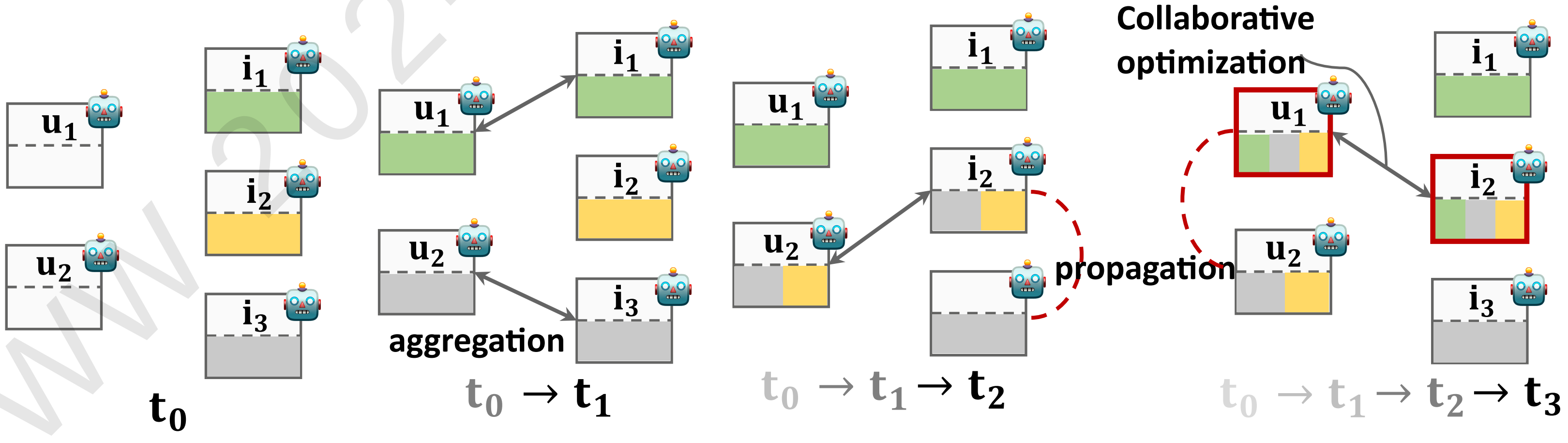
Agents as Users & Items

- **Key Points:**
  - **Can LLM-powered Agent simulate collaborative signals/user-item interactions?**



❑ **AgentCF: text-based collaborative learning**

- **Key idea: Parameter-free text-based collaborative optimization.**



Junjie Zhang et al. AgentCF: Collaborative Learning with Autonomous Language Agents for Recommender Systems. WWW 2024.

Agents as Users & Items

AgentCF

- **Key Observations:**
  - **Agents are capable of simulating user-item interactions.**

Method	CDs_{sparse}			CDs_{dense}			$Office_{sparse}$			$Office_{dense}$		
	N@1	N@5	N@10	N@1	N@5	N@10	N@1	N@5	N@10	N@1	N@5	N@10
BPR_{full}	0.1900	0.4902	0.5619	0.3900	0.6784	0.7089	0.1600	0.3548	0.4983	0.5600	0.7218	0.7625
$SASRec_{full}$	0.3300	0.5680	0.6381	0.5800	0.7618	0.7925	0.2500	0.4106	0.5467	0.4700	0.6226	0.6959
BPR_{sample}	0.1300	0.3597	0.4907	0.1300	0.3485	0.4812	0.0100	0.2709	0.4118	0.1200	0.2705	0.4576
$SASRec_{sample}$	<u>0.1900</u>	0.3948	<u>0.5308</u>	0.1300	0.3151	0.4676	0.0700	0.2775	0.4437	**0.3600**	**0.5027**	**0.6137**

Agents can faithfully simulate user-item interactions.

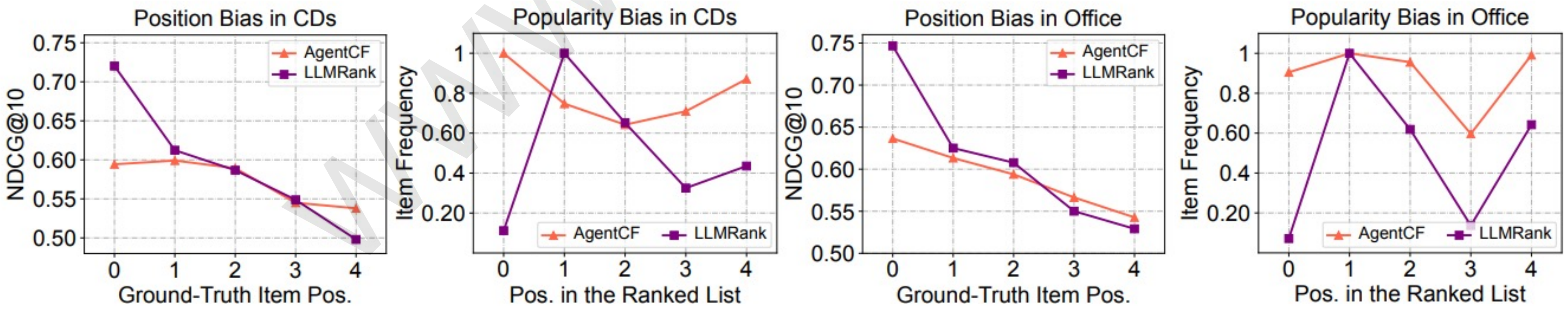
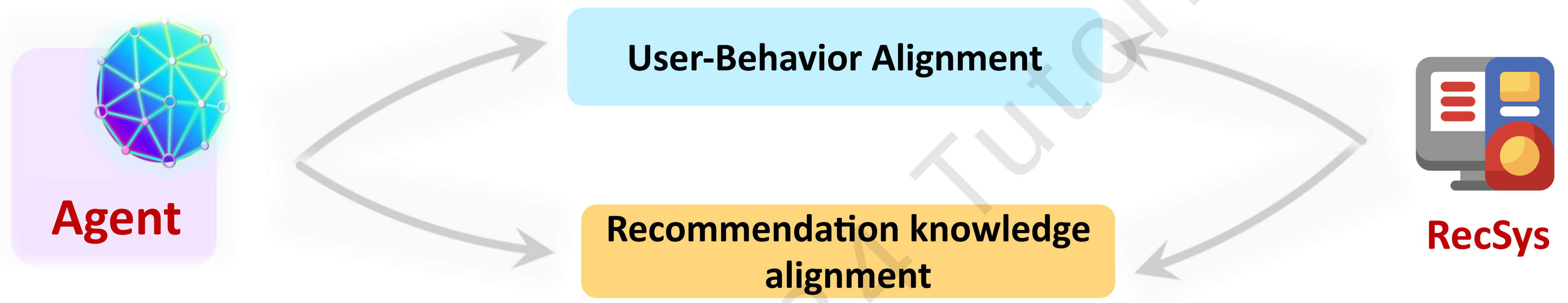


Figure 2: Analysis of whether our approach can simulate personalized agents to mitigate position bias and popularity bias.

LLM-powered Agents in Recommendation



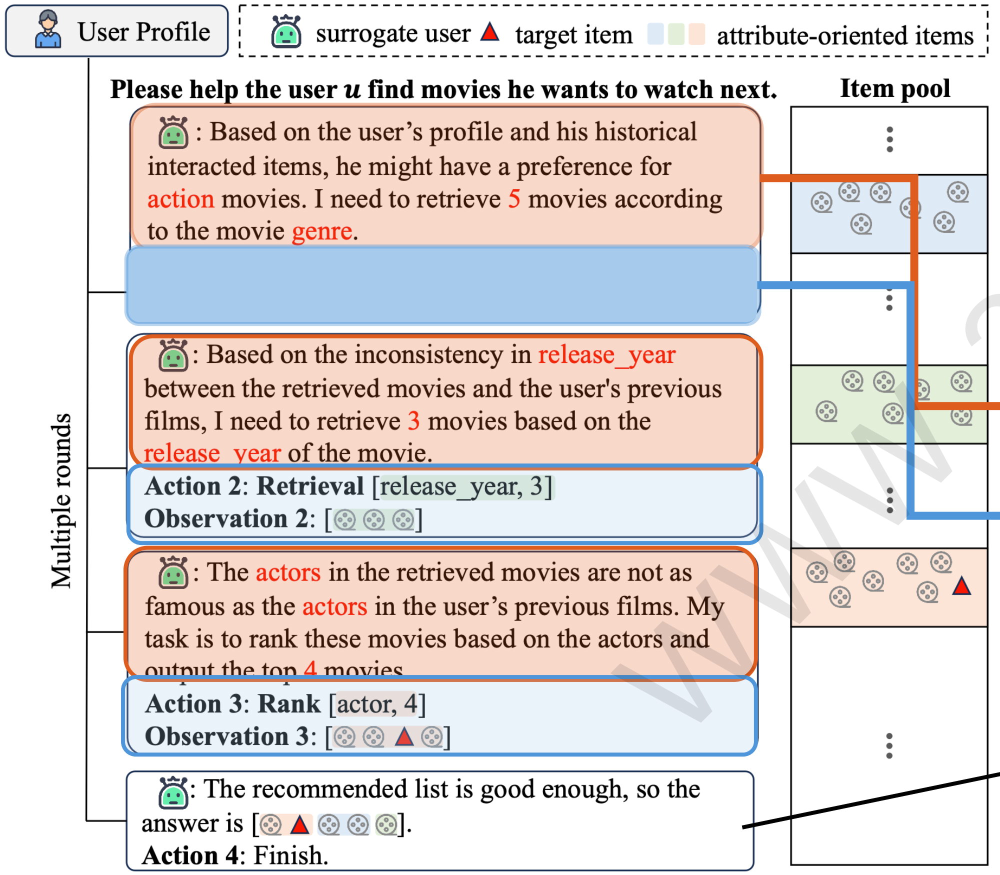
- LLM-empowered have potentials to solve long-standing problems in recommendation
  - Can an LLM-powered Agent faithfully simulate users?
    - Agent4Rec, UGen, AgentCF, RecAgent
  - Can an LLM-powered Agent be a better recommender with recommendation-specific knowledge?

Agent as Recommender
ToolRec

Agent as Recommender

❑ **ToolRec: Tool-enhanced LLM-based recommender**

- **Key Points:**
  - **Can Agents Utilize External Tools to Enhance Recommendations?**



Key Idea:
- Use **LLMs** to understand current contexts and preferences, and apply **attribute-oriented tools** to find suitable items.

Two stages:
- **Learning Preferences**: LLM-based surrogate user learns user preferences and makes decisions
- **Exploration of Items**: uses attribute-oriented tools to explore a wide range of items

❖ Process finishes when the LLM-based surrogate user is satisfied with the item list

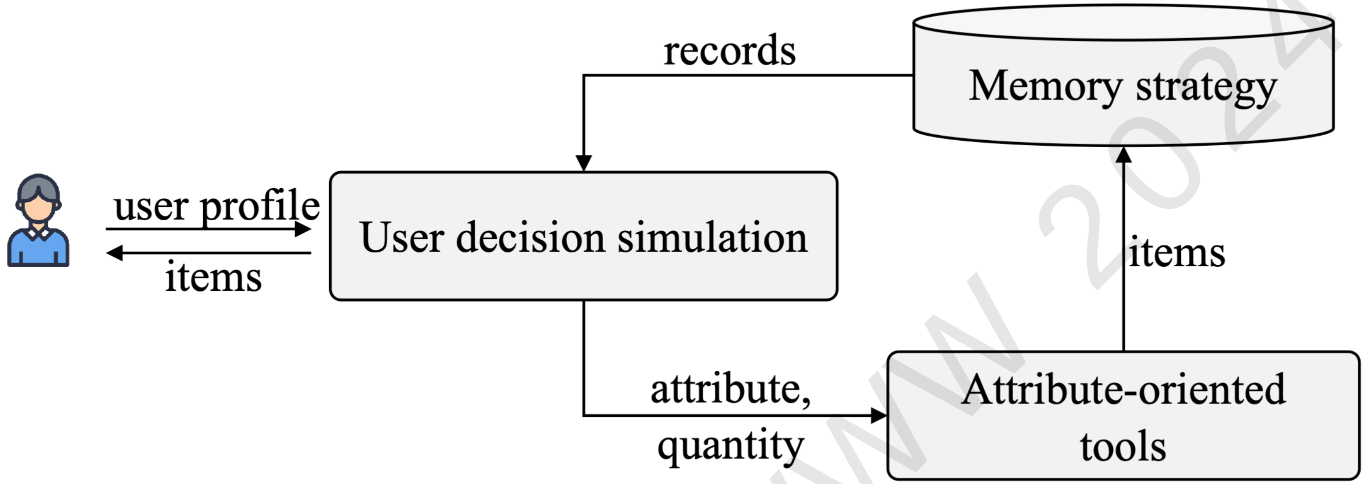
Yuyue Zhao et al. Let Me Do It For You: Towards LLM Empowered Recommendation via Tool Learning. SIGIR 2024.

Agent as Recommender
ToolRec

Agent as Recommender

❑ **ToolRec: Tool-enhanced LLM-based recommender**

- **Key Points:**
  - **Can Agents Utilize External Tools to Enhance Recommendations?**



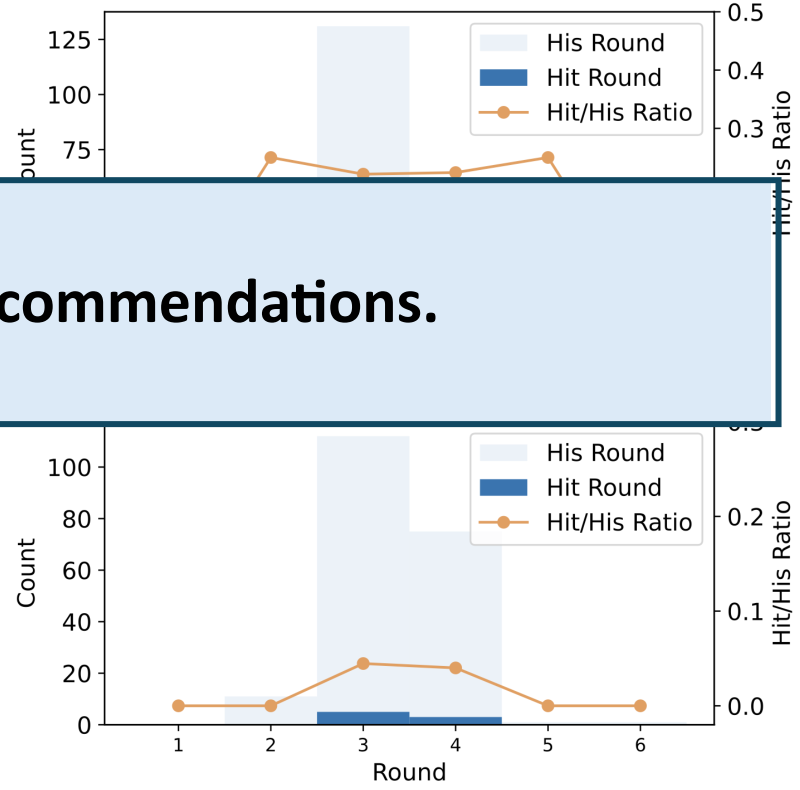
- **LLMs as the central controller, simulating the user decision.**
- **Attribute-oriented Tools: rank tools & retrieval tools.**
- **Memory strategy can ensure the correctness of generated items and cataloging candidate items.**

Yuyue Zhao et al. Let Me Do It For You: Towards LLM Empowered Recommendation via Tool Learning. SIGIR 2024.

Agent as Recommender
ToolRec

- **Key Observations:**
  - **Benefiting from rank tools and tools, ToolRec excels on the ML-1M and Amazon-Book datasets compared to baseline recommenders, demonstrating that it can better align with the users'**

	ML-1M		Amazon-Book		Yelp2018	
	Recall	NDCG	Recall	NDCG	Recall	NDCG
SASRec	0.203±0.047	0.1017±0.016	0.047±0.015	0.0205±0.006	0.030±0.005	0.0165±0.006
ToolRec	**0.215**±0.044	**0.1171**±0.018	**0.053**±0.013	**0.0259**±0.005	0.028±0.003	0.0159±0.001
ToolRec$_B$	0.185±0.018	0.0895±0.002	0.043±0.013	0.0223±0.008	0.025±0.005	0.0136±0.009
Improvement	3.36%	15.10%	14.28%	5.14%	−29.16%	−27.32%



(b) Amazon-Book.

Agents Utilizing External Tools can Enhance Recommendations.

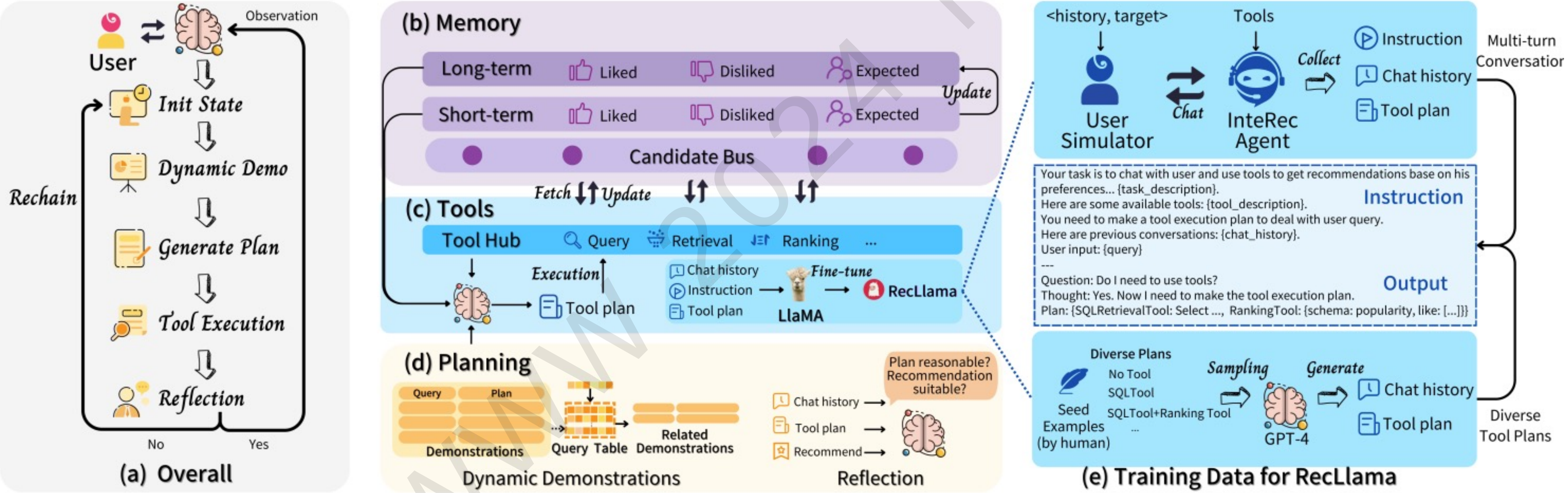
- **ToolRec shows subpar performance on the Yelp2018 dataset - local (niche) businesses.**
- **Most processes conclude in three or four rounds, indicating that the LLM can understand user preferences after a few iterations.**

Yuyue Zhao et al. Let Me Do It For You: Towards LLM Empowered Recommendation via Tool Learning. SIGIR 2024.

Agent as Recommender

❑ **InteRecAgent: Interactive Recommender.**

- **Key Points:**
  - **Agents can create a versatile and interactive recommender system.**



- **InteRecAgent enables traditional recommender systems, such as those ID-based matrix factorization models, to become interactive systems with a natural language interface.**

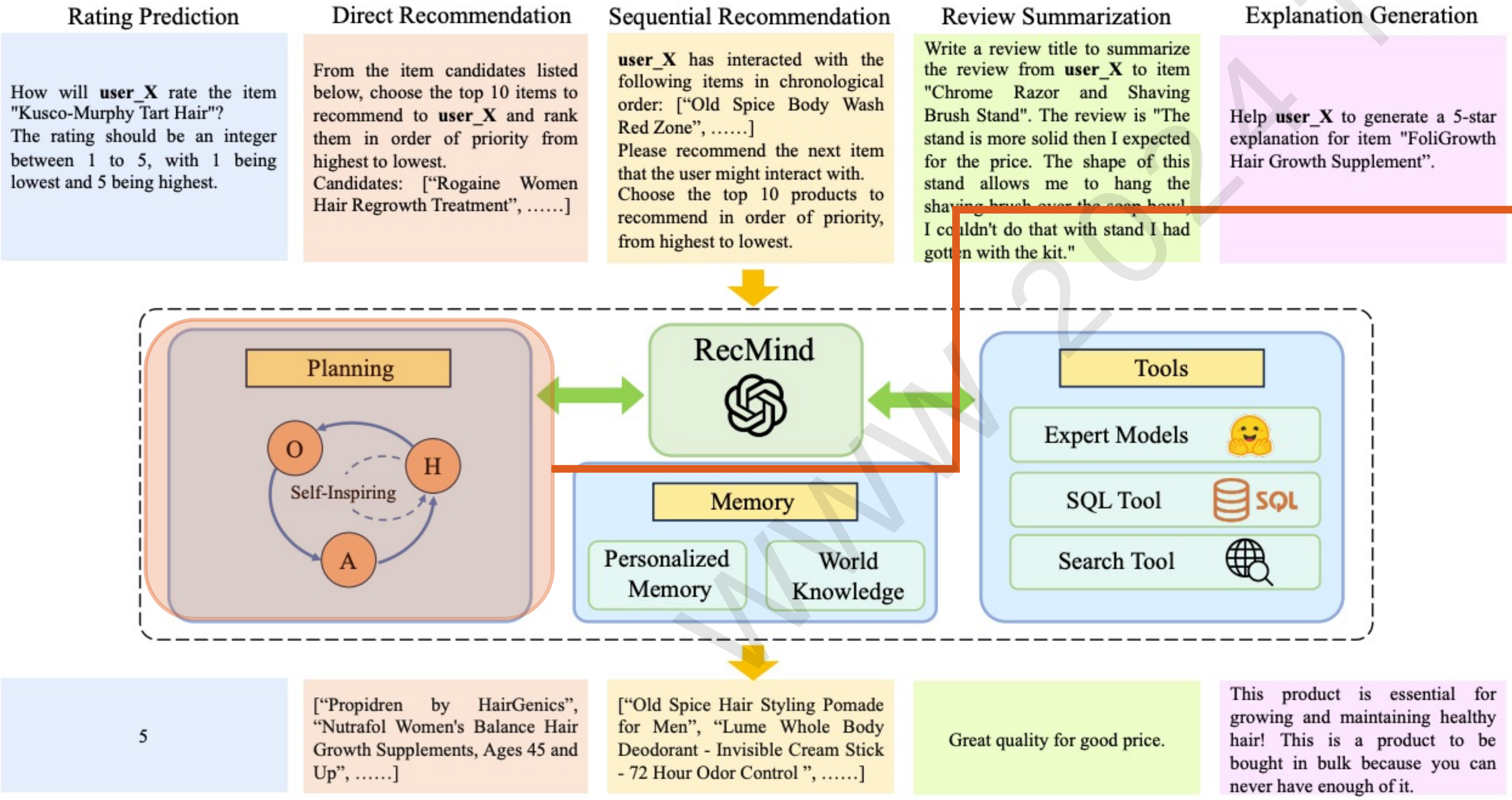
Xu Huang et al. Recommender AI Agent: Integrating Large Language Models for Interactive Recommendations. Arxiv 2023..

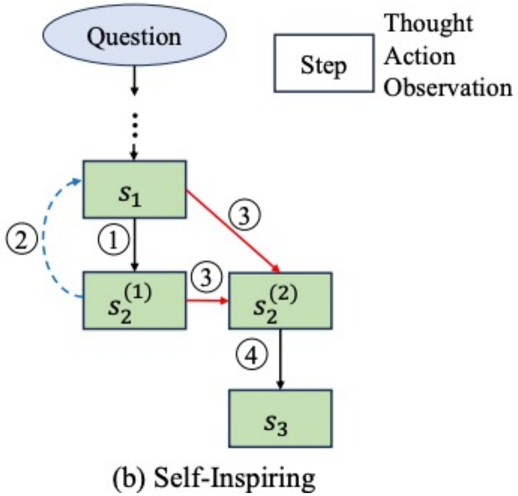
Agent as Recommender
RecMind

Agent as Recommender

- **Key Points:**
  - **Can Agents with self-inspiring planning Enhance Recommendations?**

❑ **RecMind: Recommender agent with Self-Inspiring planning ability**





(b) Self-Inspiring

- **Self-inspires:**
- **At each intermediate planning step, the agent "self-inspires" to consider all previously explored paths for the next planning, both generating alternative thoughts and backtracking.**

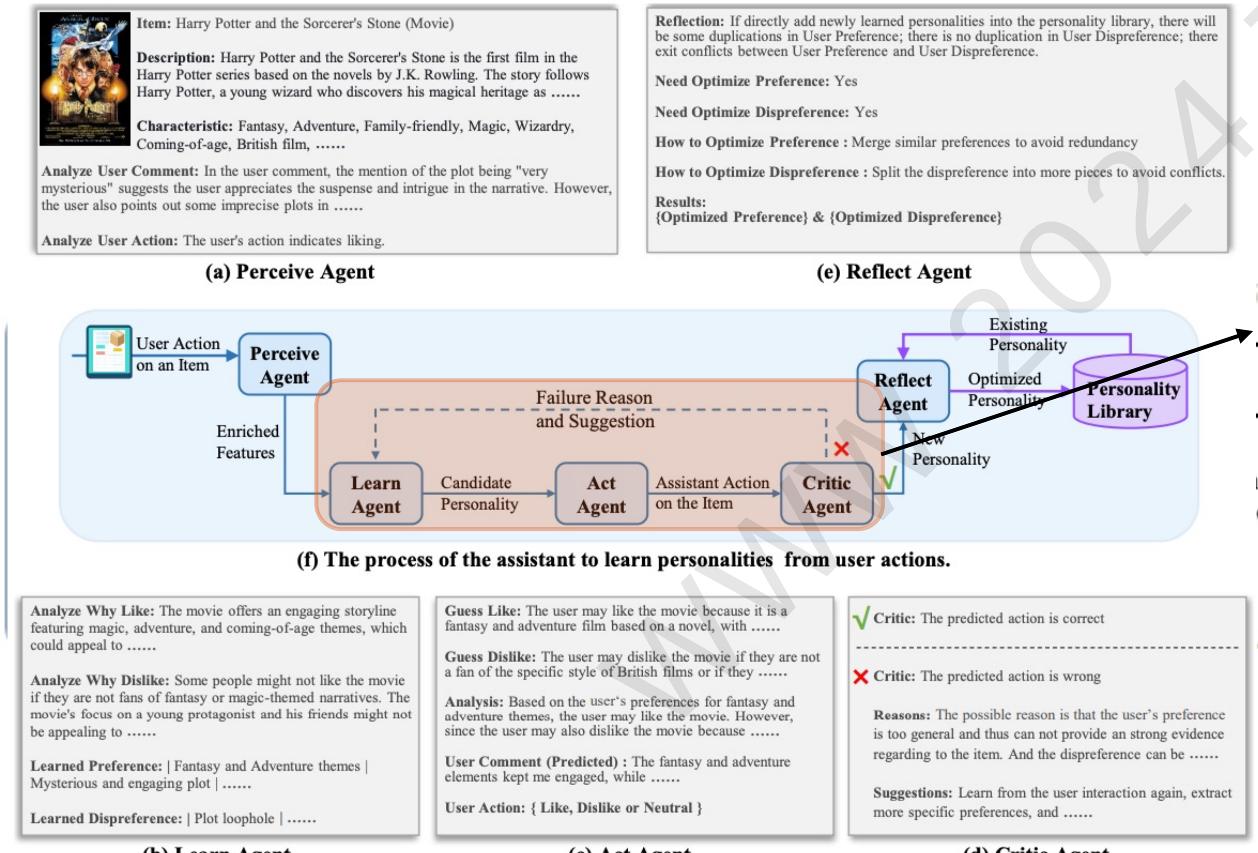
Yancheng Wang et al. "RecMind: Large Language Model Powered Agent For Recommendation.NAACL 2024.

Agent as Recommendation Assistant
RAH

Agent as Rec Assistant

RAH: Reflection-enhanced user alignment for Rec assistant

- **Key Points:**
  - Can Agents with Learn-Act-Critic loop comprehend a user's personality from their behaviors?



(a) Perceive Agent (e) Reflect Agent

(f) The process of the assistant to learn personalities from user actions.

(b) Learn Agent (c) Act Agent (d) Critic Agent

❖ **Learn-Act-Critic Loop:**
- Learn Agent collaborates with the Act and Critic Agents in an iterative process to grasp the user's personality.
- Upon receiving user feedback, Learn Agent extracts an initial personality as a candidate.
- Act Agent utilizes this candidate as input to predict the user's actual action.
- The Critic Agent then assesses the accuracy. If incorrect, Learn Agent refines the candidate's personality.

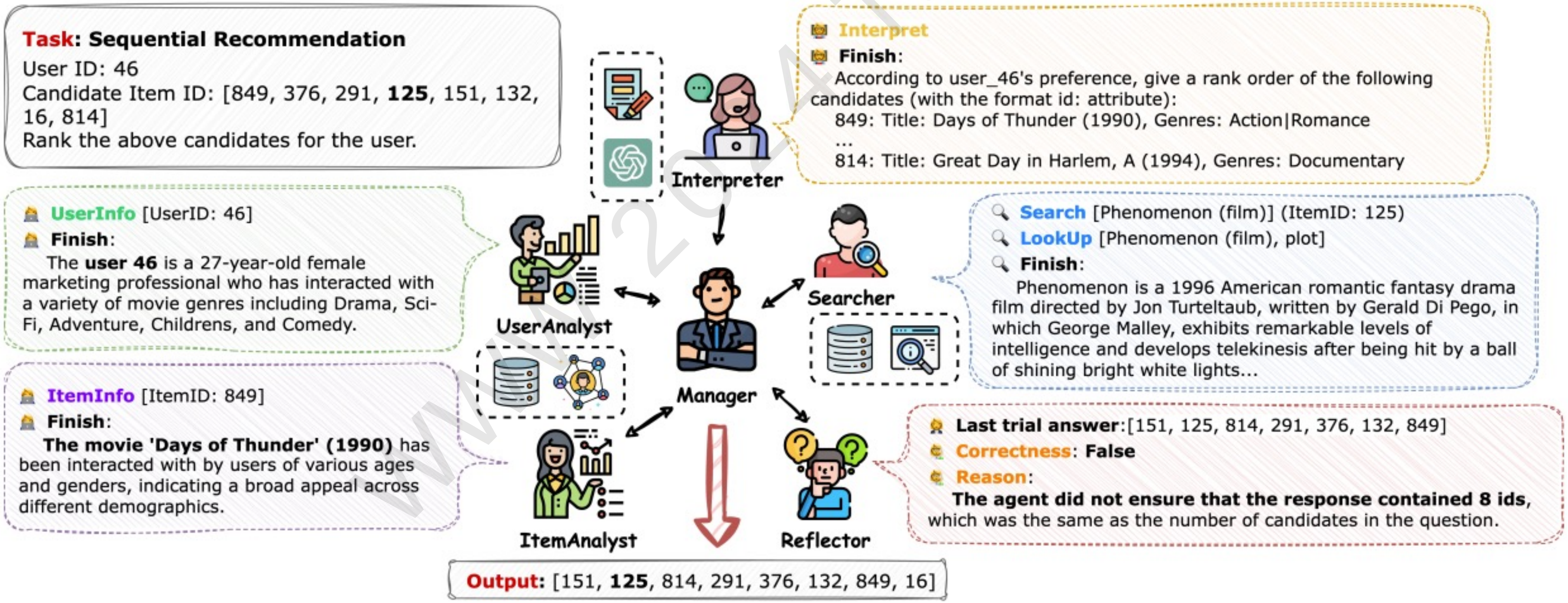
Yubo Shu et al. RAH! RecSys-Assistant-Human: A Human-Centered Recommendation Framework with LLM Agents. Arxiv 2023.

Multi-Agents as Recommender
MACRec

Multi-Agent as Recommender

❑ MACRec: enhance RecSys through multi-agent collaboration

- **Key Points:**
  - Multi-agents with different roles work collaboratively to tackle a specific recommendation task.



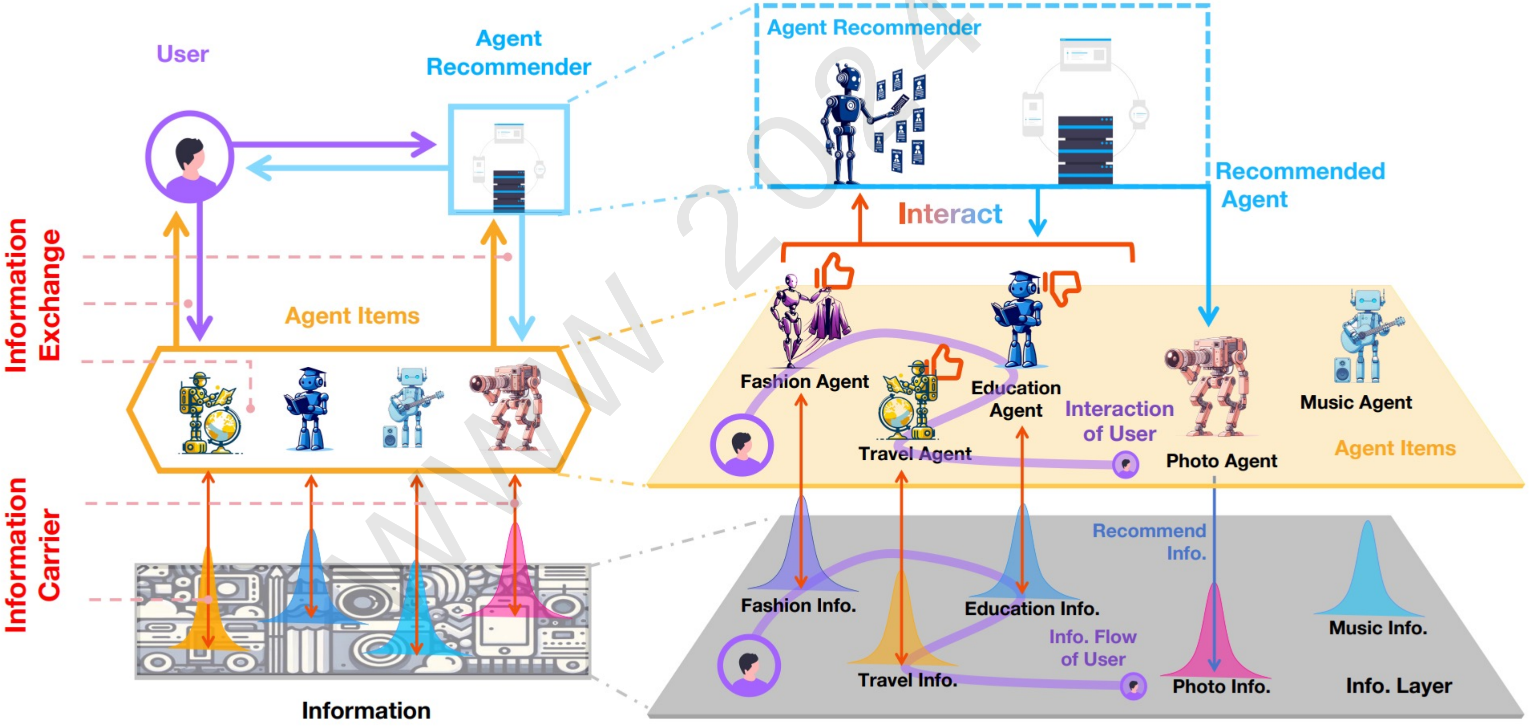
Zhefan Wang et al. Multi-Agent Collaboration Framework for Recommender Systems. Arxiv 2024.

Agent Recommender for Agent Platform
Rec4Agentverse

Agent Recommender

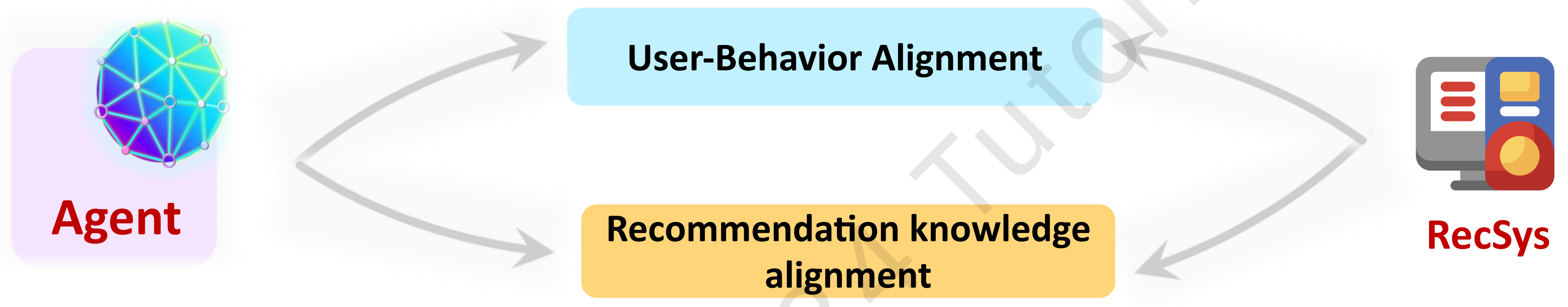
❑ **Rec4Agentverse: Agent recommender for Agent platform**

- **Key Points:**
  - **Treating LLM-based Agents in Agent platform as items in the recommender system.**
  - **Agent Recommender is employed to recommend personalized Agent Items for each user.**



Jizhi Zhange t al. Prospect Personalized Recommendation on Large Language Model-based Agent Platform. Arxiv 2024.

LLM-powered Agents in Recommendation



- LLM-empowered have potentials to solve long-standing problems in recommendation
  - Can an LLM-powered Agent faithfully simulate users?
    - Agent4Rec, UGen, AgentCF, RecAgent
  - Can an LLM-powered Agent be a better recommender with recommendation-specific knowledge?
    - ToolRec, InteRecAgent, RecMind, RAH, MACRec, Rec4Agentverse

Outline

- Background of LLM Agent-based Simulation (25minutes) 13:30-13:55
- Online behavior simulation with LLM Agents (65 minutes) 13:55-15:00
- Break (15 minutes) 15:00-15:15
- Social and economic simulation with LLM agents (50 minutes) 15:15-16:05
- City system simulation with LLM agents (45minutes) 16:05-16:50
- Open discussions (10minutes) 16:50-17:00

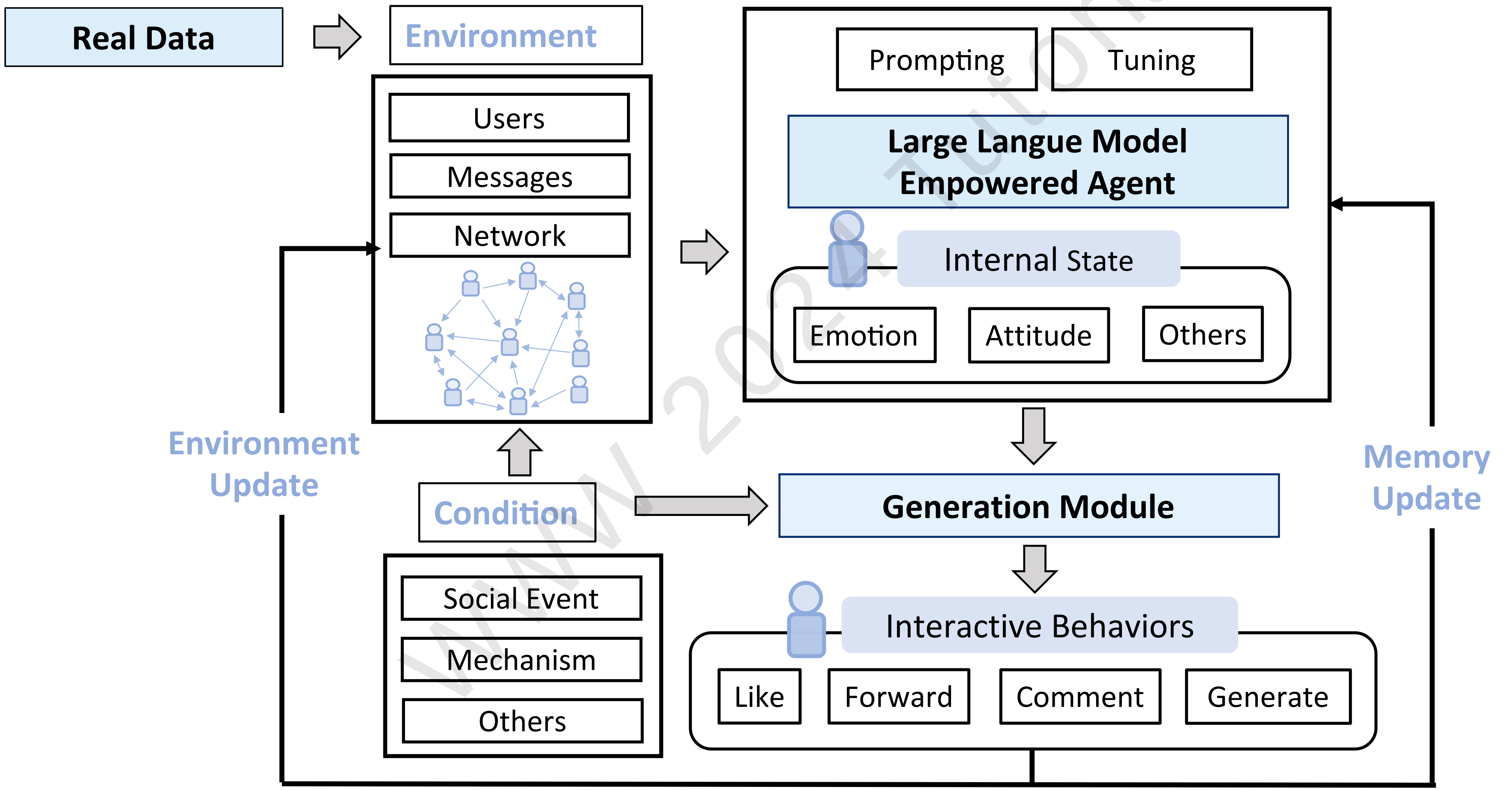
Outline

- Background of LLM Agent-based Simulation
- Online behavior simulation with LLM Agents
- Social and economic simulation with LLM agents
- City system simulation with LLM agents
- Open discussions

Outline

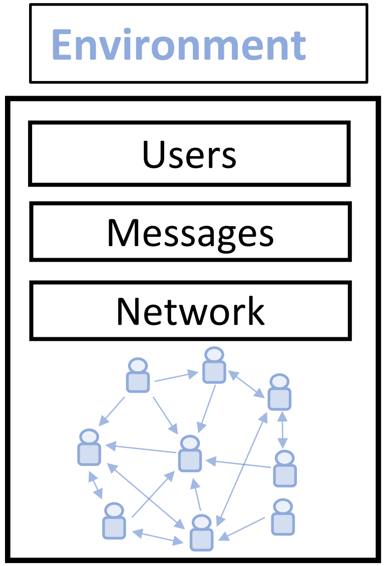
- Background of LLM Agent-based Simulation
- Online behavior simulation with LLM Agents
- Social and economic simulation with LLM agents
  - Social Simulation System (S3)
  - Attitude simulation
  - Emotion simulation
- City system simulation with LLM agents
- Open discussions

S³ （Social-network Simulation System）



S^3 (Social-network Simulation System)

1. User Demographics Prediction for Environment Construction



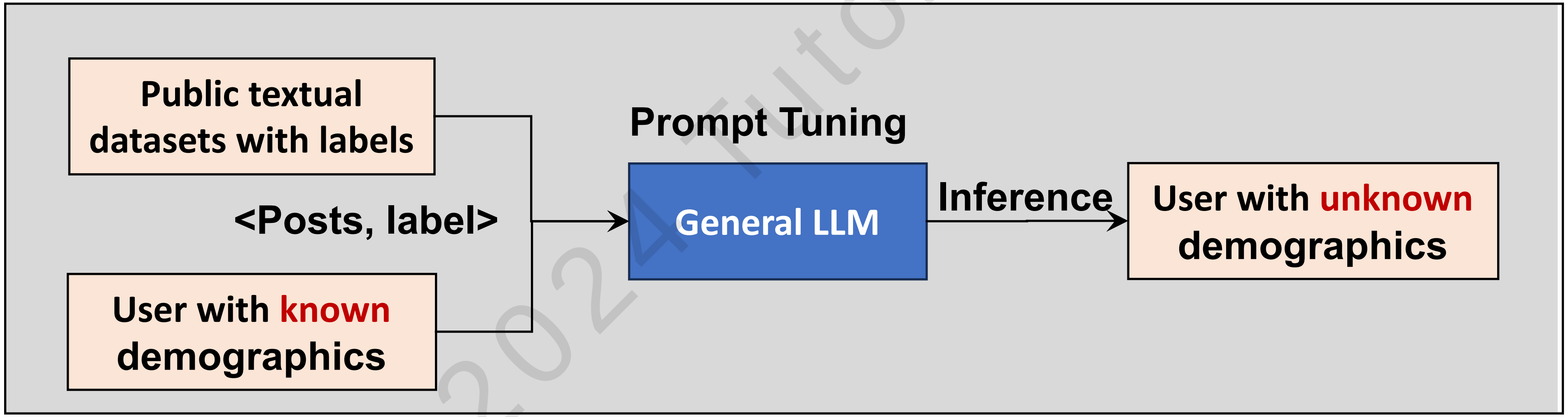


Table 4: Prediction performance of gender and age.

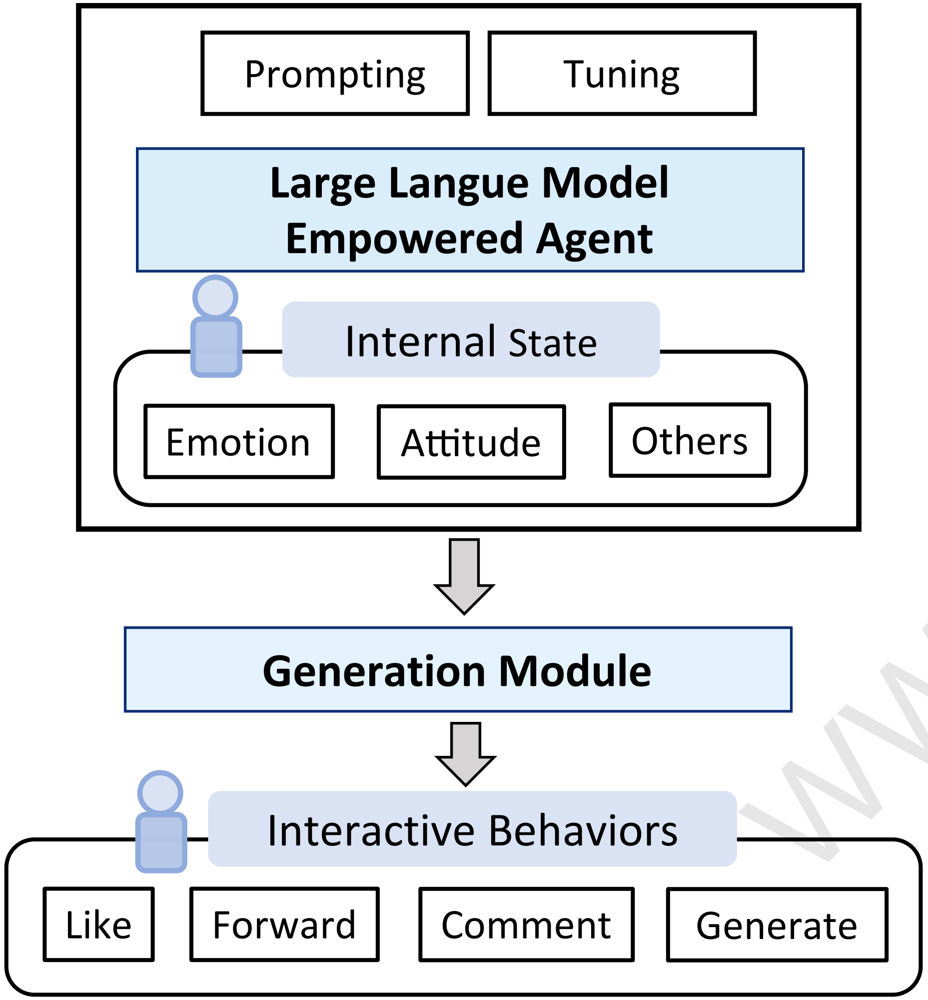
Demographic	Performance		
Gender	**Acc**	**F1**	**AUC**
	0.710	0.667	0.708
Age	**MSE**	**MAE**	**Avg % Error**
	128.0	7.53	21.50

Table 5: Ten occupations.

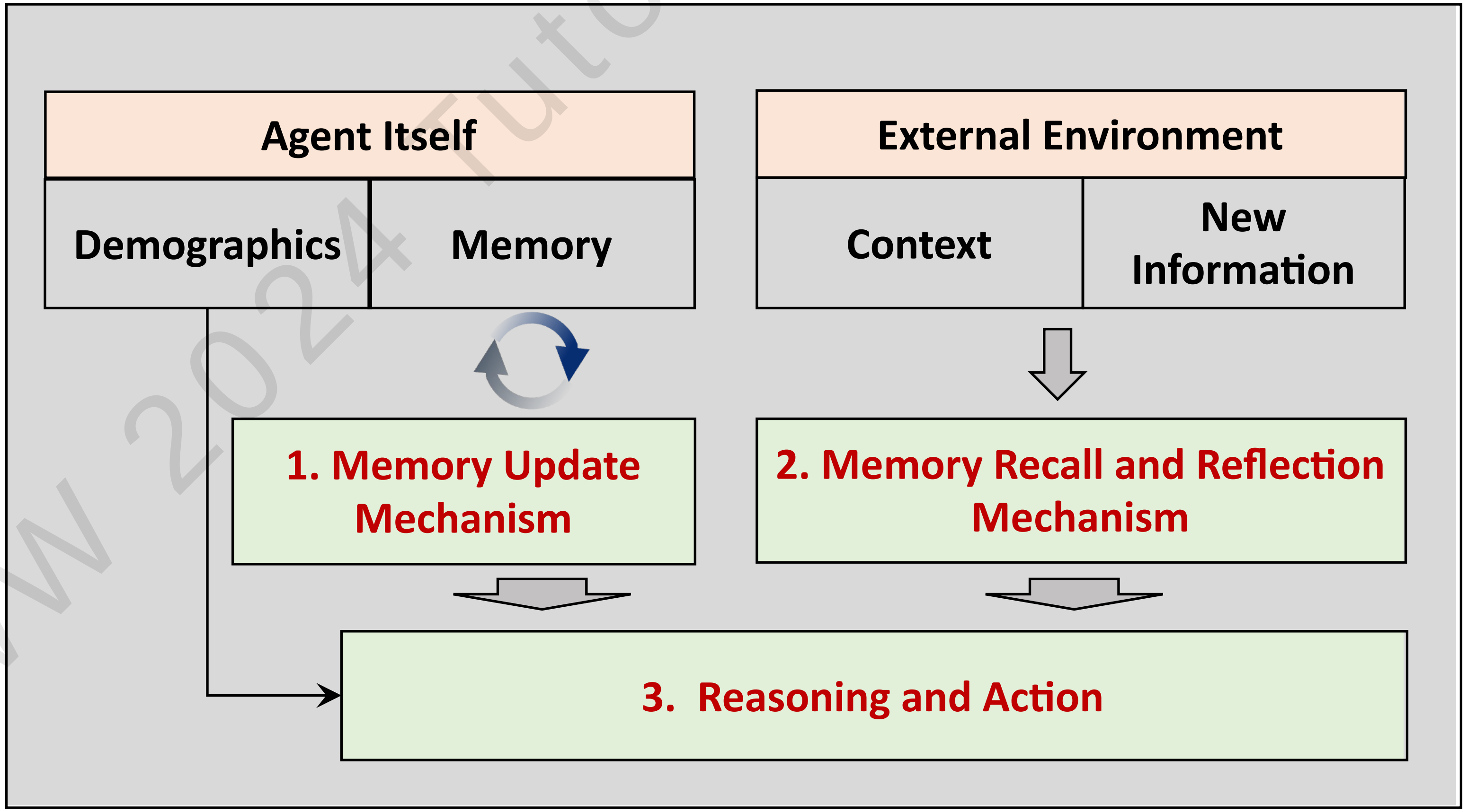
1	Education Practitioner
2	Administrative Manager / Officer
3	Unemployed / Student
4	Engineer
5	Labor Technician / Worker
6	Logistics Practitioner
7	Medical Personnel
8	Financial Practitioner
9	Media Personnel
10	Entertainment and Arts Practitioner

We consider three demographics: **age, gender, occupation**

S^3 (Social-network Simulation System)



2. Agent-based Simulation



Think like human: 1. Memory 2. Contextual information 3. Reasoning

Individual-level Simulation Ability

Scenario	Prediction Task	Accuracy	AUC	F1-Score
Gender Discrimination	Emotion Level	71.8%	—	—
	Event Propagation	66.2%	0.662	0.667
Nuclear Energy	Initial Attitude	74.3%	0.727	0.834
	Attitude Change	83.9%	0.865	0.857
	Event Propagation	69.5%	0.681	0.758

Scenario	Perplexity	Cos. Sim.
Gender Discrimination	19.289	0.723
Nuclear Energy	16.145	0.741

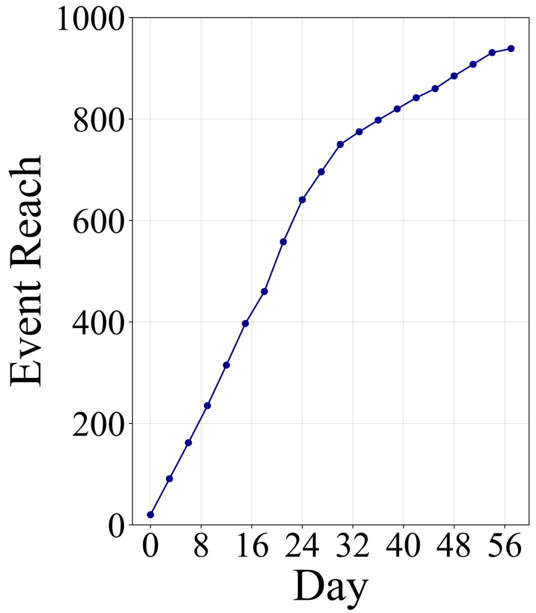
LLM-empowered simulation can well predict

- **Emotion** (Calm, Moderate, Intense)
- **Attitude** (Support nuclear energy or not)
- **Behavior** (Forward or post relevant content)

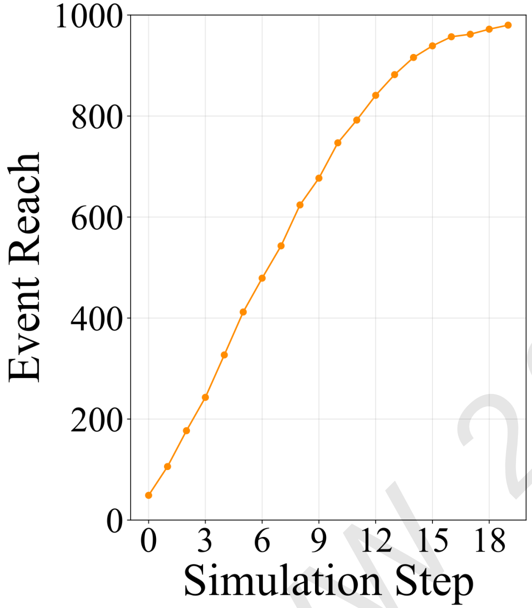
Generated content has acceptable quality

Population-level Simulation Ability

Information Propagation

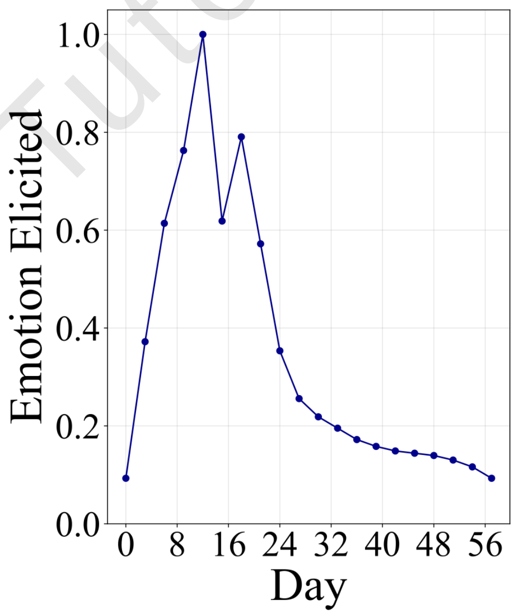


Real

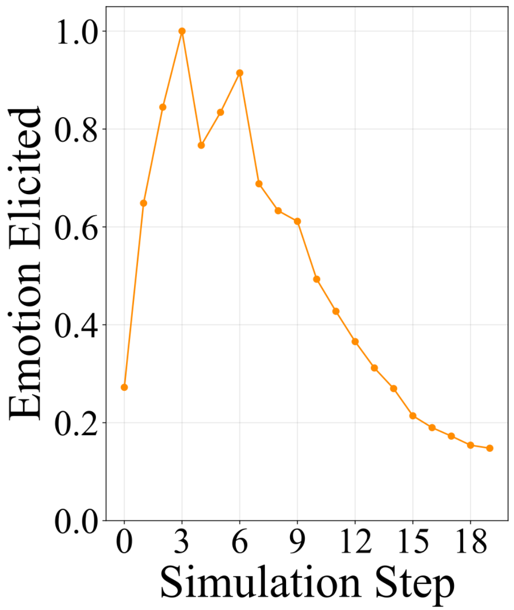


Simulation

Emotion Propagation



Real



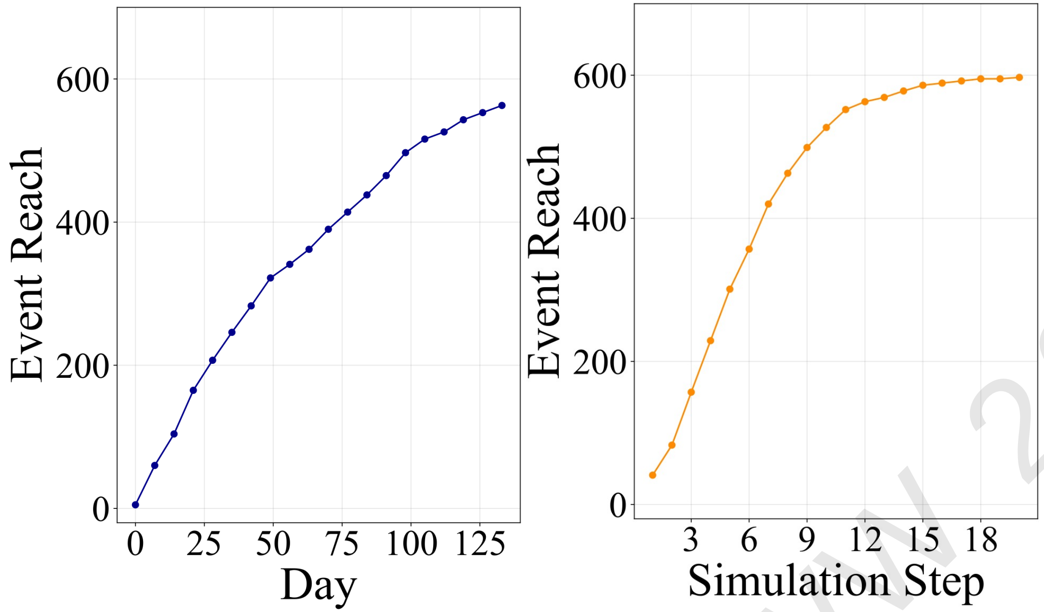
Simulation

(If the news slowly propagate to a large community, there will be 2nd peak)

Our simulation can well predict the trend in the Gender Discrimination case

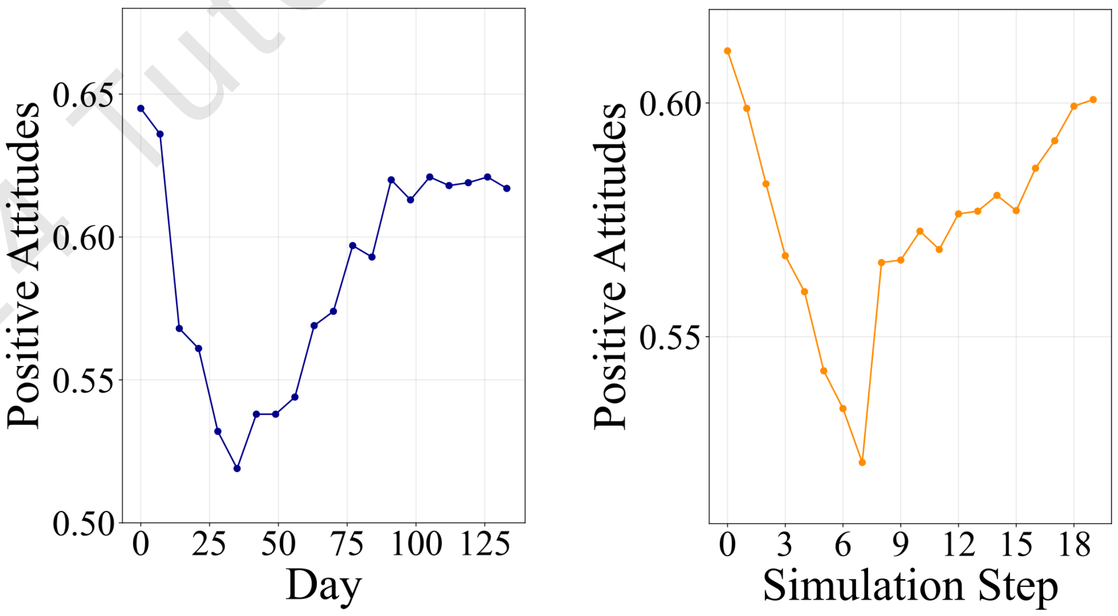
Population-level Simulation Ability

Information Propagation



Real **Simulation**

Attitude Propagation

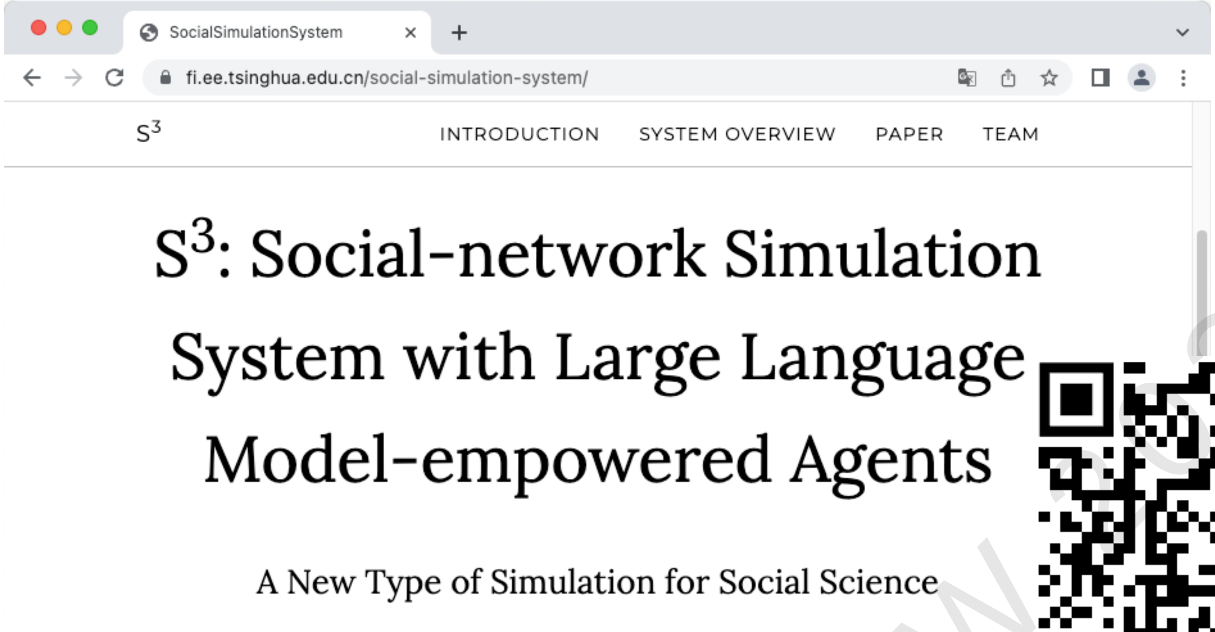


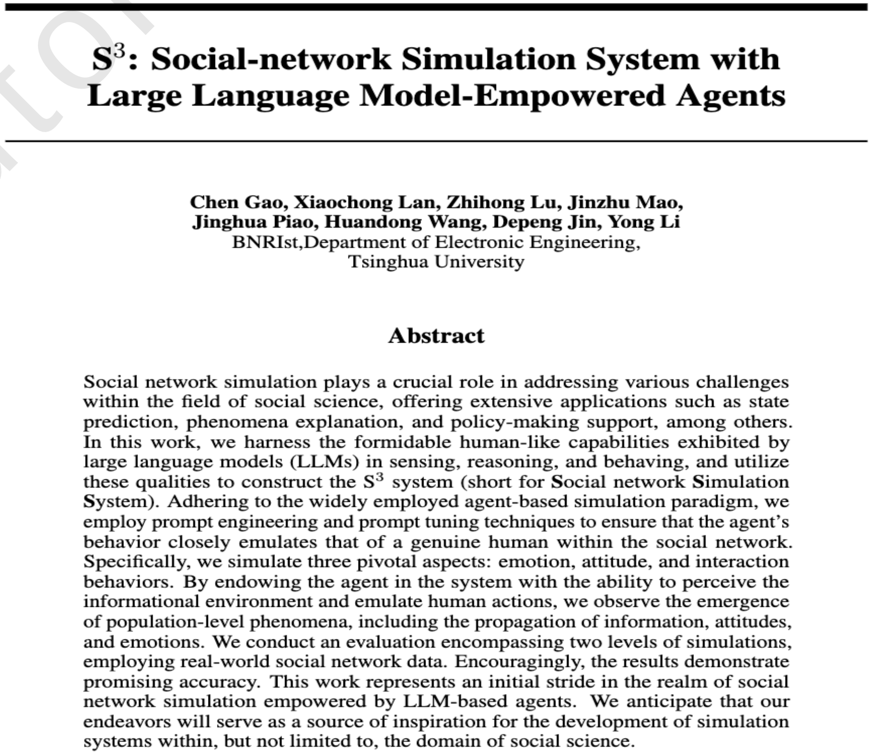
Real **Simulation**

(Public attitude tends to be normal after a period)

Our simulation can well predict the trend in the Nuclear Energy case

Official website and the preprint paper





S^3: Social-network Simulation System with Large Language Model-Empowered Agents

Chen Gao, Xiaochong Lan, Zhihong Lu, Jinzhu Mao, Jinghua Piao, Huandong Wang, Depeng Jin, Yong Li
BNRIst,Department of Electronic Engineering, Tsinghua University

Abstract

Social network simulation plays a crucial role in addressing various challenges within the field of social science, offering extensive applications such as state prediction, phenomena explanation, and policy-making support, among others. In this work, we harness the formidable human-like capabilities exhibited by large language models (LLMs) in sensing, reasoning, and behaving, and utilize these qualities to construct the S^3 system (short for **S**ocial network **S**imulation **S**ystem). Adhering to the widely employed agent-based simulation paradigm, we employ prompt engineering and prompt tuning techniques to ensure that the agent's behavior closely emulates that of a genuine human within the social network. Specifically, we simulate three pivotal aspects: emotion, attitude, and interaction behaviors. By endowing the agent in the system with the ability to perceive the informational environment and emulate human actions, we observe the emergence of population-level phenomena, including the propagation of information, attitudes, and emotions. We conduct an evaluation encompassing two levels of simulations, employing real-world social network data. Encouragingly, the results demonstrate promising accuracy. This work represents an initial stride in the realm of social network simulation empowered by LLM-based agents. We anticipate that our endeavors will serve as a source of inspiration for the development of simulation systems within, but not limited to, the domain of social science.

Official Website:
https://fi.ee.tsinghua.edu.cn/social-simulation-system

Paper:
https://fi.ee.tsinghua.edu.cn/social-simulation-system/paper

S^3: Social-network Simulation System with Large Language Model-empowered Agents. *C. Gao, X. Lan, Z. Lu, J. Mao, J. Piao, H. Wang, D. Jin, and **Yong Li**, preprint 2023.*

Outline

- Background of LLM Agent-based Simulation
- Online behavior simulation with LLM Agents
- Social and economic simulation with LLM agents
    - Social Simulation System (S3)
    - Attitude simulation
    - Emotion simulation
- City system simulation with LLM agents
- Open discussions

Stance Detection and Simulation with LLM

LLM is promising on stance detection; **BUT**:
Directly applying large language models to stance detection may yield poor results.

Model	Score on Sem16-CC	Score on Sem16-A
BERT-GCN	35.5	53.6
PT-HCL	38.9	56.5
GPT-3.5	**31.1**	**9.1**

SOTA zero-shot stance detection baselines

To show the performance of directly applying GPT-3.5, we strictly follow the prompt design in [1] and [2].

[1] Zhang B, Ding D, Jing L. How would stance detection techniques evolve after the launch of chatgpt?[J]. arXiv preprint arXiv:2212.14548, 2022.
[2] Ziems C, Held W, Shaikh O, et al. Can Large Language Models Transform Computational Social Science?[J]. arXiv preprint arXiv:2305.03514, 2023.

Stance Simulation with LLM: Challenges

Why directly applying LLMs does not work?
There are challenges to be tackled!

Challenge 1:
Stance detection demands multi-aspect knowledge.

Tweet:
Time to reclaim our nation! No more Republicans! #ByeByeGOP

Target: Donald Trump **Stance:** Against

Required knowledge:

1. On social media, the hashtag #ByeByeGOP expresses disagreement with the Republican Party.
2. Donald Trump is a Republican.

Stance Simulation with LLM: Challenges

Why directly applying LLMs does not work?
There are challenges to be tackled!

Challenge 2: Stance detection necessitates advanced reasoning.	
Tweet: It's a problem when explaining feminism, even in a calm and complex level, cannot be understood.	
Target: Feminism Movement	**Stance:** Favor
Logical chain: The lack of understanding of feminism is problematic.→ Feminism should be understood and accepted → Support feminism	

We need to make some specific and proper design.

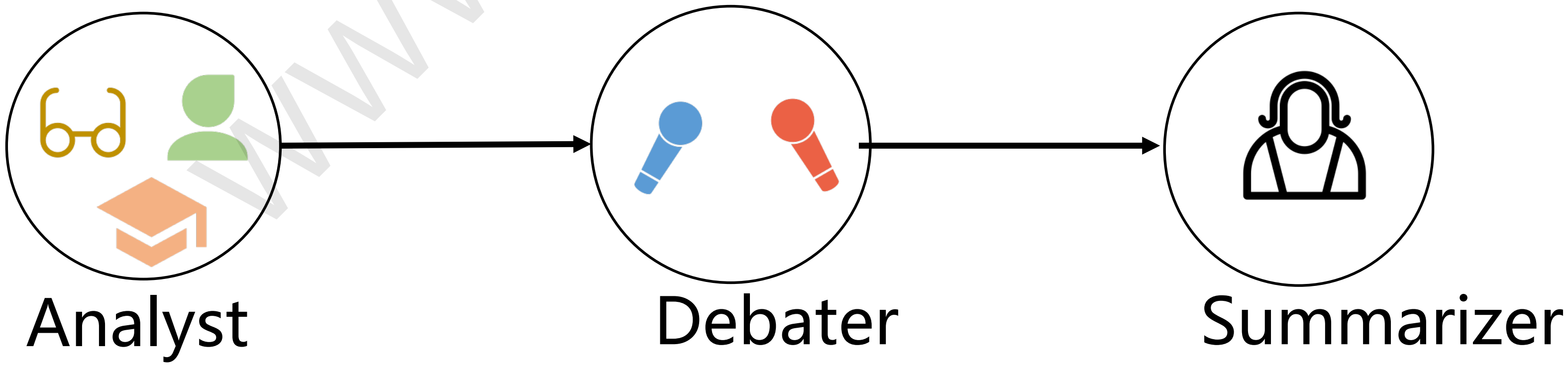
Stance Simulation with LLM: Method

Analyst-Debater-Summarizer Framework

- Stance detection demands multi-aspect knowledge.(Challenge 1)
  →Step1: Analysts analyze text from various perspectives.

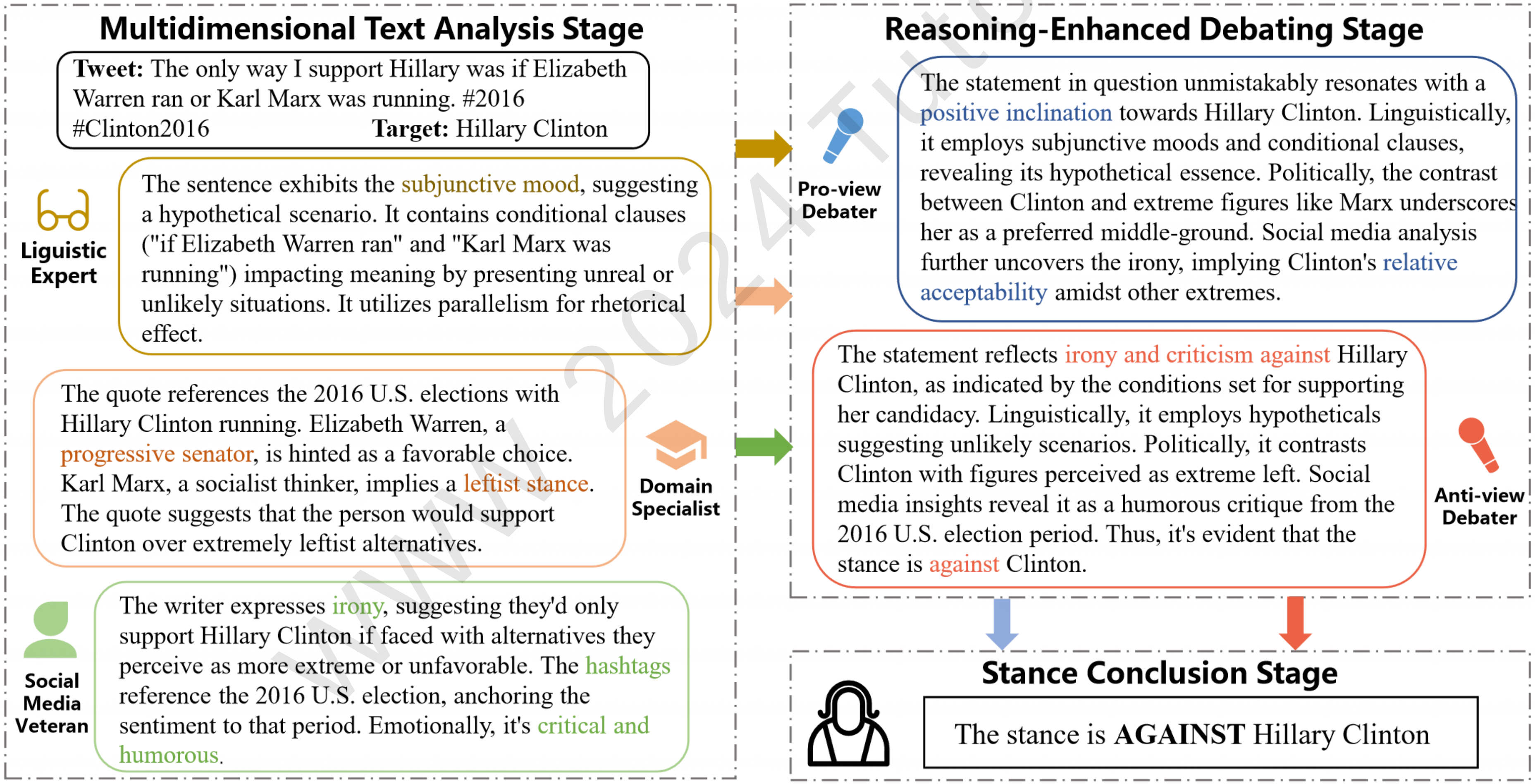
- Stance detection necessitates advanced reasoning. (Challenge 2)
  →Step 2: Debaters debate for each potential stance category.

A summarizer get conclusion from the debater's debate, determining the final result.



Stance Simulation with LLM: Method

Overall Framework



Stance Simulation with LLM: Method

Multidimensional Text Analysis Stage

Input: A text with a stance.

Output: The individual analyses of the text by three agents:
the linguistic expert, the domain specialist, and the social media veteran.

- **Linguistic Expert**
  - Dissects the text from a linguistic standpoint.
  - Focus on: grammatical structure; tense and inflection; rhetorical devices; lexical choices, etc.
- **Domain Specialist**
  - Explains domain relevant knowledge.
  - Focus on: characters; events; organizations; parties; religions, etc.
- **Social Media Veteran**
  - Delves into the nuances of social media expression.
  - Focus on: hashtags; Internet slang and colloquialisms; emotional tone.

Stance Simulation with LLM: Method

Reasoning-Enhanced Debating Stage

Agents: Advocators for each possible stance

Input: A text with a stance. The analyses of the text by the linguistic expert, the domain specialist, and the social media veteran.

Output: The debate from each agent for the stance they support, including the evidence it chooses and its logical chain.

Stance Conclusion Stage

Agent: A Judge

Input: A text with an embedded stance. Arguments from each agent, including evidence and their logical reasoning.

Output: The identified stance of the text.

Stance Simulation with LLM: Experiments

Accuracy

Model	SEM16(%) DT	HC	FM	LA	A	CC	P-Stance(%) Trump	Biden	Sanders	VAST(%) All
TOAD	49.5	51.2	54.1	46.2	46.1	30.9	53.0	68.4	62.9	41.0
TGA Net	40.7	49.3	46.6	45.2	52.7	36.6	-	-	-	65.7
BERT-GCN	42.3	50.0	44.3	44.2	53.6	35.5	-	-	-	68.6
PT-HCL	50.1	54.5	54.6	50.9	56.5	38.9	-	-	-	71.6
JointCL	50.5	54.8	53.8	49.5	54.5	39.7	62.0	59.0	73.0	72.3
GPT-3.5	62.5	68.7	44.7	51.5	9.1	31.1	62.9	80.0	71.5	62.3
GPT-3.5+COT	63.3	70.9	47.7	53.4	13.3	34.0	63.9	81.2	73.2	68.9
COLA(ours)	**68.5**	**81.7***	**63.4***	**71.0***	**70.8***	**65.5***	**86.6***	**84.0**	**79.7***	**73.0**

Category	Model	SEM16(%) DT	HC	FM	LA	A	CC	P-Stance(%) Trump	Biden	Sanders
In-target Labeled Data Dependent Methods	BiCond	59.0	56.1	52.9	61.2	55.3	35.6	73.0	69.4	64.6
	BERT	57.9	61.3	59.0	63.1	60.7	38.8	67.7	73.1	68.2
	CrossNet	60.2	60.2	55.7	61.3	56.4	40.1	58.0	65.0	53.0
	ATT-LSTM	55.3	59.8	55.3	62.6	55.9	39.2	-	-	-
	ASGCN	58.7	61.0	58.7	63.2	59.5	40.6	77.0	78.4	70.8
	TPDG	63.0	73.4	**67.3**	**74.7**	64.7	42.3	76.8	78.1	71.0
Zero-shot Method	COLA(ours)	**68.5**	**81.7***	63.4	71.0	**70.8**	**67.5***	**86.6***	**84.0***	**79.7***

- Our method achieves the current **SOTA** level in most scenarios **without the need for additional training**.

Stance Simulation with LLM: Experiments

Effectiveness

Model	SEM16(%)					
	DT	HC	FM	LA	A	CC
COLA	68.5	81.7	63.4	71.0	70.8	67.5
w/o LE	64.3	80.5	63.3	68.9	69.9	65.5
w/o DS	66.5	79.2	64.4	67.9	70.7	65.4
w/o SMV	64.8	76.8	64.5	64.1	67.7	63.5
w/o TAS	64.4	77.2	65.7	63.8	67.0	62.3
w/o DS	64.7	74.9	62.5	39.2	59.6	53.4

- The removal of any expert agent results in a certain degree of performance degradation.
- The removal of both stages leads to a significant performance drop.

Stance Simulation with LLM: Experiments

Explainability: Can COLA provide reasonable explanations for its decisions?

Tweet: @GovtsTheProblem This is what I see: Make way 4 ur queen peasants! Don't touch or talk 2 her U filth! #NoHillary2016 #Benghazi
Target: Hillary Clinton

Detected Stance: Against
Explaination: The use of derogatory language, the mention of the Benghazi incident in a negative context, and the hashtag #NoHillary-2016 all indicate thje stance against Clinton and a desire for her not to be elected.

Tweet: The ruling by @Scotus is a major setback for @EPA & the environment. #dirtycoal
Target: Climate Change is Real Concern

Detected Stance: Favor
Explaination: The use of the hashtag #dirtycoal conveys a negative stance on coal, associated with environmental harm and climate change. Mentioning the Environmental Protection Agency (@EPA) and describing a ruling as an environmental setback highlights a concern for environmental protection, including climate change mitigation. The emotional tone of frustration and negative environmental hashtags further reinforce this pro-environmental perspective.

- Cases show COLA can provide reasonable and high-quality explanations for its decisions.

Method	SEM16(%)					
	DT	HC	FM	LA	A	CC
GPT-3.5	69.0	75.5	60.8	55.3	10.3	25.2
COLA	**71.2**	75.9	69.1	**71.0**	**62.3**	**64.0**
GPT-3.5+COLA's Explainations	69.4	**77.7**	**70.7**	66.7	61.9	54.5

- The explanations generated COLA can greatly help GPT 3.5 to identify stances.

→ COLA can provide reasonable and high-quality explanations for its decisions.

Stance Simulation with LLM: Experiments

Versatility: Can COLA perform well on other measurement tasks?

Category	Model	Restaurant14(%)		Laptop14(%)		Restaurant15(%)	
		Accuracy	Macro-F1	Accuracy	Macro-F1	Accuracy	Macro-F1
Labeled Data Dependent Methods	DGEDT	**86.3**	80.0	79.8	75.6	84.0	71.0
	dotGCN	86.2	**80.5**	81.0	**78.1**	85.2	72.7
Zero-shot Methods	GPT-3.5-Turbo	74.3	69.6	69.9	61.0	80.4	67.7
	Ours	84.1	77.7	**81.6**	77.0	**85.4**	**74.9**

Model	Accuracy(%)	F1-Score(%)
Hybrid RCNN	74.8	59.6
GPT-3.5 Turbo	67.6	56.0
Ours	**76.5**	**63.9**

- **Sentiment analysis:** Determine the sentiment expressed in the text.
- **Persuasion prediction:** Determine whether one party in a conversation will be persuaded after a discussion.

Our method achieves or surpasses the current state-of-the-art levels in both tasks, without requiring additional training.

-> Our method demonstrates strong versatility and can be applied to a range of measurement tasks in social networks.

Stance Simulation with LLM

Stance Detection with Collaborative Role-Infused LLM-Based Agents

Xiaochong Lan, Chen Gao, Depeng Jin, Yong Li
Department of Electronic Engineering, BNRist, Tsinghua University, China
lanxc22@mails.tsinghua.edu.cn, {chgao96, jindp, liyong07}@tsinghua.edu.cn

Abstract

Stance detection automatically detects the stance in a text towards a target, vital for content analysis in web and social media research. Despite their promising capabilities, LLMs encounter challenges when directly applied to stance detection. First, stance detection demands multi-aspect knowledge, from deciphering event-related terminologies to understanding the expression styles in social media platforms. Second, stance detection requires advanced reasoning to infer authors' implicit viewpoints, as stance are often subtly embedded rather than overtly stated in the text. To address these challenges, we design a three-stage framework COLA (short for **C**ollaborative r**O**le-infused **L**LM-based **A**gents) in which LLMs are designated distinct roles, creating a collaborative system where each role contributes uniquely. Initially, in the multidimensional text analysis stage, we configure the LLMs to act as a linguistic expert, a domain specialist, and a social media veteran to get a multifaceted analysis of texts, thus overcoming the first challenge. Next, in the reasoning-enhanced debating stage, for each potential stance, we designate a specific LLM-based agent to advocate for it, guiding the LLM to detect logical connections between text features and stance, tackling the second challenge. Finally, in the stance conclusion stage, a final decision maker agent consolidates prior insights to determine the stance. Our approach avoids extra annotated data and model training and is highly usable. We achieve state-of-the-art performance across multiple datasets. Ablation studies validate the effectiveness of each design role in handling stance detection. Further experiments have demonstrated the explainability and the versatility of our approach. Our approach excels in usability, accuracy, effectiveness, explainability and versatility, highlighting its value.

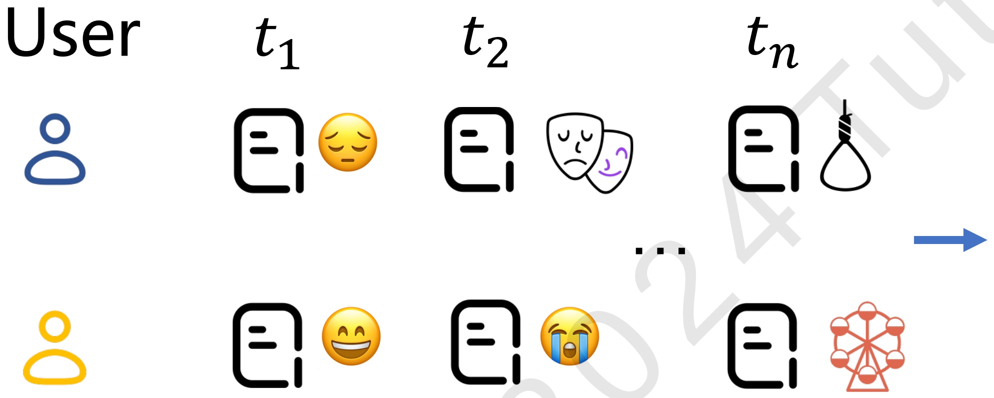
论文:
https://arxiv.org/abs/2310.10467

Stance Detection with Collaborative Role-Infused LLM-Based Agents. *X. Lan, **Chen Gao**, D. Jin, and **Yong Li**, ICWSM 2024 (Spotlight, top 4% among all submissions).*

Emotional Vulnerability Simulation with LLM

Vulnerable User Detection：

Assessing users’ emotional vulnerability levels based on texts posted on social media.



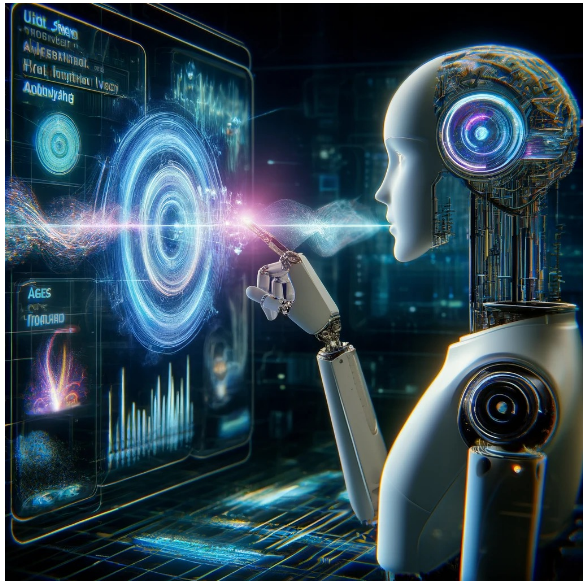
Which user is the vulnerable one?

- Different individuals exhibit **different patterns** of emotional changes.
- The threshold of events that cause emotional changes **varies**.
- Identifying emotional vulnerabilities is beneficial for accurately predicting individual **emotional changes** during the simulation.

Emotional Vulnerability Simulation with LLM

Generating Personalized User Characteristics in Simulation Environments

- **Manual Creation**
  - ➢ Not scalable for large scale simulation
- **Random Generation**
  - ➢ Lacks accuracy
- **Using Groundtruth Data**
  - ➢ Very difficult to directly access users' real characteristics
- **Automated Inference Based on Past User Behavior**
  - ➢ Reasonably accurate and scalable for large scale simulation
  - ✓ **Our approach**



Emotional Vulnerability Simulation: Challenges

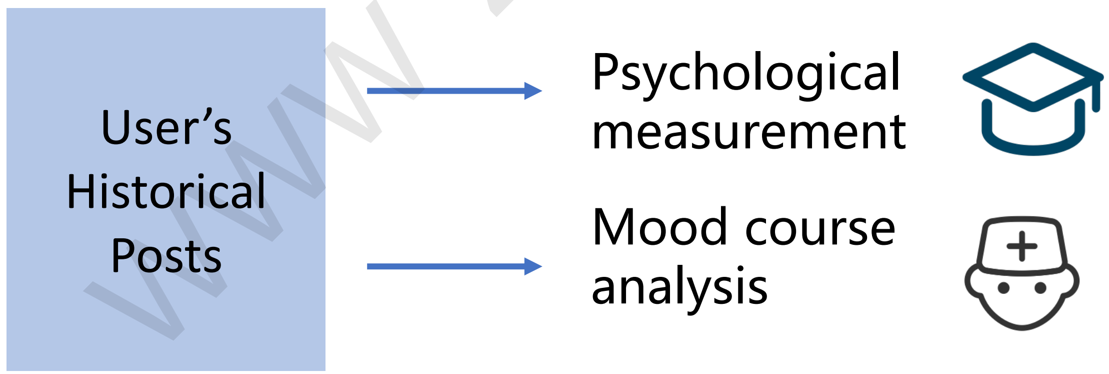
Challenges:

- Relying on professional psychology knowledge.
  - Modeling of emotions and patterns of emotional change cannot be entirely data-driven; it relies on professional knowledge in psychology.

- Necessary to ensure both high accuracy and high explainability.
  - While LLM classifiers naturally generate explanations for their decisions, their accuracy is lower than domain-specific models.
  - Traditional classifiers have high accuracy but poor explainability.

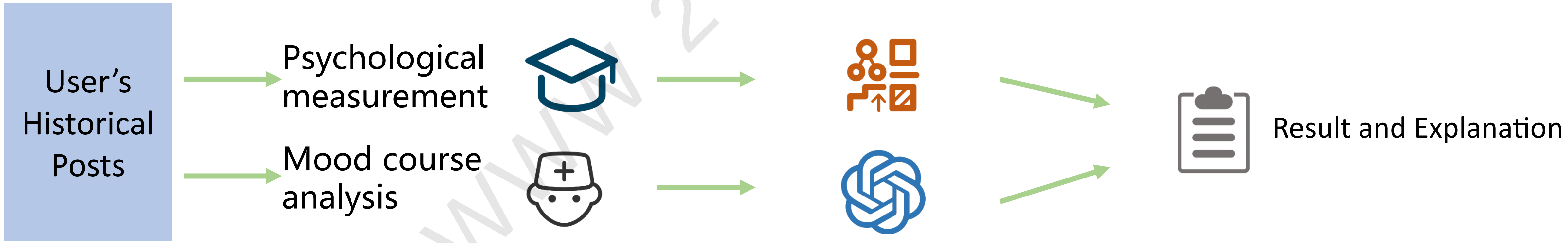
Emotional Vulnerability Simulation: Method

- The task relies on expertise in psychology (Challenge 1).
  - ✓ Using professional psychological scales, emotional vulnerability tags are applied to users with LLM.
  - ✓ Incorporating the concept of "mood course" from psychiatry, the LLM is used to explicitly model users' emotional history.



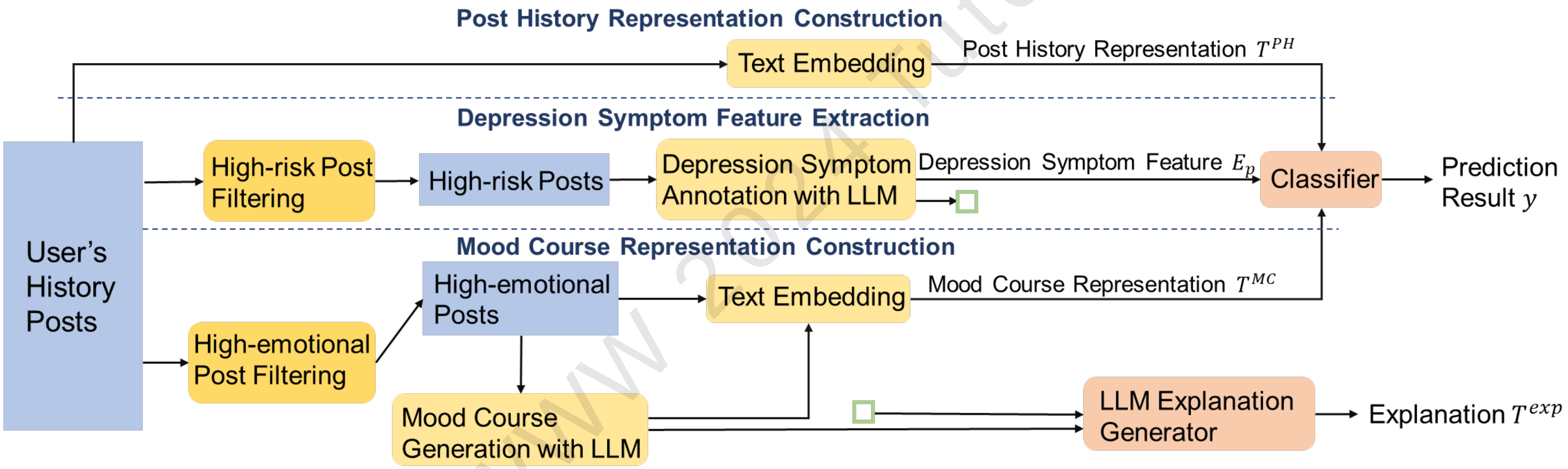
Emotional Vulnerability Simulation: Method

- The task needs to ensure high accuracy and high explainability (Challenge 2)
  - ✓ Classification with well-trained traditional classifiers to improve accuracy.
  - ✓ Utilizing LLM to produce explanations that include both arguments and reasoning, enhancing explainability.



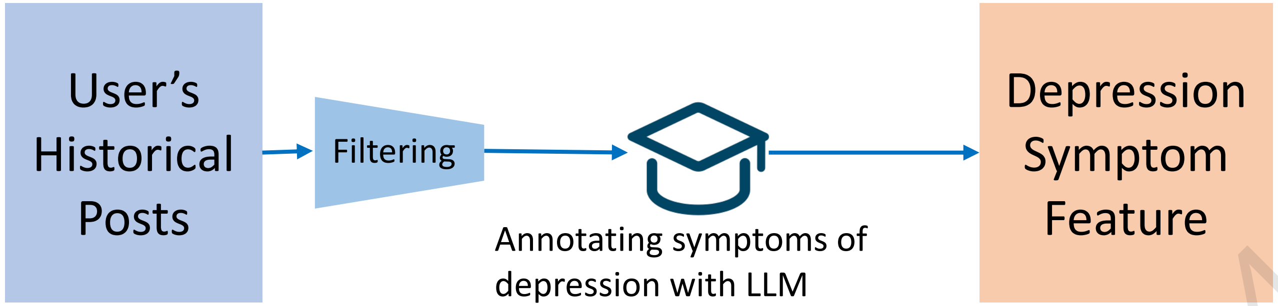
Collaboration between traditional classifiers and LLMs for classification

Emotional Vulnerability Simulation: Method



Emotional Vulnerability Simulation: Method

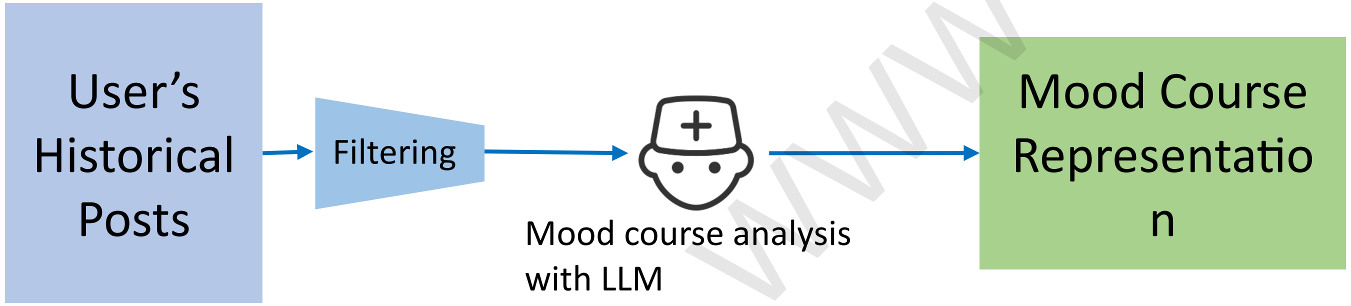
Diagnostic Criteria Feature Extraction



A. Depressed mood
B. Loss of interest/pleasure
C. Weight loss or gain
D. Insomnia or hypersomnia
E. Psychomotor agitation or retardation
F. Fatigue
G. Inappropriate guilt
H. Decreased concentration
I. Thoughts of suicide

Symptoms of depression defined in the DSM-5

Mood Course Representation Construction



- Mood course, defined as the temporal pattern and progression of emotional states, is critical in diagnosing clinical depression.
- It delineates the onset, duration, and recurrence of mood episodes, providing insights into the disorder's nature and trajectory.

Emotional Vulnerability Simulation: Method

Post History Representation Construction

User's Historical Posts → Text Embedding Model → Post History Representation

- Modeling a user's entire post history

Prediction and Explanation

Depression Symptom

Mood Course

Post History

- Classifying with GBT
- Explaining with LLM

Result and Explanation

Emotional Vulnerability Simulation: Experiments

Performance of our Method

	Depressed	Control
Num. of users	1000	19000
Num. of posts	69548	1314874
Avg. num. of posts per user	69.55	69.20

Category	Method	Precision	Recall	F1-score	AUROC	AUPRC
Traditional Method	TF-IDF+XGBoost	0.3644	0.4300	0.3945	0.9023	0.4303
Deep Learning-Based Methods	HAN	0.5702	0.6500	0.6075	0.8929	0.5864
	Mood2Content	0.7216	0.7000	0.7106	0.9537	0.7774
PLM-Based Methods	FastText	0.7467	0.5600	0.6400	0.9441	0.6255
	gte-small	0.6359	0.6526	0.6200	0.9499	0.6959
	BERT	0.6667	0.6400	0.6531	0.9481	0.7102
	MentalRoBERTa	0.7326	0.6300	0.6774	0.9423	0.6880
LLM-Based Methods	ChatGPT	0.0875	0.7100	0.1559	0.6603	0.0767
	MentalLLama	0.0899	0.7800	0.1612	0.6821	0.0811
Our Method	DORIS	**0.7596**	**0.7900**	**0.7596**	**0.9715**	**0.8134**

- Our method outperforms the current **SOTA** on all metrics.
- Due to the incorporation of knowledge, our method still maintains **high performance** on highly imbalanced datasets (1:19).

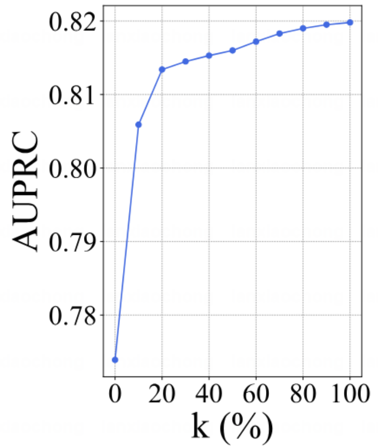
Emotional Vulnerability Simulation: Experiments

Ablation Study

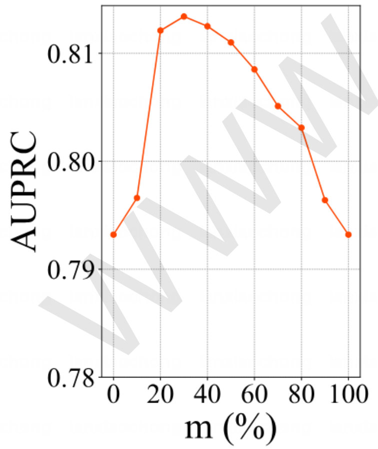
	F1-score	AUROC	AUPRC
Full Design	0.7596	0.9715	0.8134
w/o DC Feature	0.6867	0.9679	0.7739
w/o MC Representation	0.7415	0.9660	0.7932
w/o PH Representation	0.7200	0.9660	0.7817

- Every module in our approach contributes to performance gain.

Hyperparameter Study



(a) The impact of k.



(b) The impact of m.

- The parameters k and m determine the proportion of posts we process with high-performance LLM.
- The trend of k and m indicates that under our filtering mechanism, processing about 20% of tweets with LLMs yields the best or near-best results.
- Our method conserves computational resources.

Emotional Vulnerability Simulation: Experiments

Explainability of our Method

Can our method provide high-quality explanations for its decisions?

Diagnostic Criteria Feature Construction

Post Time	Text	Symptoms
April 23rd, 2015	The incident from two years ago, I'll never forget it for the rest of my life.	None
April 30th, 2015	Driven to the brink of agitation, consumed by an overwhelming restlessness.	Psychomotor agitation
May 3rd, 2015	I'm going back to school tomorrow, and I haven't touched volleyball in a month.	None
May 8th, 2015	The bitterness in my heart, no one understands. Even if I cry until tears run dry.	Depressed mood
...	...	...
July 10th, 2015	I didn't realize I could lose so much weight without deliberately dieting.	Weight loss
July 16th, 2015	I really dislike people who lack integrity, and I find it very repulsive.	None
July 22th, 2015	I just want to find a place to hide and cry my heart out right now.	Depressed mood

Mood Course Representation Construction

Post Time	Text with High Emotional Content
April 30th, 2015	Driven to the brink of agitation, consumed by an overwhelming restlessness.
May 8th, 2015	The bitterness in my heart, no one understands. Even if I cry until tears run dry.
...	...
July 16th, 2015	I really dislike people who lack integrity, and I find it very repulsive.
July 22th, 2015	I just want to find a place to hide and cry my heart out right now.

Mood Course Description

The individual's social media expressions from April to July 2015 depict a concerning trajectory of emotional distress. Initially, they report intense agitation and restlessness, suggesting acute emotional turmoil. As weeks progress, there's a deepening sense of isolation and misunderstood bitterness, indicating potential depressive symptoms. Episodes of wanting to withdraw socially and expressions of disdain towards perceived moral failings in others hint at both a heightened sensitivity to interpersonal dynamics and a profound sense of disillusionment. The culmination of these feelings into a desire for seclusion and emotional release by mid-July underscores a significant emotional struggle, possibly pointing towards depressive episodes or acute stress reactions. This mood course, characterized by fluctuations in emotional intensity and themes of alienation, distress, and a quest for solace, warrants professional attention for potential mental health interventions.

Generated Explanation

Based on the user's mood course and specific posts, the judgment of depression appears to be accurate. The reported symptoms—psychomotor agitation, a persistently depressed mood, unintentional weight loss, and a desire for isolation—are hallmark indicators of depression. These symptoms not only cover a range of depressive disorder criteria, including mood disturbances, physical changes, and social withdrawal, but also span several months, suggesting a sustained period of emotional distress rather than a transient phase. The pattern and intensity of symptoms described align with clinical understanding of depressive episodes, warranting a diagnosis of depression rather than a normal fluctuation of emotions.

The explanation include:

- Psychological Scale Annotation Results
- High Emotional Sensitivity Text Extraction Results
- Mood Course Description
- Decision Explanation

Our method provides high-quality explanations for its judgments **from multiple perspectives** and **with evidence**.

Emotional Vulnerability Simulation: Preprint Paper

Depression Detection on Social Media with Large Language Models

Xiaochong Lan, Yiming Cheng, Li Sheng, Chen Gao, Yong Li
Department of Electronic Engineering, BNRist, Tsinghua University, China
lanxc22@mails.tsinghua.edu.cn, {chgao96, liyong07}@tsinghua.edu.cn

arXiv:2403.10750v1 [cs.CL] 16 Mar 2024

Abstract

Depression harms. However, due to a lack of mental health awareness and fear of stigma, many patients do not actively seek diagnosis and treatment, leading to detrimental outcomes. Depression detection aims to determine whether an individual suffers from depression by analyzing their history of posts on social media, which can significantly aid in early detection and intervention. It mainly faces two key challenges: **1)** it requires professional medical knowledge, and **2)** it necessitates both high accuracy and explainability. To address it, we propose a novel depression detection system called DORIS, combining medical knowledge and the recent advances in large language models (LLMs). Specifically, to tackle the first challenge, we proposed an LLM-based solution to first annotate whether high-risk texts meet medical diagnostic criteria. Further, we retrieve texts with high emotional intensity and summarize critical information from the historical mood records of users, so-called *mood courses*. To tackle the second challenge, we combine LLM and traditional classifiers to integrate medical knowledge-guided features, for which the model can also explain its prediction results, achieving both high accuracy and explainability. Extensive experimental results on benchmarking datasets show that, compared to the current best baseline, our approach improves by 0.036 in AUPRC, which can be considered significant, demonstrating the effectiveness of our approach and its high value as an NLP application.

1 Introduction

Depression profoundly affects humanity, with WHO estimates indicating that 5% of adults suffer from it[1], significantly contributing to the global suicide rate[2]. Given the harmful nature of depression, timely diagnosis and intervention are necessary. However, traditional hospital-based approaches for diagnosing depression face several issues. Firstly, patients often avoid evaluations due to stigma or not recognizing their need for help [34]. What's more, self-reported diagnoses can be unreliable due to intentional concealment [4]. Additionally, the high cost of hospital evaluations places a burden on patients [22]. For these reasons, many individuals with depression remain undiagnosed and untreated, with over 75% of those in low- and middle-income countries receiving no treatment at all[3].

Depression detection on social media identifies potential depression through users' post histories [29, 35], as shown in Figure 1, which is one typical beneficial application of NLP techniques. That is, it leverages public posts from online social networks for broad detection coverage, benefits from more genuine expressions than those in clinical settings [23], and reduces economic costs compared to professional diagnoses.

[1]www.who.int/news-room/fact-sheets/detail/depression

论文:
https://arxiv.org/pdf/2403.10750

Depression Detection on Social Media with Large Language Models *X. Lan, Y. Cheng, L. Sheng,* ***Chen Gao*** *and* ***Yong Li****, preprint 2024.*

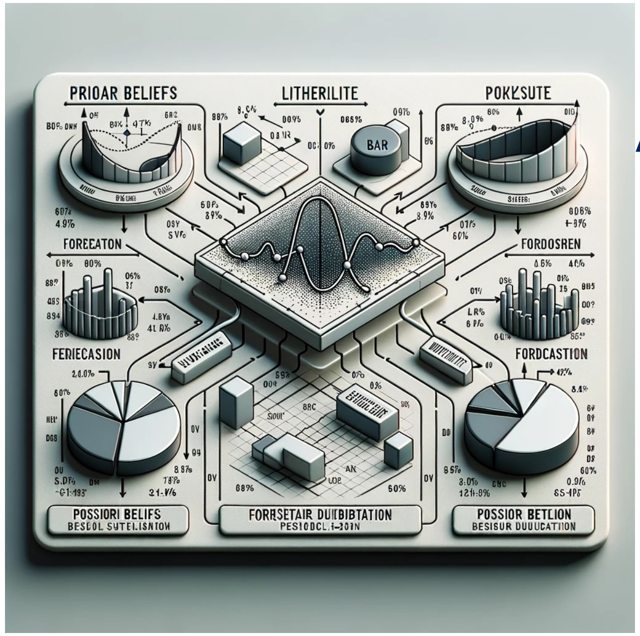
Outline

- Background of LLM Agent-based Simulation
- Online behavior simulation with LLM Agents
- Social and economic simulation with LLM agents
  - Economic simulation
- City system simulation with LLM agents
- Open discussions

LLM-based Macroeconomic Simulator：Background

Traditional Forecasting

Approach 1: Statistical Model



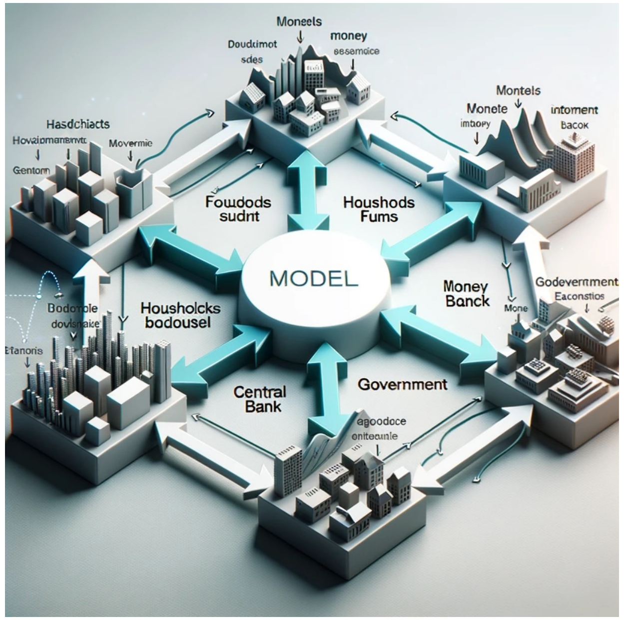
Assumption of Stability

"The policy predictions of the models that are in use aren't wrong, they are simply non-existent."

Traditional Forecasting

Approach 2: DSGE

Dynamic Stochastic General Equilibrium



Assumption of Perfect World

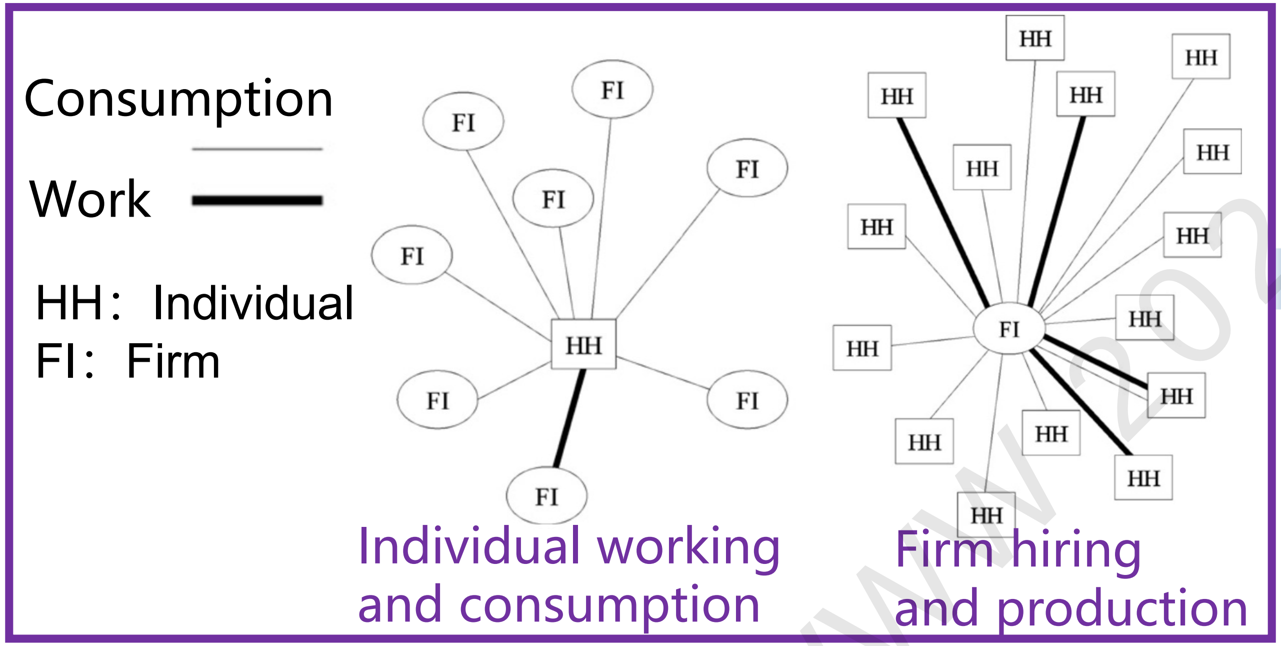
Simulation with ABM:

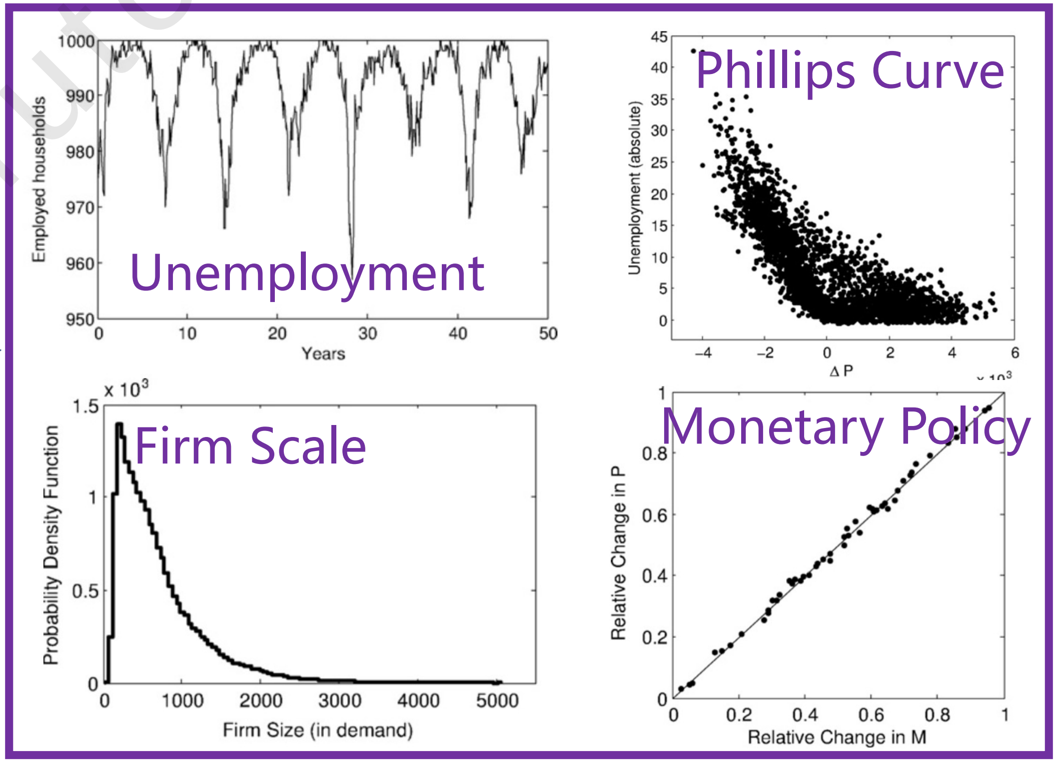
Macroeconomic phenomena emerge from individual behavior

Farmer, J. Doyne, and Duncan Foley. "The economy needs agent-based modelling." *Nature* 460.7256 (2009): 685-686.

LLM-based Macroeconomic Simulator：LEN (Baseline)

Business Cycle emerges from rule-based individual and firm behavior



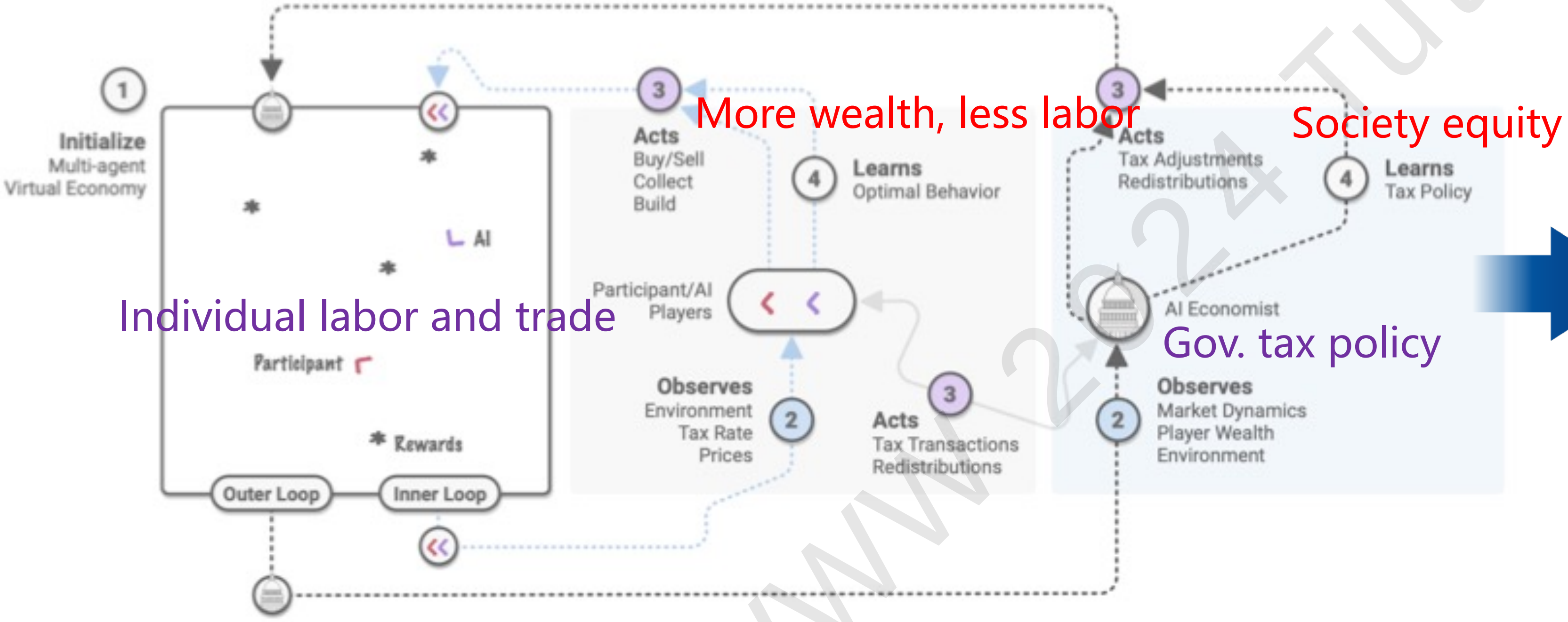


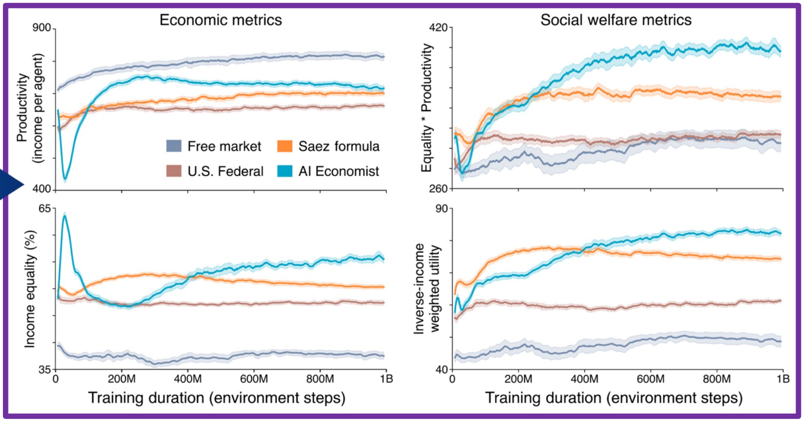
Macroeconomic phenomena emerge from individual behavior, without rationality and equilibrium assumption

Lengnick, Matthias. "Agent-based macroeconomics: A baseline model." *Journal of Economic Behavior & Organization* 86 (2013): 102-120.

LLM-based Macroeconomic Simulator：RL Approach

RL: maximize average wealth and society equity





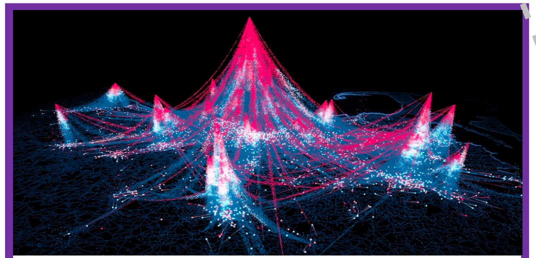
Taking social welfare as the RL optimization goal to achieve optimal tax policy

Zheng, Stephan, et al. "The AI Economist: Taxation policy design via two-level deep multiagent reinforcement learning." *Science advances* 8.18 (2022): eabk2607.

Shortcomings of Existing ABM Economic Simulation

1. Perception

Heterogeneous individuals' perception of the real-world economic environment



2. Reflection

Expectation of dynamic changes in the market environment

3. Decision

Consider the impact of various economic factors

Perception

Shortcomings & Challenges

Reflection

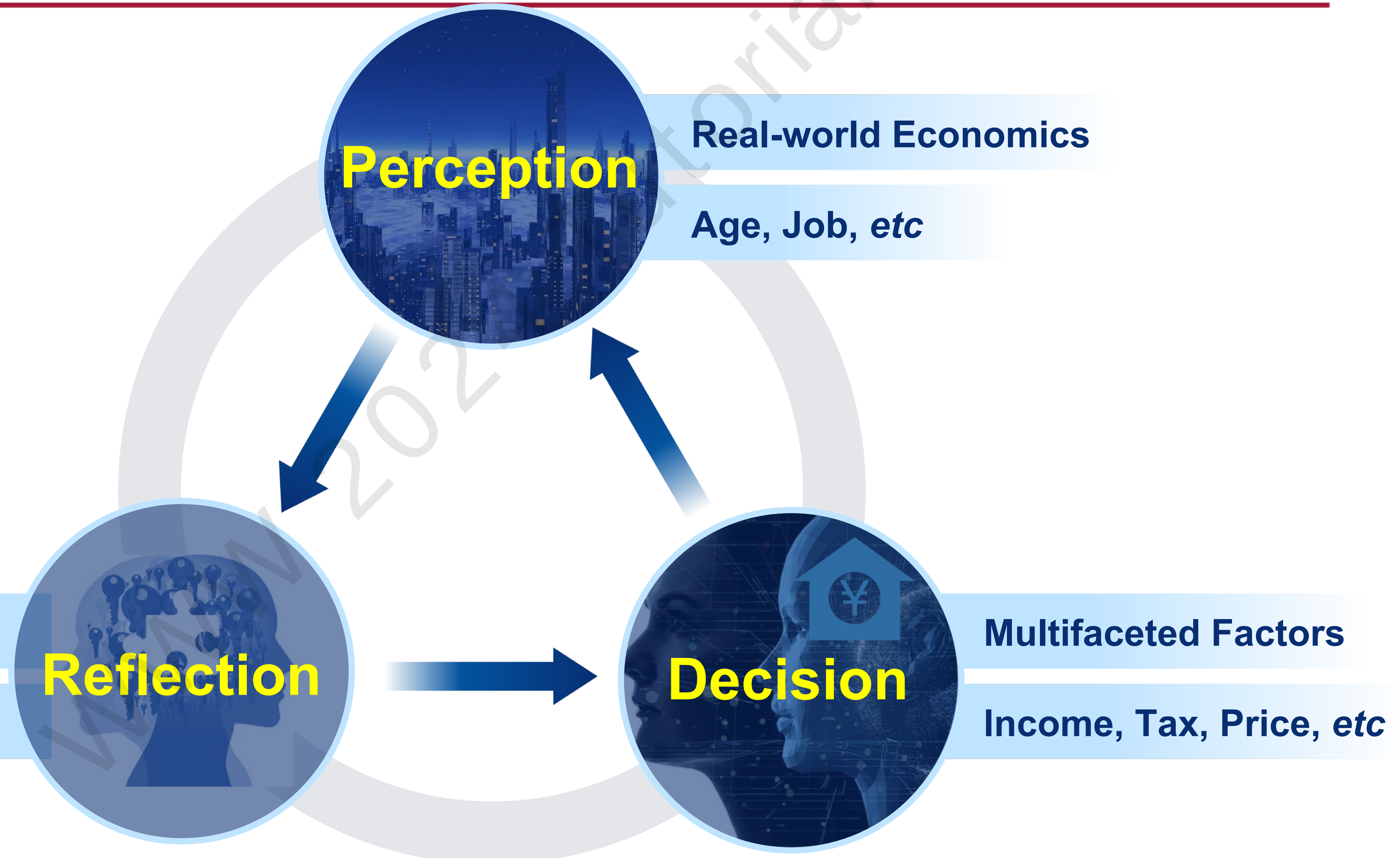
Decision

Perception, reflection and decision-making support the emergence of macroeconomic phenomena

Why LLM for Economic Simulation

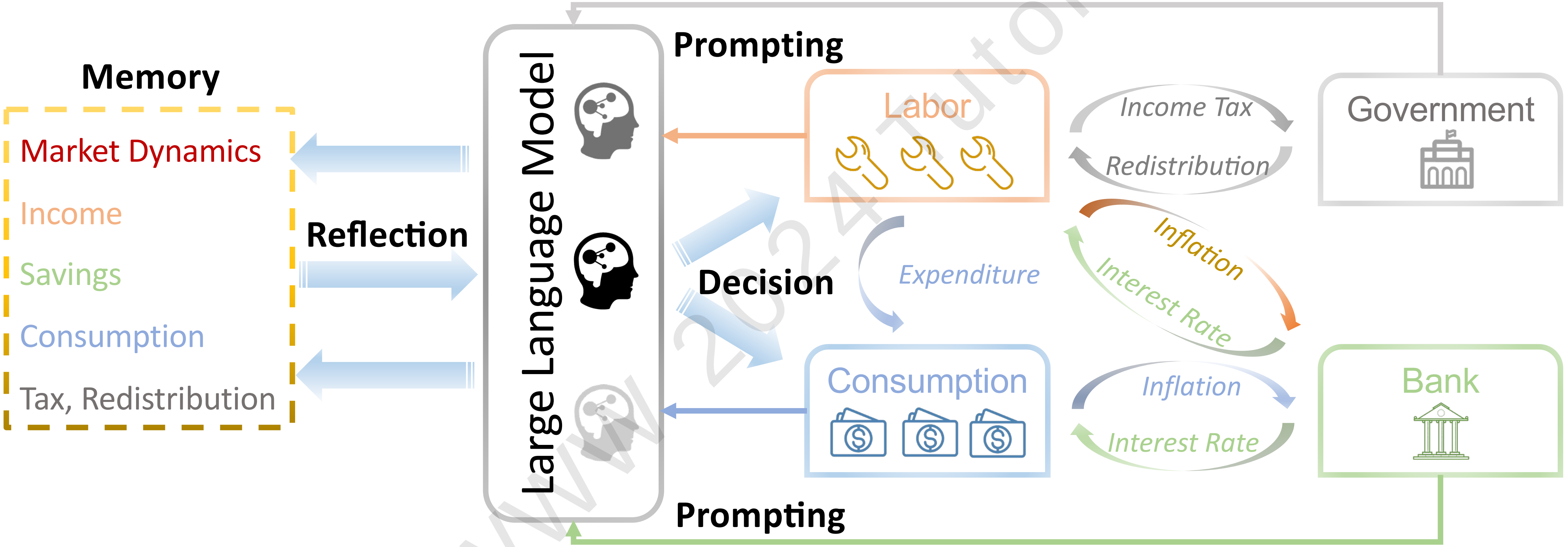
LLM-empowered agent behavior

- Perception
- Reflection
- Decision

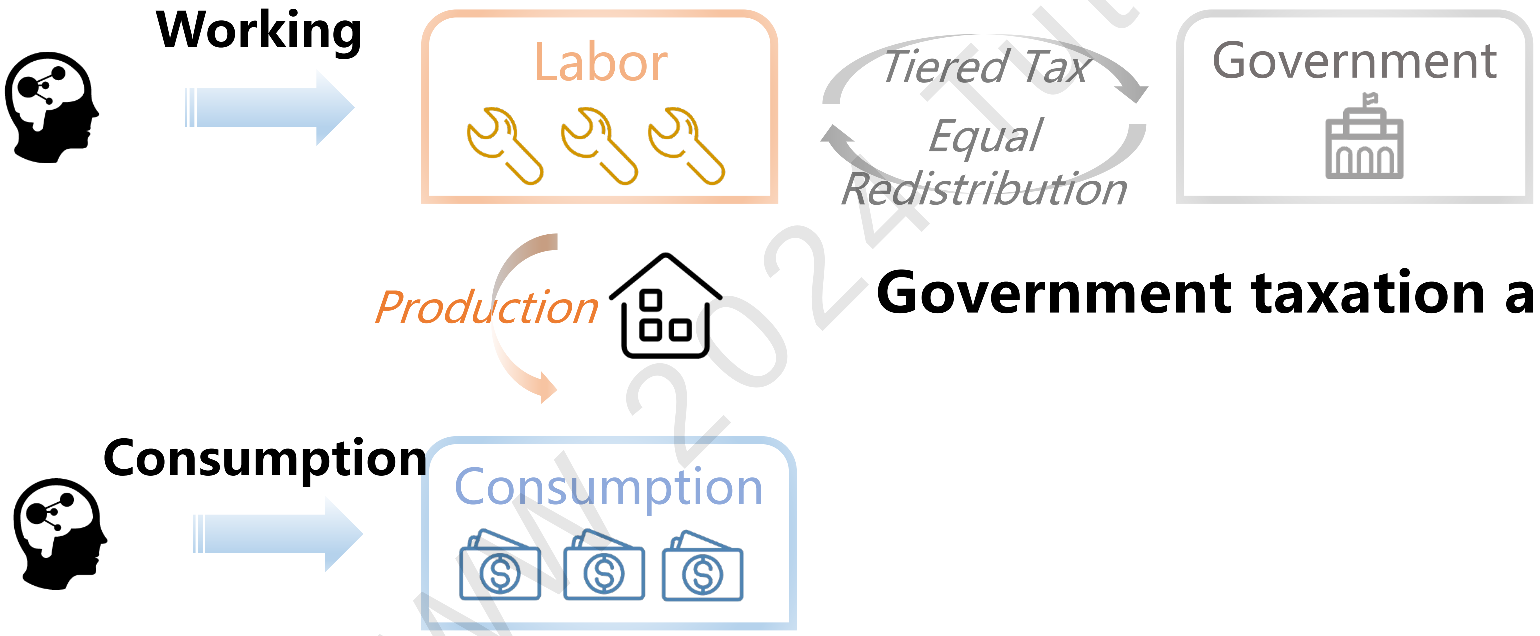


LLM agent has human-like economic behavior characteristics

LLM-based Macroeconomic Simulator：Framework



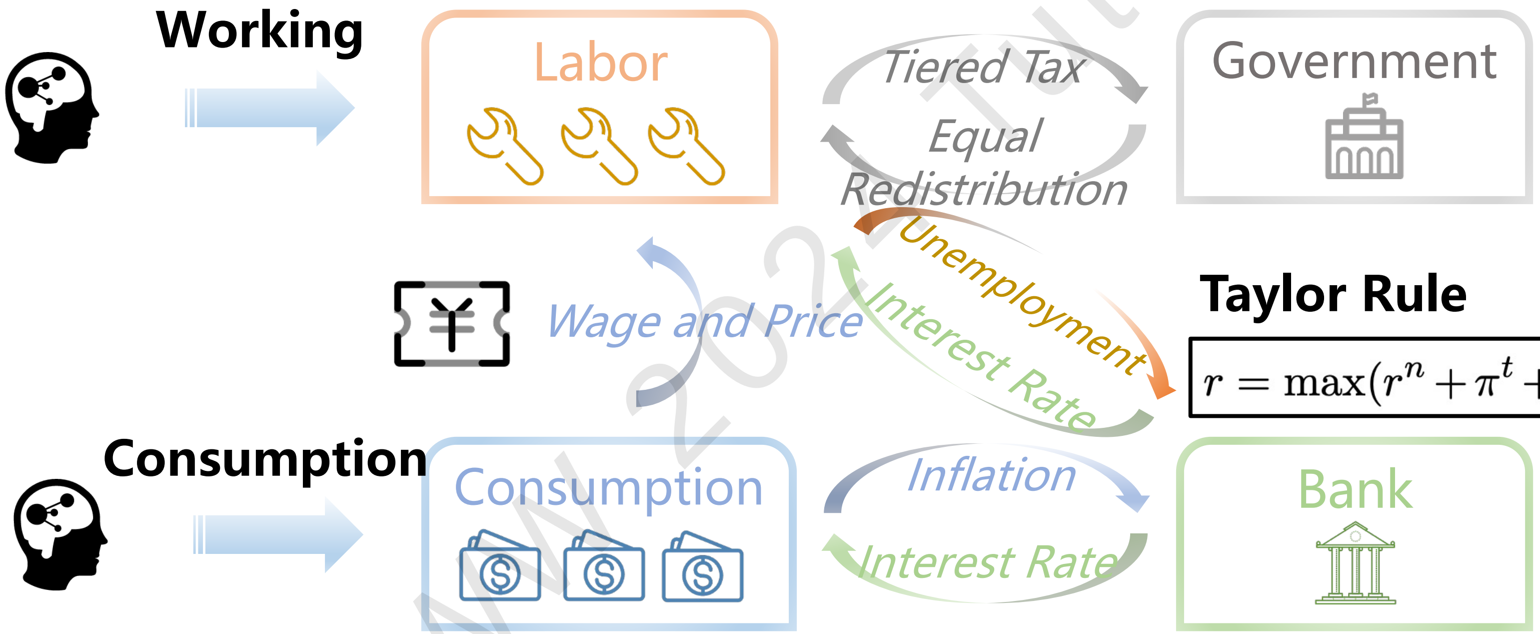
LLM-based Macroeconomic Simulator：Environment



Government taxation and redistribution

Goods production and consumption

LLM-based Macroeconomic Simulator: Environment



Taylor Rule

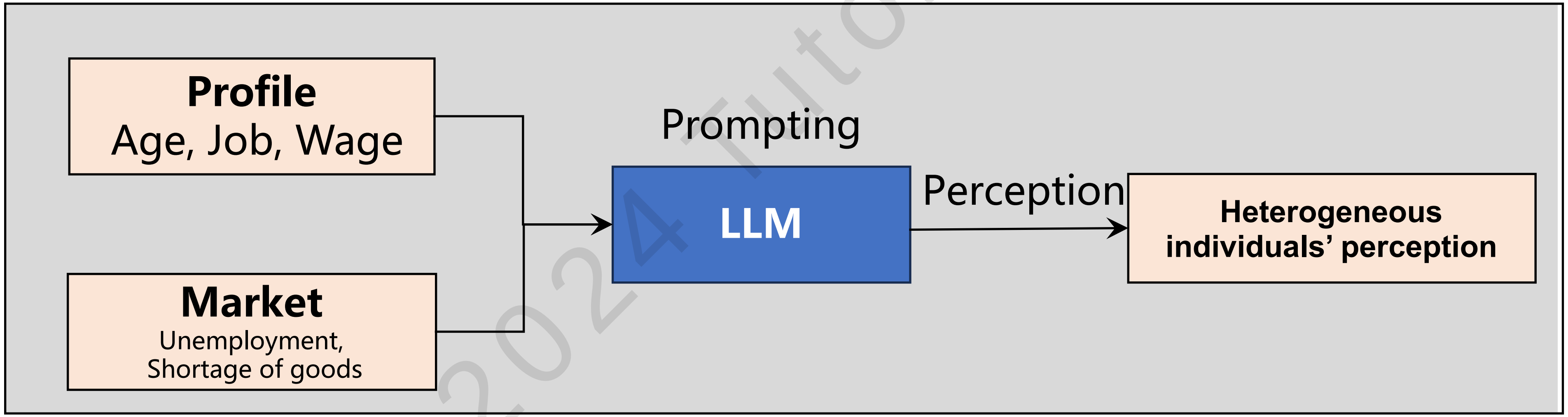
$$r = \max(r^n + \pi^t + \alpha^\pi(\pi - \pi^t) + \alpha^u(u^n - u), 0)$$

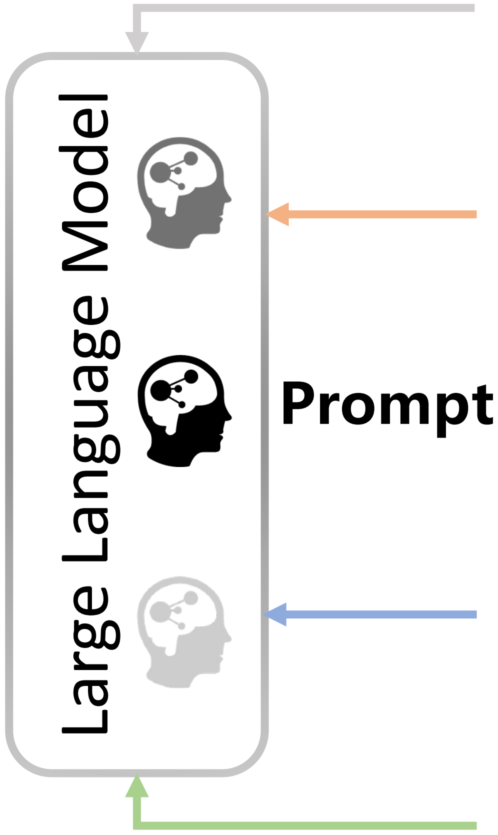
The impact of goods production and consumption on wages and prices

The impact of market conditions on interest rates (and savings)

LLM-based Macroeconomic Simulator：Agent Design

1. Perception of real-world economics





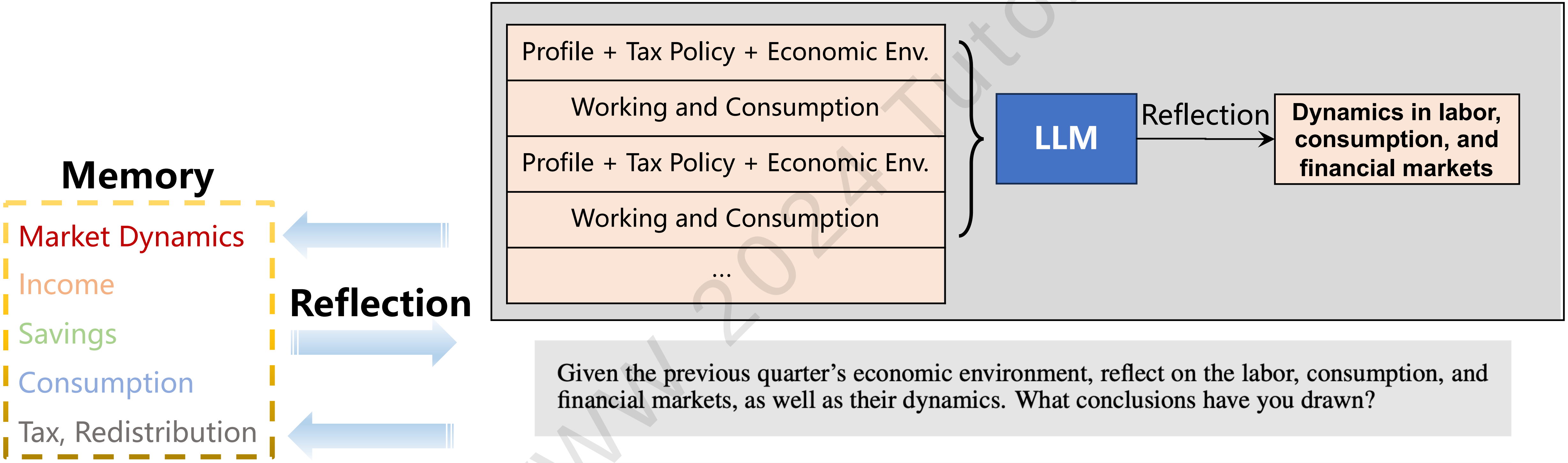
Prompting

In the previous month, you became unemployed and had no income. Now, you are invited to work as a(an) [*offer*] with a monthly salary of [*wage*].

Risks from unemployment last month increase willingness to work this month

LLM-based Macroeconomic Simulator：Agent Design

2. Reflect on the past economics



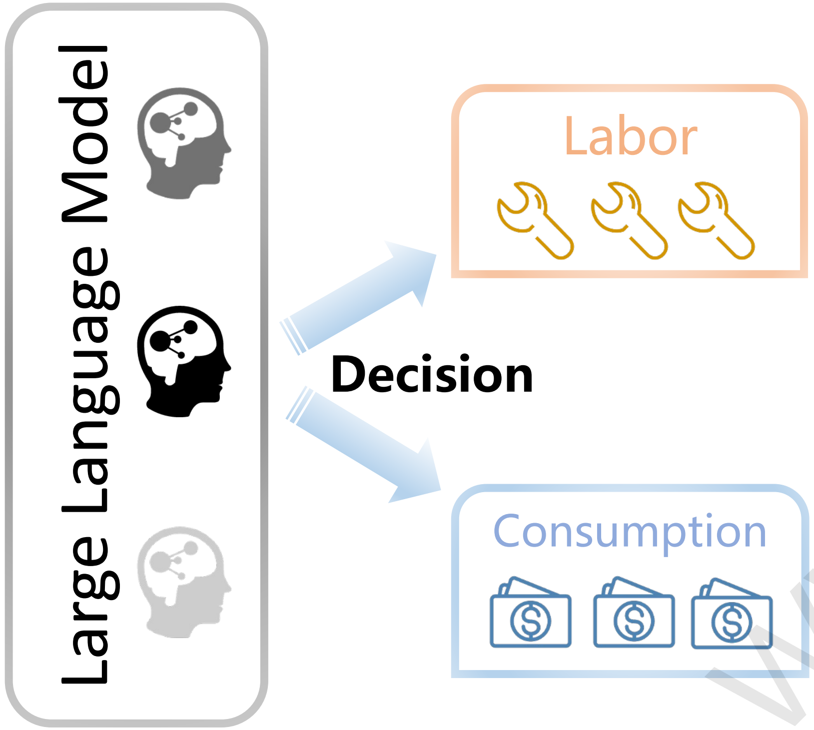
Given the previous quarter's economic environment, reflect on the labor, consumption, and financial markets, as well as their dynamics. What conclusions have you drawn?

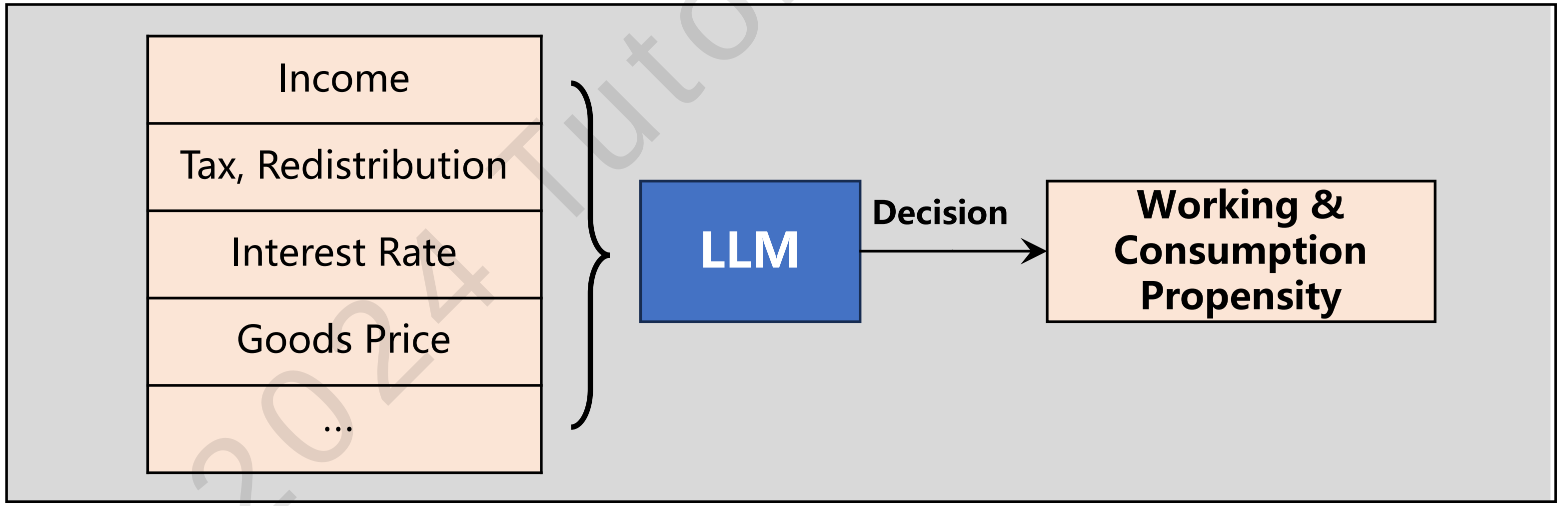
Based on the previous quarter's data, the **labor market experienced deflation**... The consumption market also saw a **decrease in prices for essential goods**... The financial market's interest rates **remained unchanged at 3.00%**. Overall, the quarter highlighted the need for **careful financial planning and adaptability in response to market fluctuations**.

Respond to dynamics in labor and consumer markets with adaptive decision-making

LLM-based Macroeconomic Simulator：Agent Design

3. Consider the impact of multifaceted economic factors



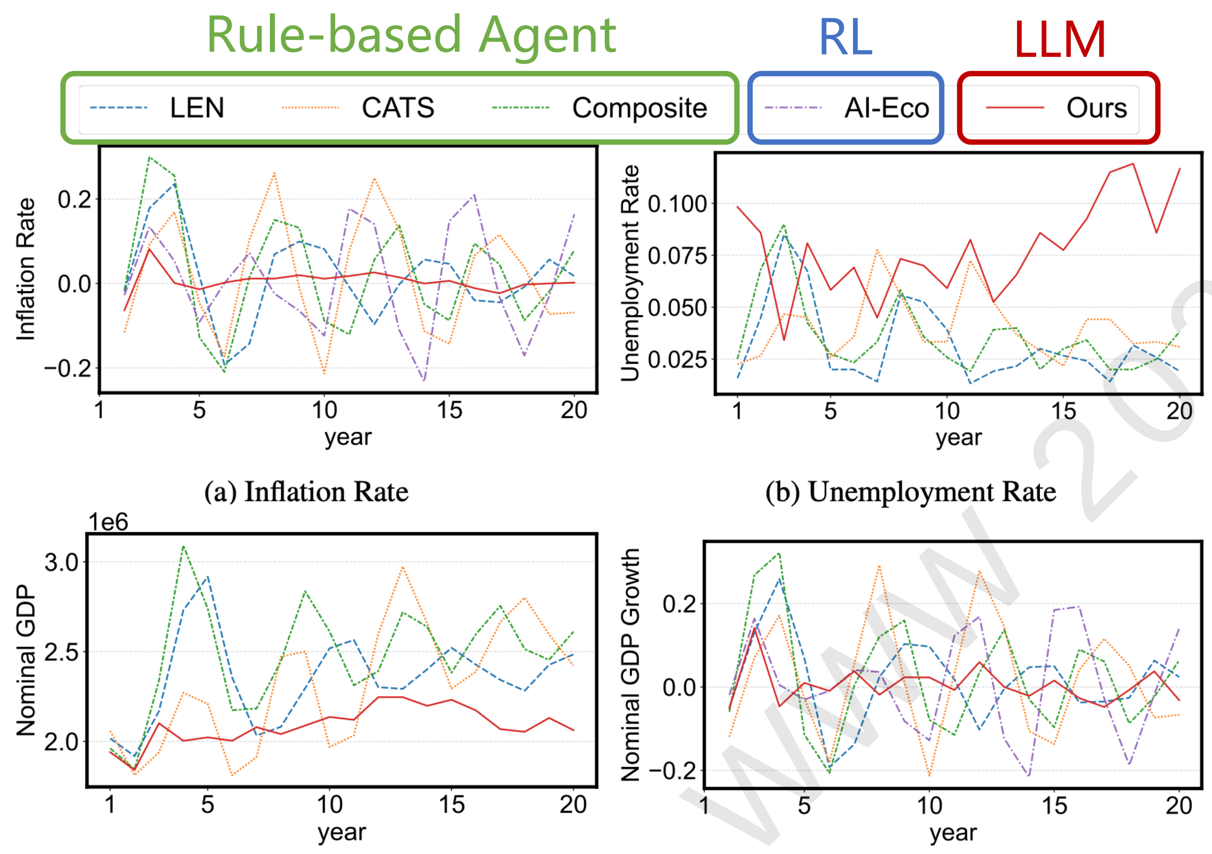


With all these factors in play, and considering aspects like your living costs, any future aspirations, and the broader economic trends, how is your willingness to work this month? Furthermore, how would you plan your expenditures on essential goods, keeping in mind goods price?

Work and consumption decisions that consider multifaceted economic factors

LLM-based Macroeconomic Simulator: Experiments

Inflation, Unemployment Rate, GDP, GDP Growth Rate

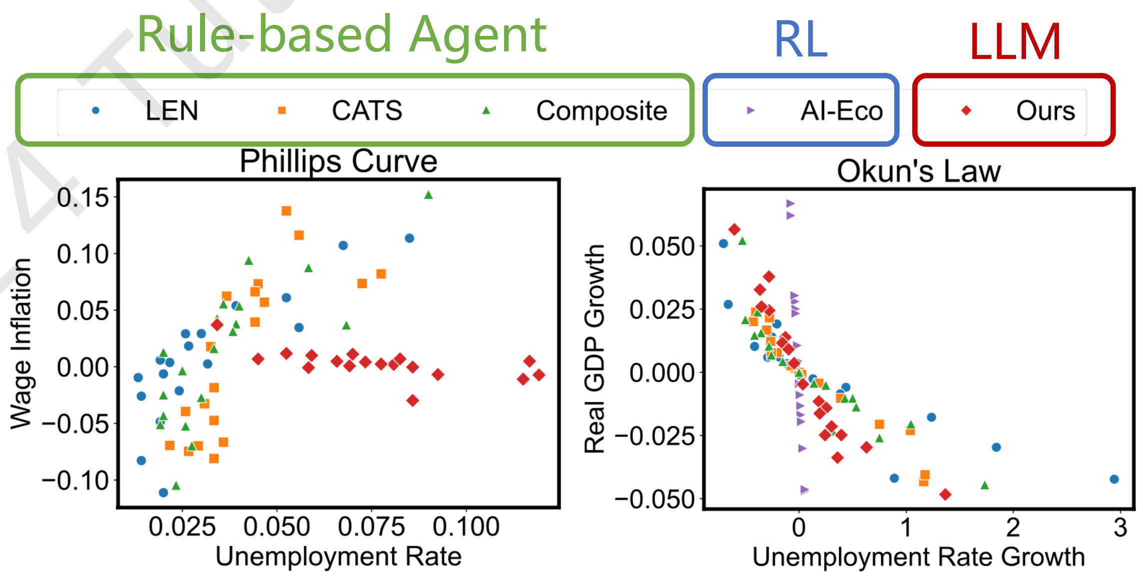


(a) Inflation Rate (b) Unemployment Rate

(c) Nominal GDP (d) Nominal GDP Growth Rate

Macroeconomic indicators with smaller fluctuations and more reasonable numerical ranges

Phillips Curve, Okun's Law



Macroeconomic regularities consistent with empirical facts

↓

Emerging macroeconomic phenomena

LLM-based Macroeconomic Simulator：Experiments

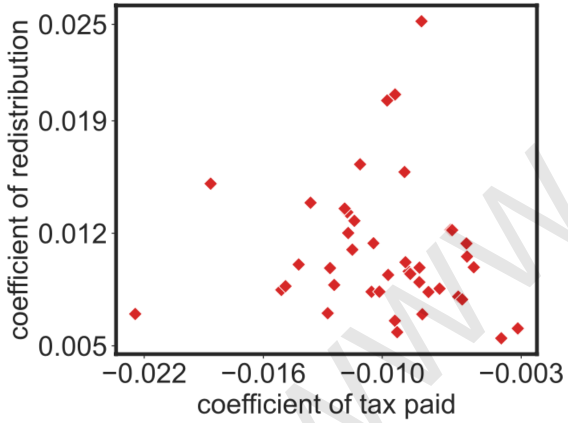
Regression

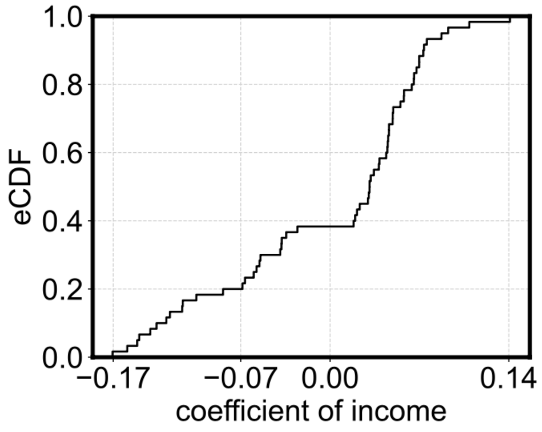
Working & Consumption Decision ~ Salary, Tax, Redistribution, Savings, Price, Interest Rate

#agents with significant regression coefficients

		v_i	$\hat{c}_i$	$T(z_i)$	z^r	P	s_i	r
Working	p_i^w	60	37	60	65	58	56	31
Consumption	p_i^c	65	73	51	52	62	100	49

For **work** decisions, regression coefficients are used to study the **impact of monthly salary, taxes, and financial rebates**.





For all agents, **paying less taxes and receiving more financial rebates** increases willingness to work

For more than 60% of agents, **increasing monthly salary** will increase work willingness

LLM-based Macroeconomic Simulator：Experiments

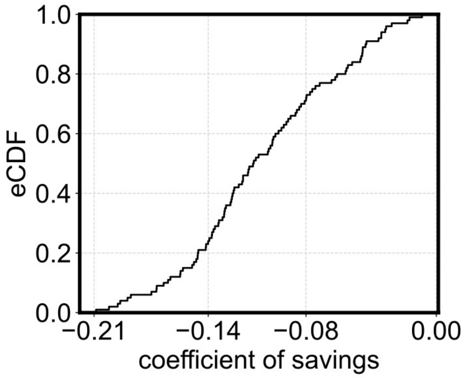
Regression

Working & Consumption Decision ~ Salary, Tax, Redistribution, Savings, Price, Interest Rate

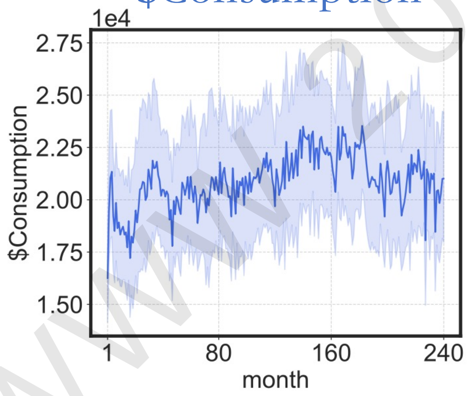
#agents with significant regression coefficients

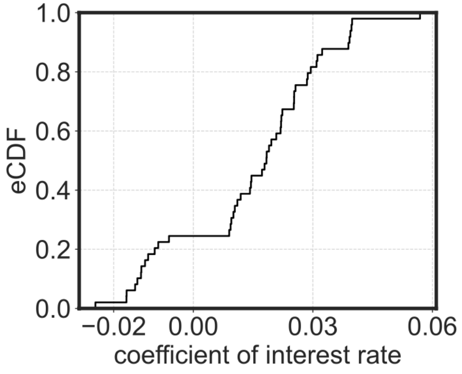
		v_i	$\hat{c}_i$	$T(z_i)$	z^r	P	s_i	r
Working	p_i^w	60	37	60	65	58	56	31
Consumption	p_i^c	65	73	51	52	62	100	49

For **consumption** decisions, use regression coefficients to study the **impact of savings and interest rates.**



$Consumption





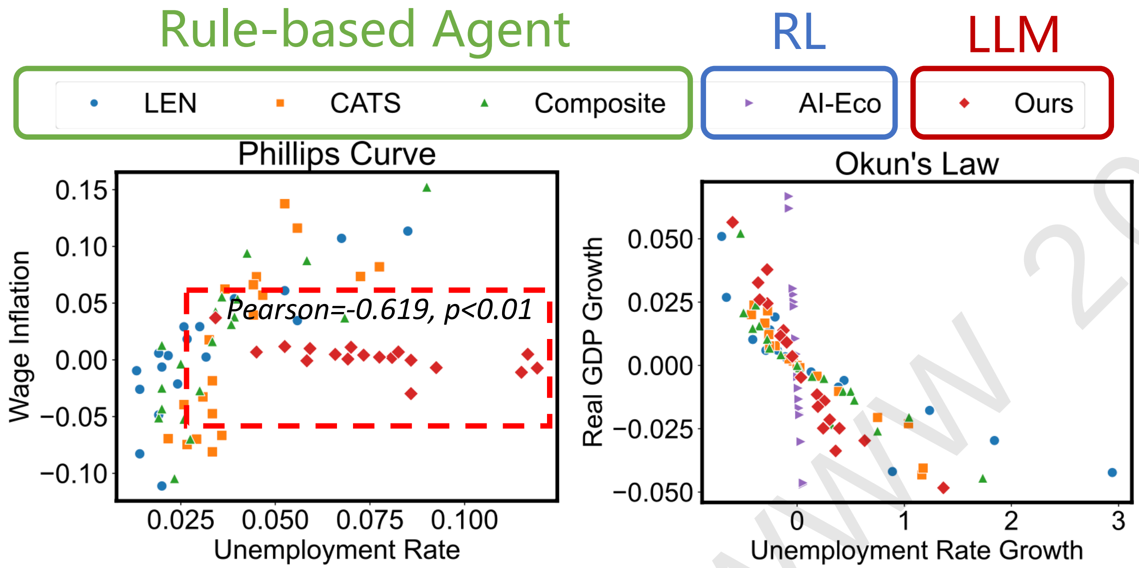
Savings↑ → Consumption Propensity (Proportion)↓
→ Stable consumption

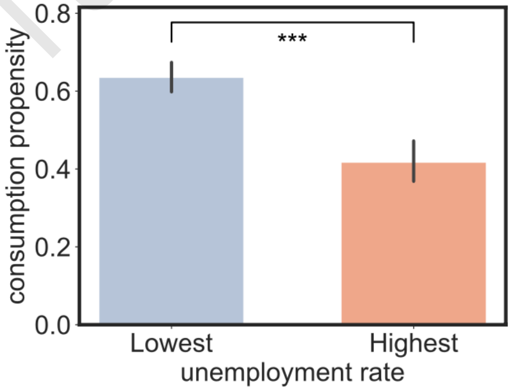
For more than 70% of agents, **high interest rates** will increase consumption willingness

LLM-based Macroeconomic Simulator：Experiments

Only LLM agent decision-making gives the correct Phillips Curve

- The negative relationship between unemployment rate and wage inflation





Comparison of consumption willingness in the two years with the highest/lowest unemployment rate

In the last quarter, I have adjusted my **willingness to work** and my **planned expenditures** on essential goods slightly **downwards**. This decision is primarily influenced by the **continued deflation in the labor market**, resulting in a decrease in my expected income. With a lower income, I need to be **cautious about my spending** and ensure that I have enough savings for **future expenses and unforeseen circumstances**. . .

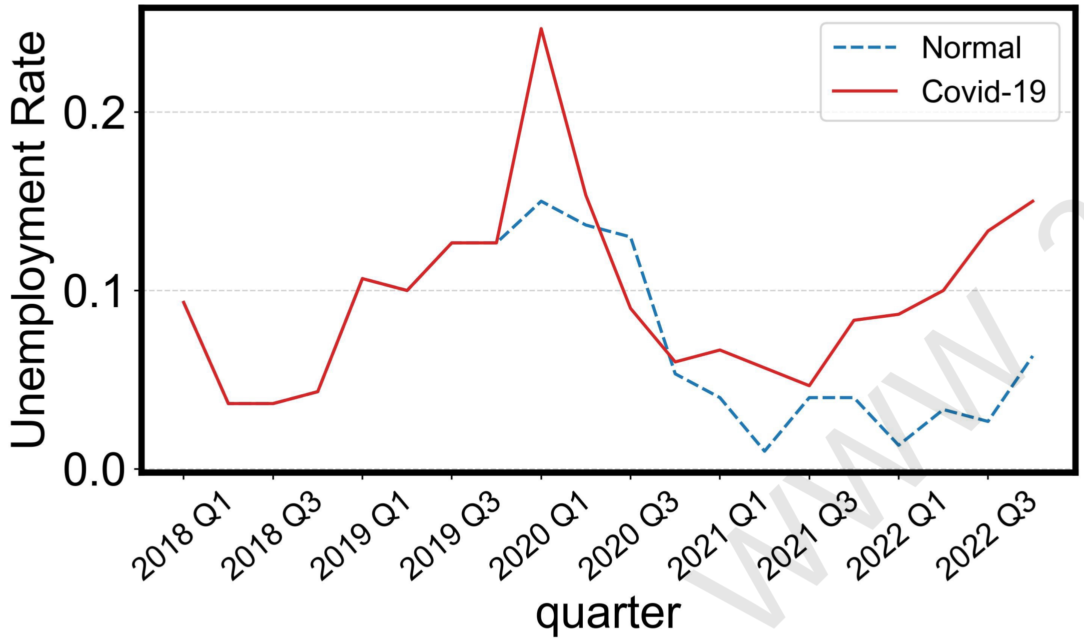
Ask about the reasons for LLM decisions

Grasp the market environment:
consumption reduction caused by economic downturn under high unemployment rate

LLM-based Macroeconomic Simulator：Experiments

Intervention: The impact of external shocks

➢ Take COVID-19 as an example



New prompt

"In response to the large-scale outbreak of Covid-19 in the United States, the federal government has declared a national emergency since March 2020."

Agent Reflection

*... However, the outbreak of Covid-19 and the subsequent national emergency declaration had a **significant impact on the labor market**. Many businesses were forced to close or reduce their operations, resulting in **widespread unemployment and uncertainty**. This situation has likely **affected my willingness to work**, as job security and health concerns become more prominent ...*

COVID-19 restrictions bring surge in unemployment

LLM-based Macroeconomic Simulator：Preprint Paper

Large Language Model-Empowered Agents for Simulating Macroeconomic Activities

Nian Li, Chen Gao, Yong Li, Qingmin Liao
Tsinghua University
linian21@mails.tsinghua.edu.cn, {chgao96, liyong07, liaoqm}@tsinghua.edu.cn

Abstract

The advent of the Web has brought about a paradigm shift in traditional economics, particularly in the digital economy era, enabling the precise recording and analysis of individual economic behavior. This has led to a growing emphasis on data-driven modeling in macroeconomics. In macroeconomic research, Agent-based modeling (ABM) emerged as an alternative, evolving through rule-based agents, machine learning-enhanced decision-making, and, more recently, advanced AI agents. However, the existing works are suffering from three main challenges when endowing agents with human-like decision-making, including agent heterogeneity, the influence of macroeconomic trends, and multifaceted economic factors. Large language models (LLMs) have recently gained prominence in offering autonomous human-like characteristics. Therefore, leveraging LLMs in macroeconomic simulation presents an opportunity to overcome traditional limitations. In this work, we take an early step in introducing a novel approach that leverages LLMs in macroeconomic simulation. We design prompt-engineering-driven LLM agents to exhibit human-like decision-making and adaptability in the economic environment, with the abilities of perception, reflection, and decision-making to address the abovementioned challenges. Simulation experiments on macroeconomic activities show that LLM-empowered agents can make realistic work and consumption decisions and emerge more reasonable macroeconomic phenomena than existing rule-based or AI agents. Our work demonstrates the promising potential to simulate macroeconomics based on LLM and its human-like characteristics.

Paper:
https://arxiv.org/abs/2310.10436

Large Language Model-Empowered Agents for Simulating Macroeconomic Activities.
*N. Li, **Chen Gao, Yong Li**, and Q. Liao. arXiv preprint arXiv:2310.10436 (2023).*

Simulating Human Society with Large Language Model Agents: City, Social Media, and Economic System

Chen Gao[1], Fengli Xu[1], Xu Chen[2], Xiang Wang[3], Xiangnan He[3] and Yong Li[1]

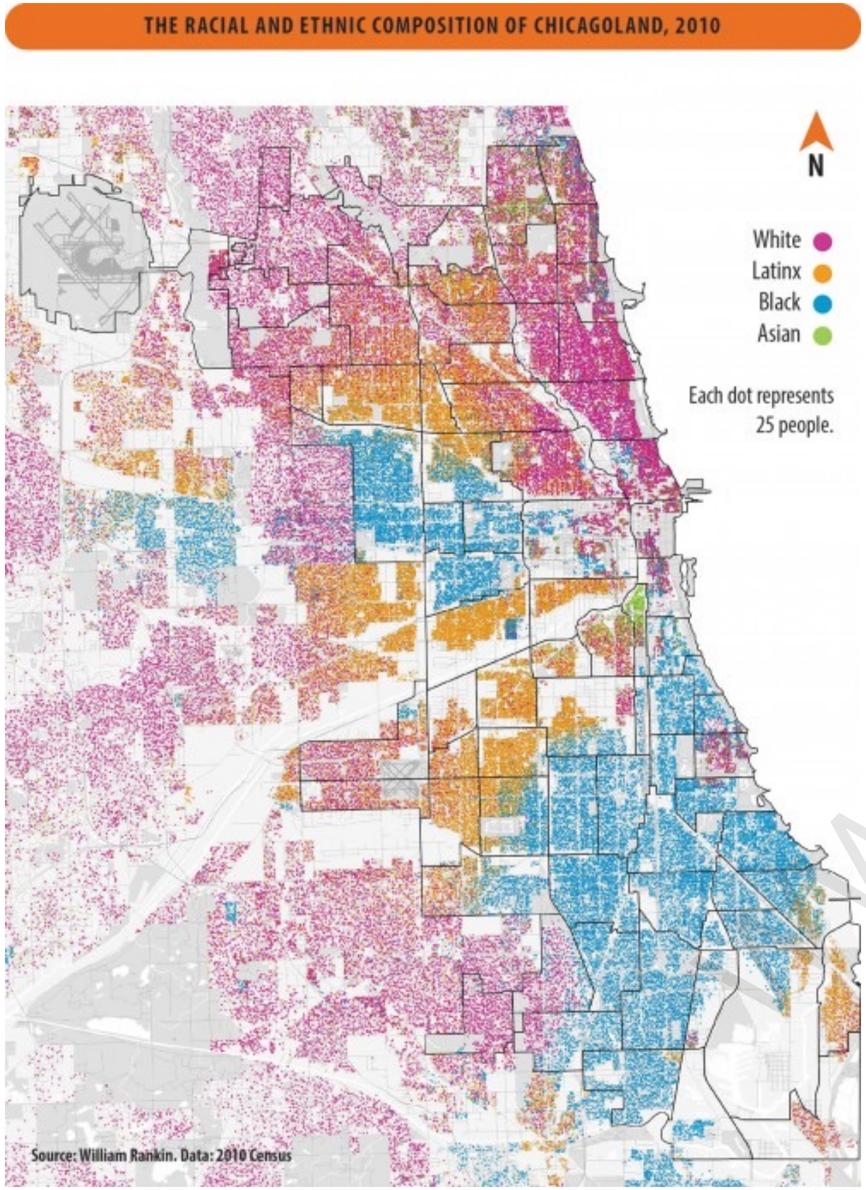
[1]Tsinghua University

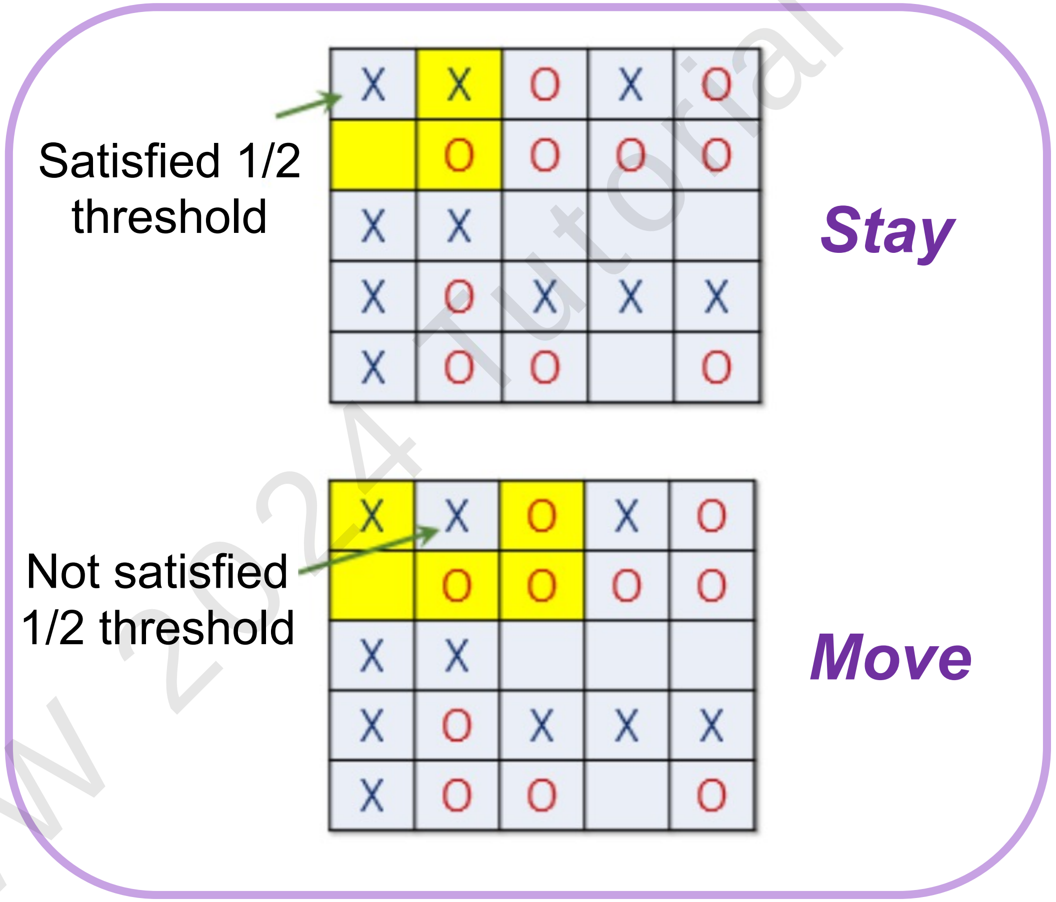
[2]Renmin University of China

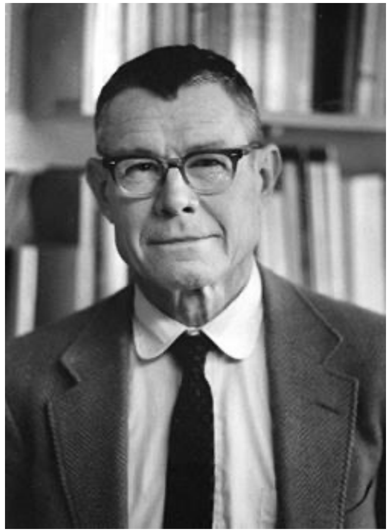
[3]University of Science and Technology of China

Simulate Urban Dynamics with LLM agents

One of the Earliest Agent-based Models: Urban Segregation





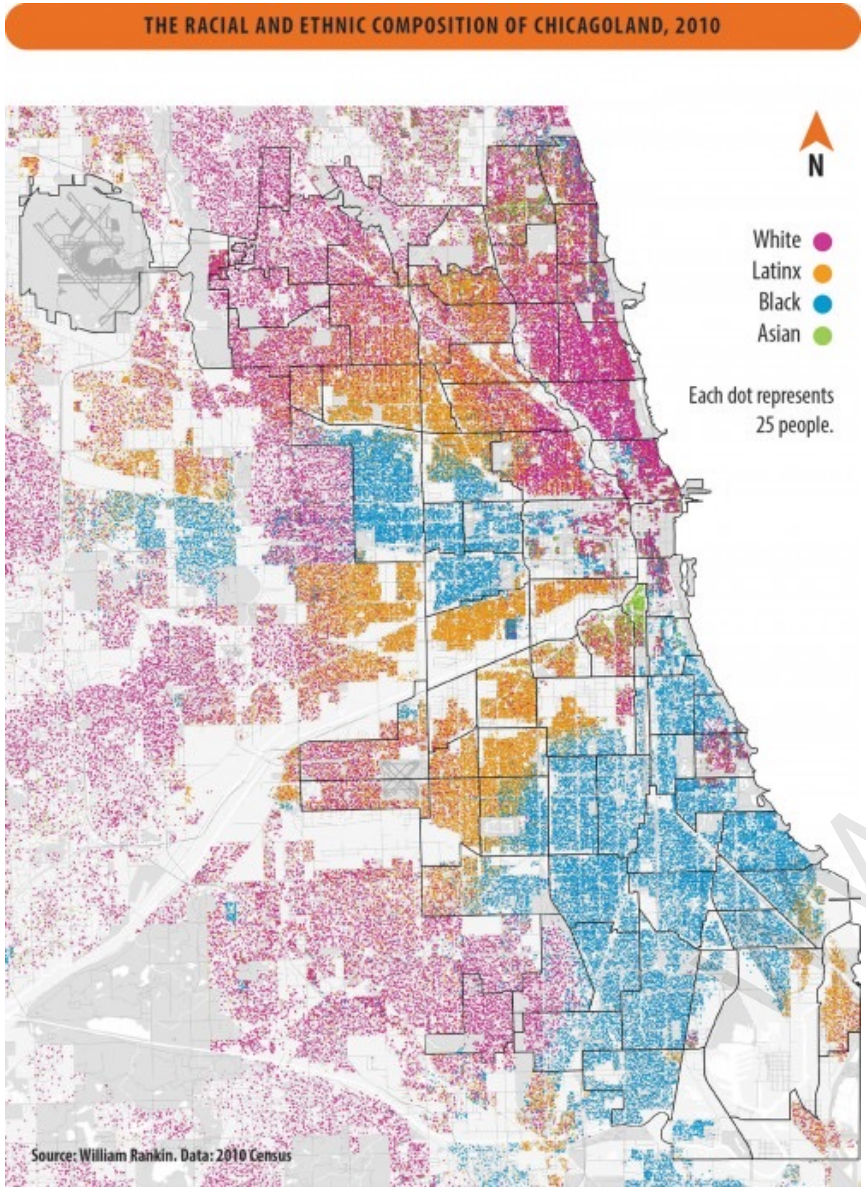


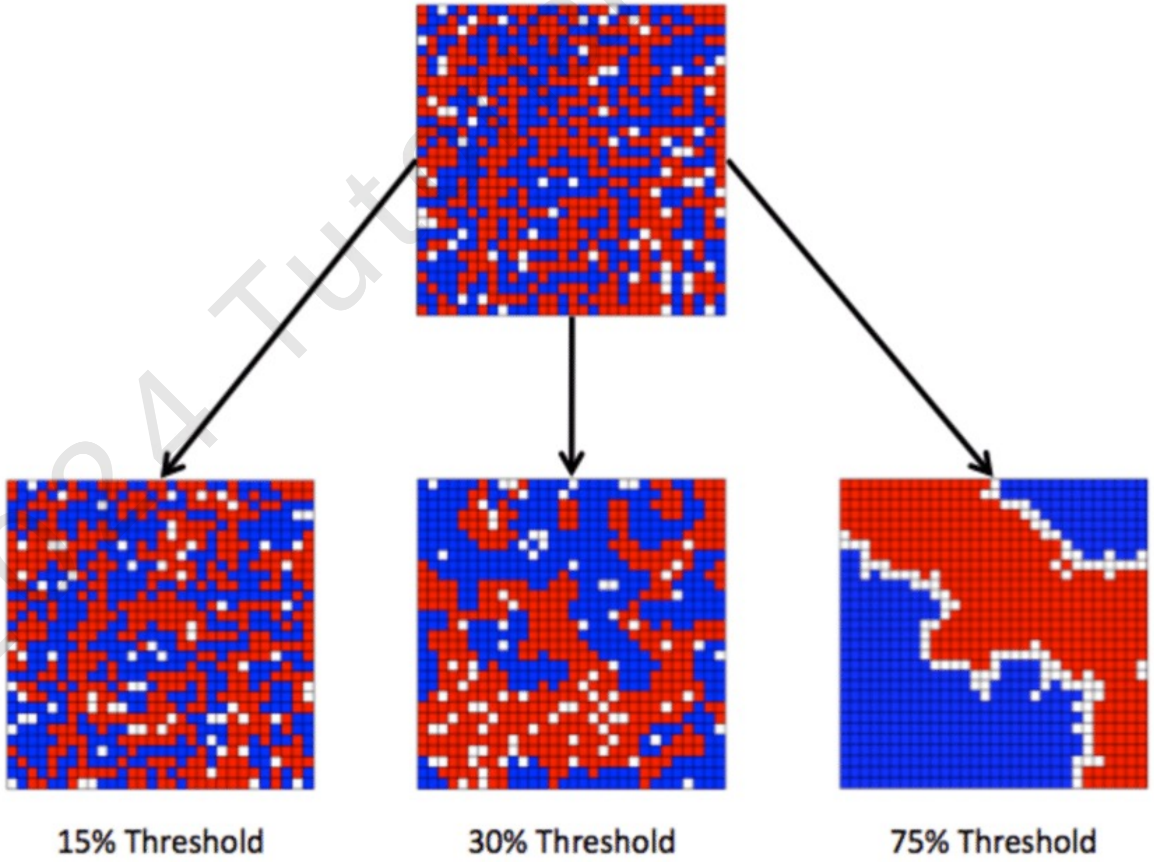
Thomas Schelling (1921 - 2016)

Using Coins and graph paper to simulate ***urban segregation*** with autonomous agents.

Schelling, Thomas C. "Dynamic models of segregation." *Journal of mathematical sociology* 1.2 (1971): 143-186.

One of the Earliest Agent-based Models: Urban Segregation



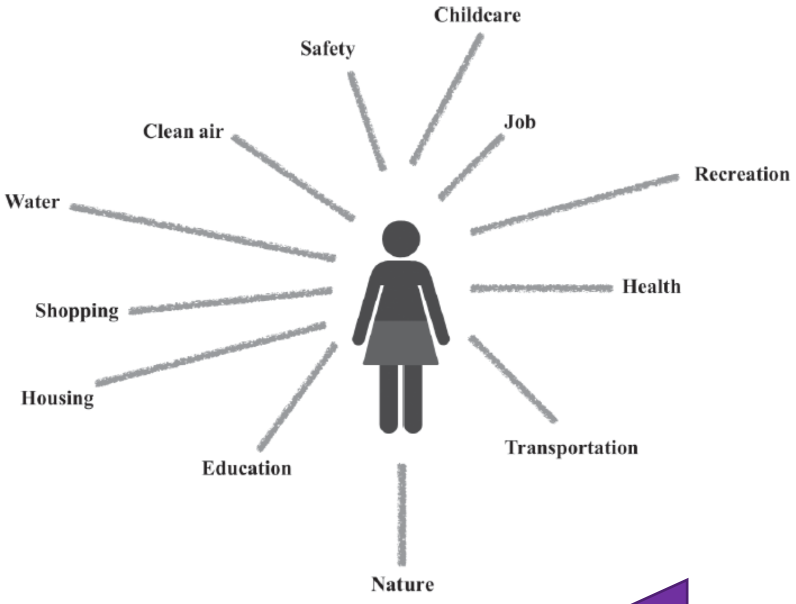


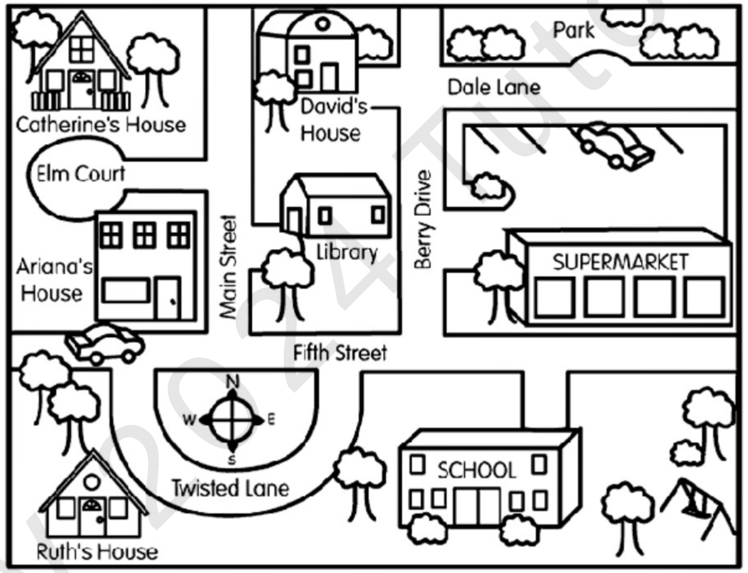
Mild preference could lead to the emergence of severe segregation at aggregate level.

Schelling, Thomas C. "Dynamic models of segregation." *Journal of mathematical sociology* 1.2 (1971): 143-186.

Agent-based Models in Urban Studies

Billions | **Millions** | **Thousands**

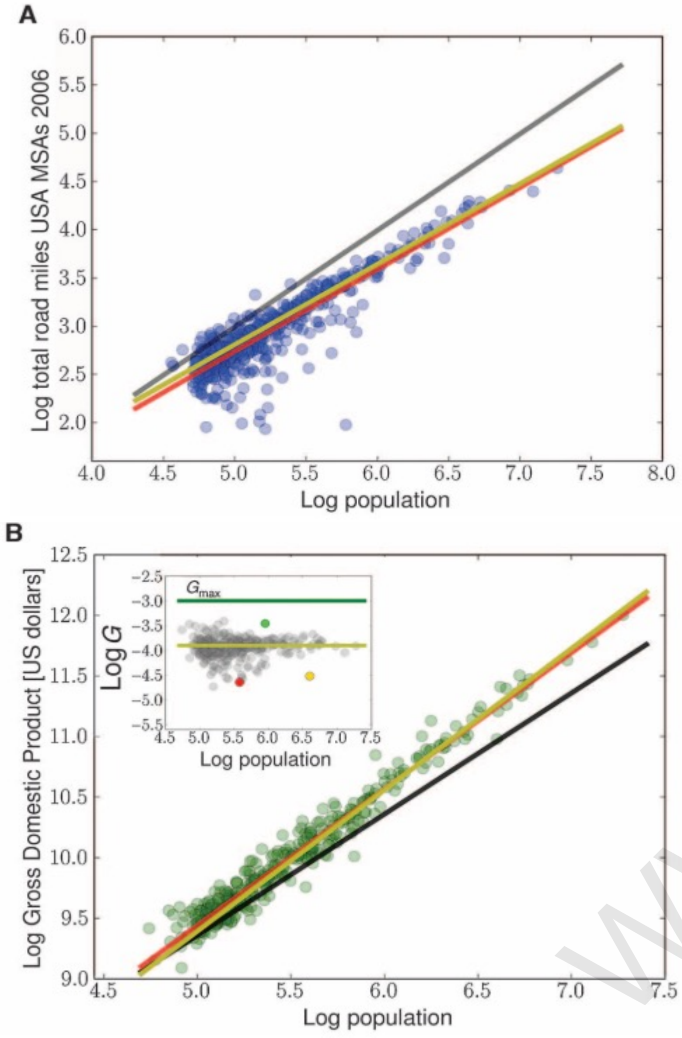




Micro ⟷ **Macro**

Agent-based Models in Cities: How **Complex but Universal** <u>macro patterns</u> emerge from the interaction of <u>individual agents</u>?

Agent-based Models in Urban Studies

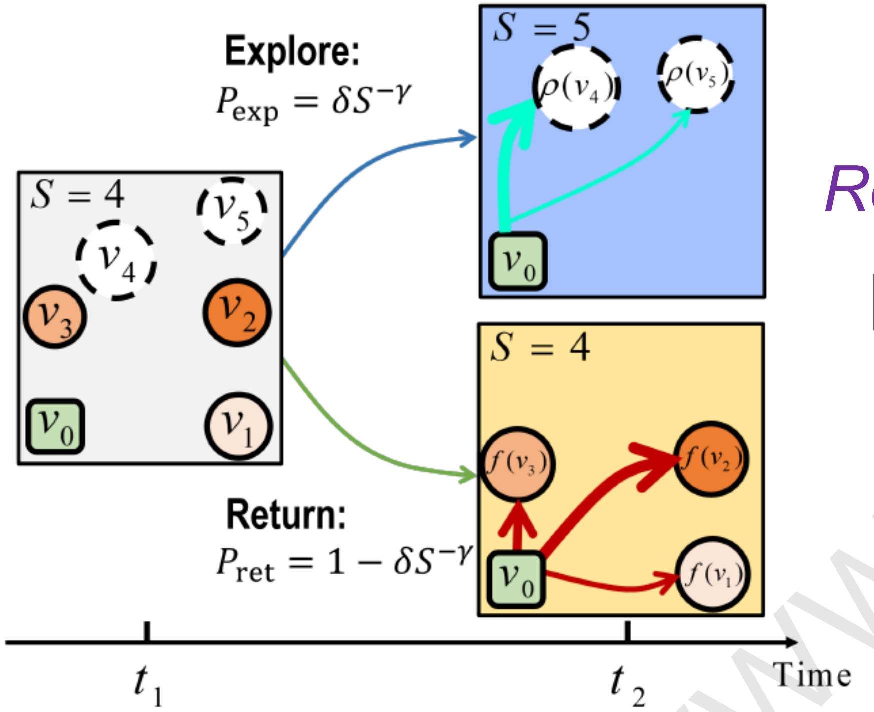


Urban scaling relations	Observed exponent range	Model ($D = 2$, $H = 1$)	Model D, H
Land area $A = aN^{\alpha}$	[0.56,1.04]	$\alpha = \frac{2}{3}$	$\alpha = \frac{D}{D+H}$
Network volume $A_n = A_0 N^{\nu}$	[0.74,0.92]	$\nu = \frac{5}{6}$	$\nu = 1 - \delta$
Network length $L_n = L_0 N^{\lambda}$	[0.55,0.78]	$\lambda = \frac{2}{3}$	$\lambda = \alpha$
Interactions per capita $\bar{I}_i = I_0 N^{\delta}$	[0.00,0.25]	$\delta = \frac{1}{6}$	$\delta = \frac{H}{D(D+H)}$
Socioeconomic rates $Y = Y_0 N^{\beta}$	[1.01,1.33]	$\beta = \frac{7}{6}$	$\beta = 1 + \delta$
Network power dissipation $W = W_0 N^{\omega}$	[1.05,1.17]	$\omega = \frac{7}{6}$	$\omega = 1 + \delta$
Average land rents $P_L = P_0 N^{\delta_L}$	[0.46,0.52]	$\delta_L = \frac{1}{2}$	$\delta_L = 1 - \alpha + \delta$

Researchers have long sought to design agent-based model to explain the empirical laws in cities.

Bettencourt, Luís MA. "The origins of scaling in cities." science 340.6139 (2013): 1438-1441.

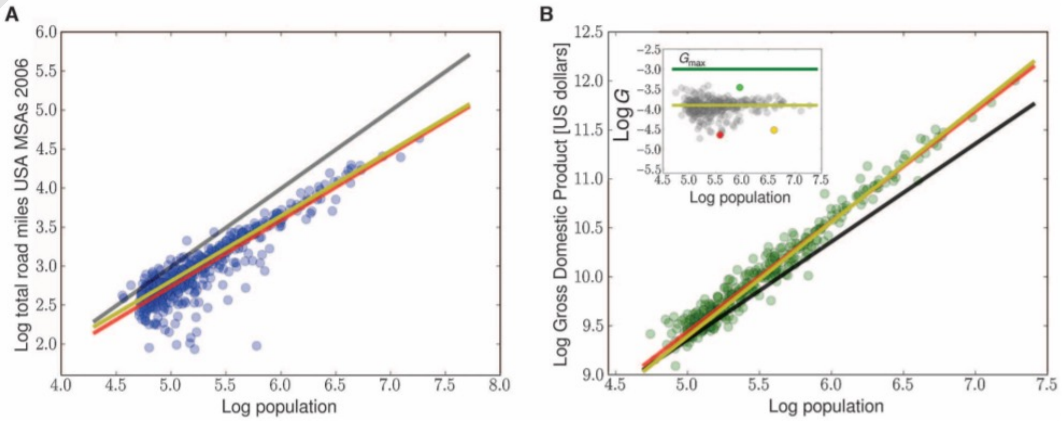
Agent-based Models in Urban Studies



Agent-based Models for Urban Mobility

Reproduce

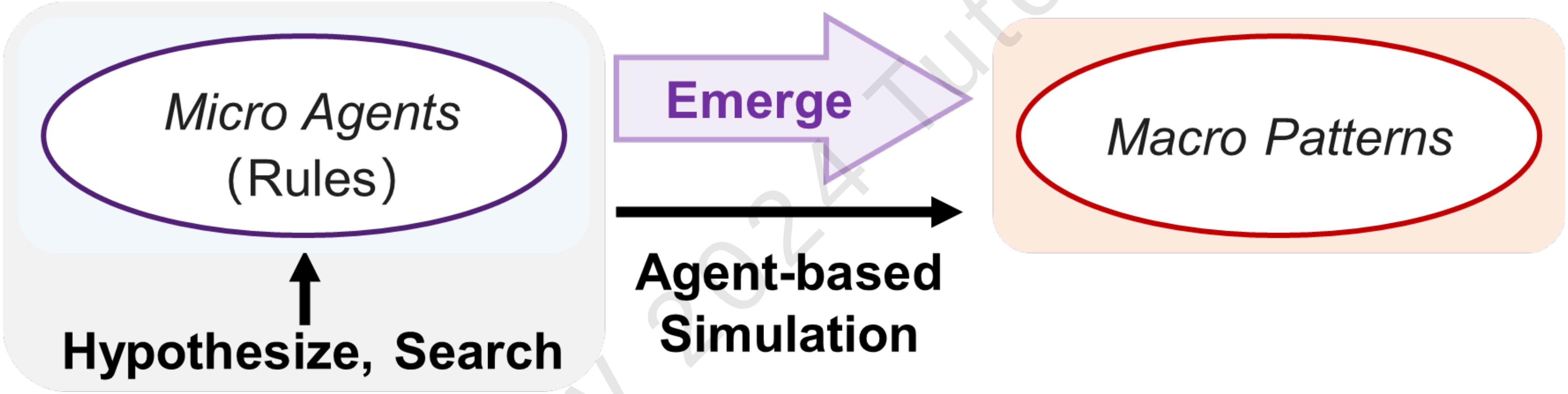
Urban scaling relations	Observed exponent range	Model ($D = 2$, $H = 1$)	Model D, H
Land area $A = aN^{\alpha}$	[0.56,1.04]	$\alpha = \frac{2}{3}$	$\alpha = \frac{D}{D+H}$
Network volume $A_n = A_0 N^{\nu}$	[0.74,0.92]	$\nu = \frac{5}{6}$	$\nu = 1 - \delta$
Network length $L_n = L_0 N^{\lambda}$	[0.55,0.78]	$\lambda = \frac{2}{3}$	$\lambda = \alpha$
Interactions per capita $\bar{I}_i = I_0 N^{\delta}$	[0.00,0.25]	$\delta = \frac{1}{6}$	$\delta = \frac{H}{D(D+H)}$
Socioeconomic rates $Y = Y_0 N^{\beta}$	[1.01,1.33]	$\beta = \frac{7}{6}$	$\beta = 1 + \delta$
Network power dissipation $W = W_0 N^{\omega}$	[1.05,1.17]	$\omega = \frac{7}{6}$	$\omega = 1 + \delta$
Average land rents $P_L = P_0 N^{\delta_L}$	[0.46,0.52]	$\delta_L = \frac{1}{2}$	$\delta_L = 1 - \alpha + \delta$



Scaling Laws in Cities

Xu, Fengli, et al. "Emergence of urban growth patterns from human mobility behavior." *Nature Computational Science* 1.12 (2021): 791-800.

The Classic Research Paradigm of Rule-based Agents in City

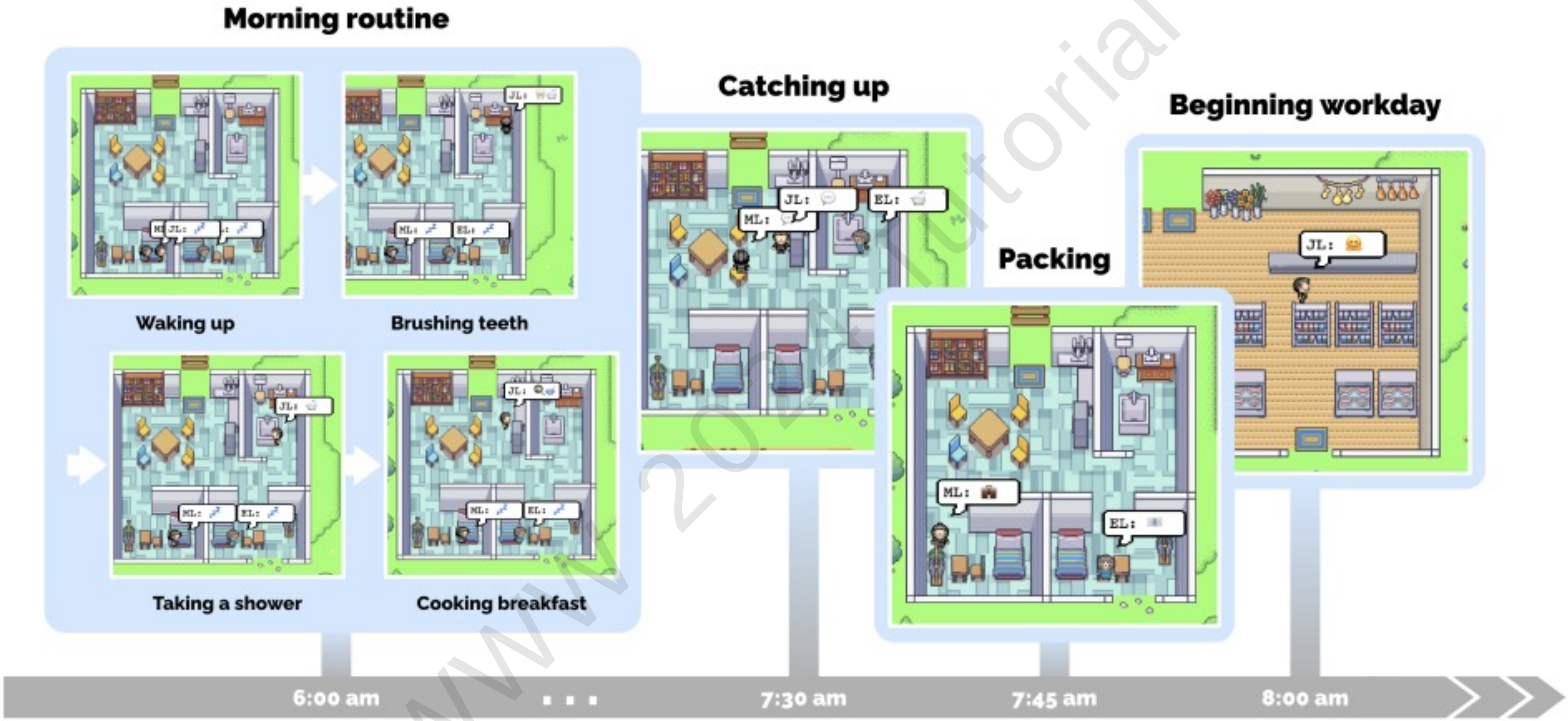


Rule-based agents are used as the proxies of micro rules/mechanisms, examining whether they can explain interested macro patterns.

LLM-empowered Generative Agents

Park, Joon Sung, et al. "Generative agents: Interactive simulacra of human behavior." UIST. 2023.

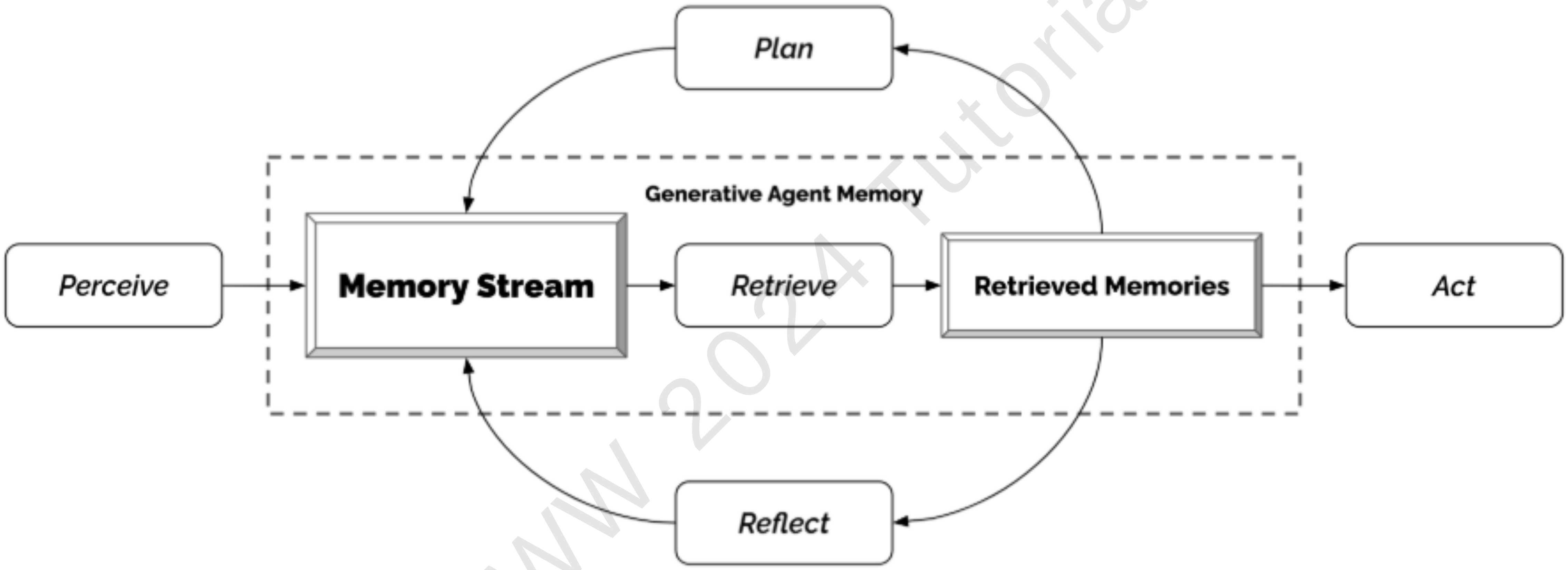
LLM-empowered Generative Agents



Leveraging the common sense reasoning power of LLMs to simulate 25 agents in a virtual town.

Park, Joon Sung, et al. "Generative agents: Interactive simulacra of human behavior." UIST. 2023.

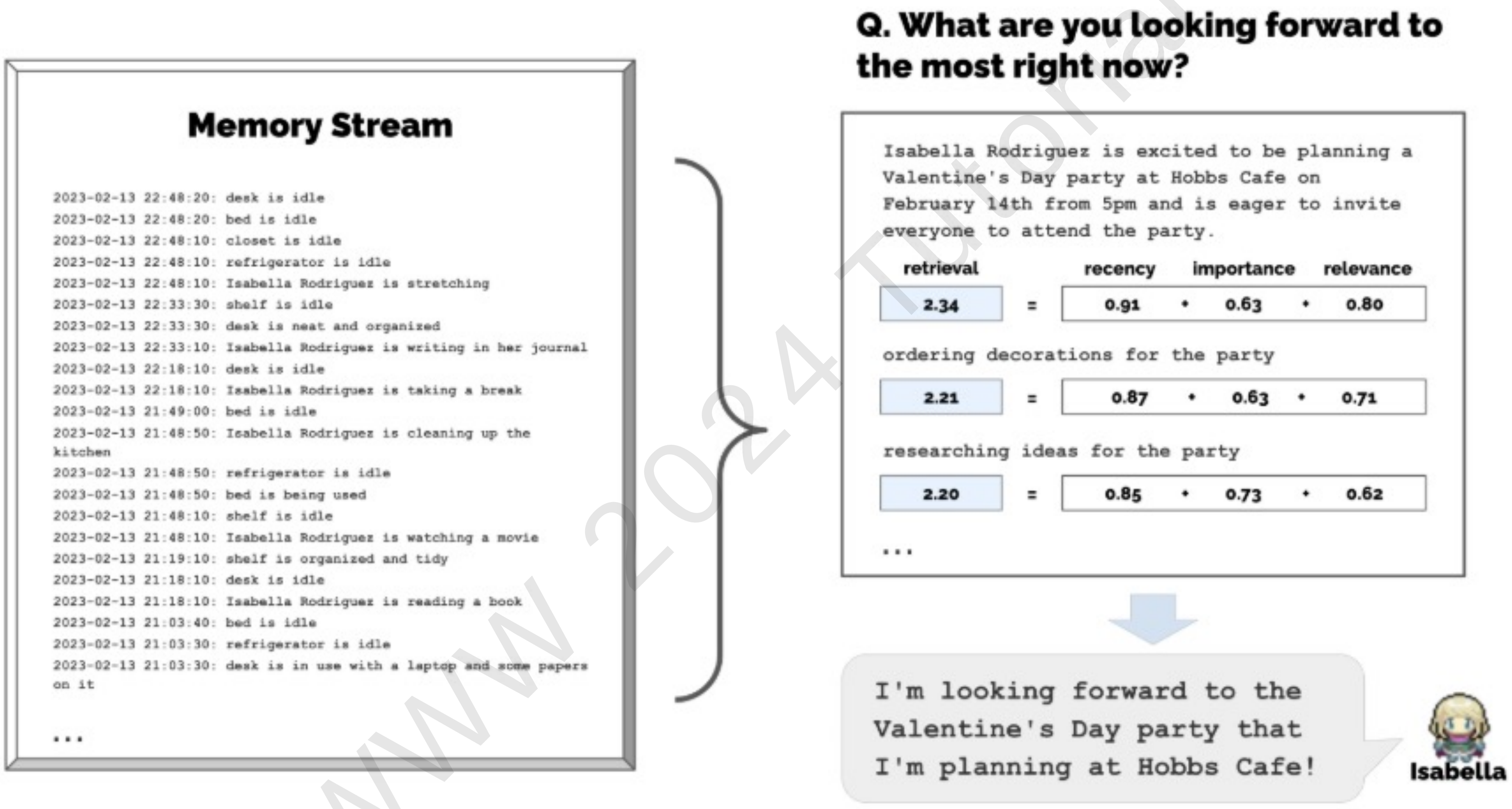
LLM-empowered Generative Agents



The architecture of generative agents, five modules: ***Plan***, ***Perceive***, ***Reflect***, ***Act***, ***Retrieve***

Park, Joon Sung, et al. "Generative agents: Interactive simulacra of human behavior." UIST. 2023.

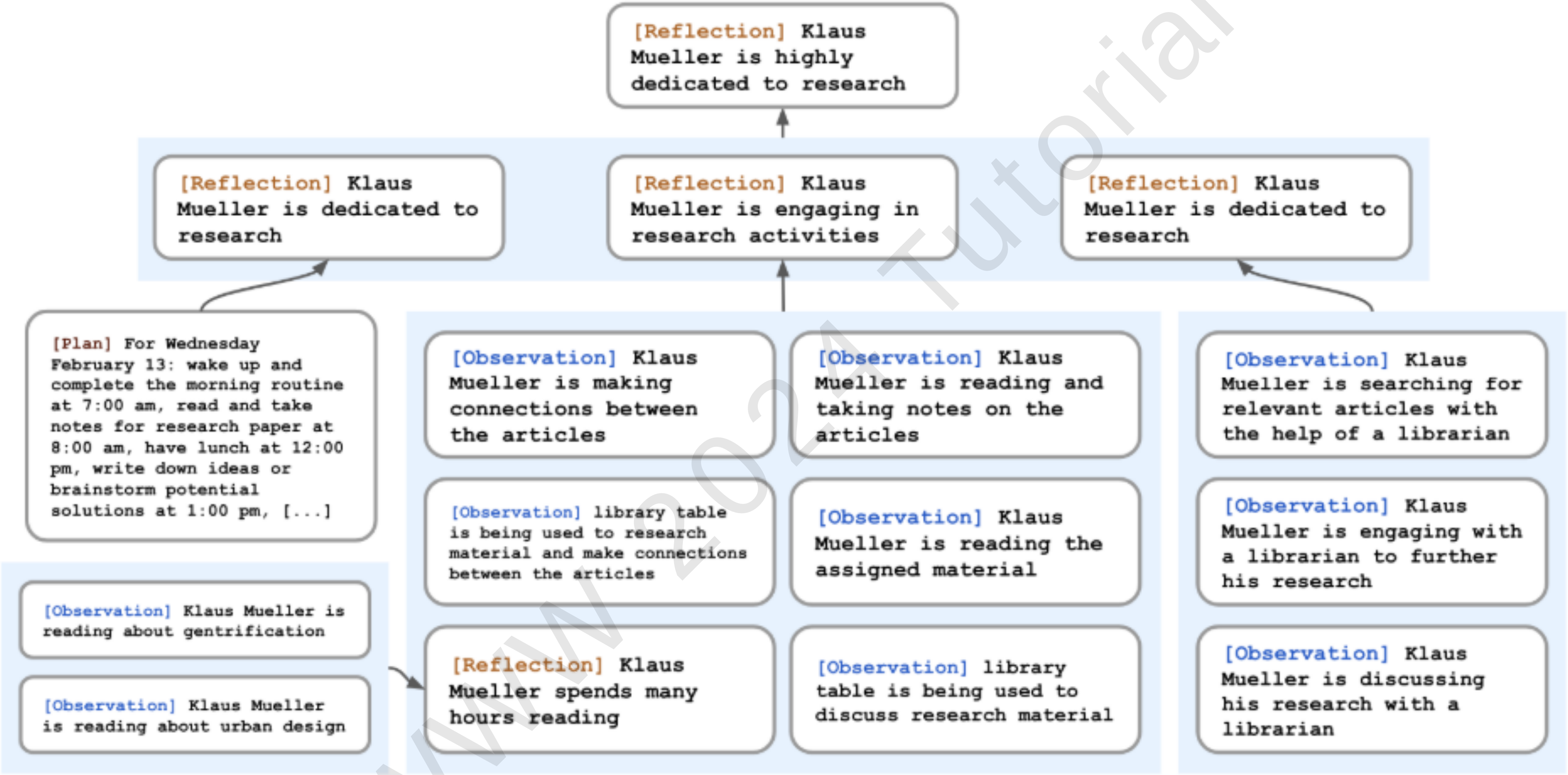
LLM-empowered Generative Agents



Retrieve relevant memories to augment agent behavior simulation.

Park, Joon Sung, et al. "Generative agents: Interactive simulacra of human behavior." UIST. 2023.

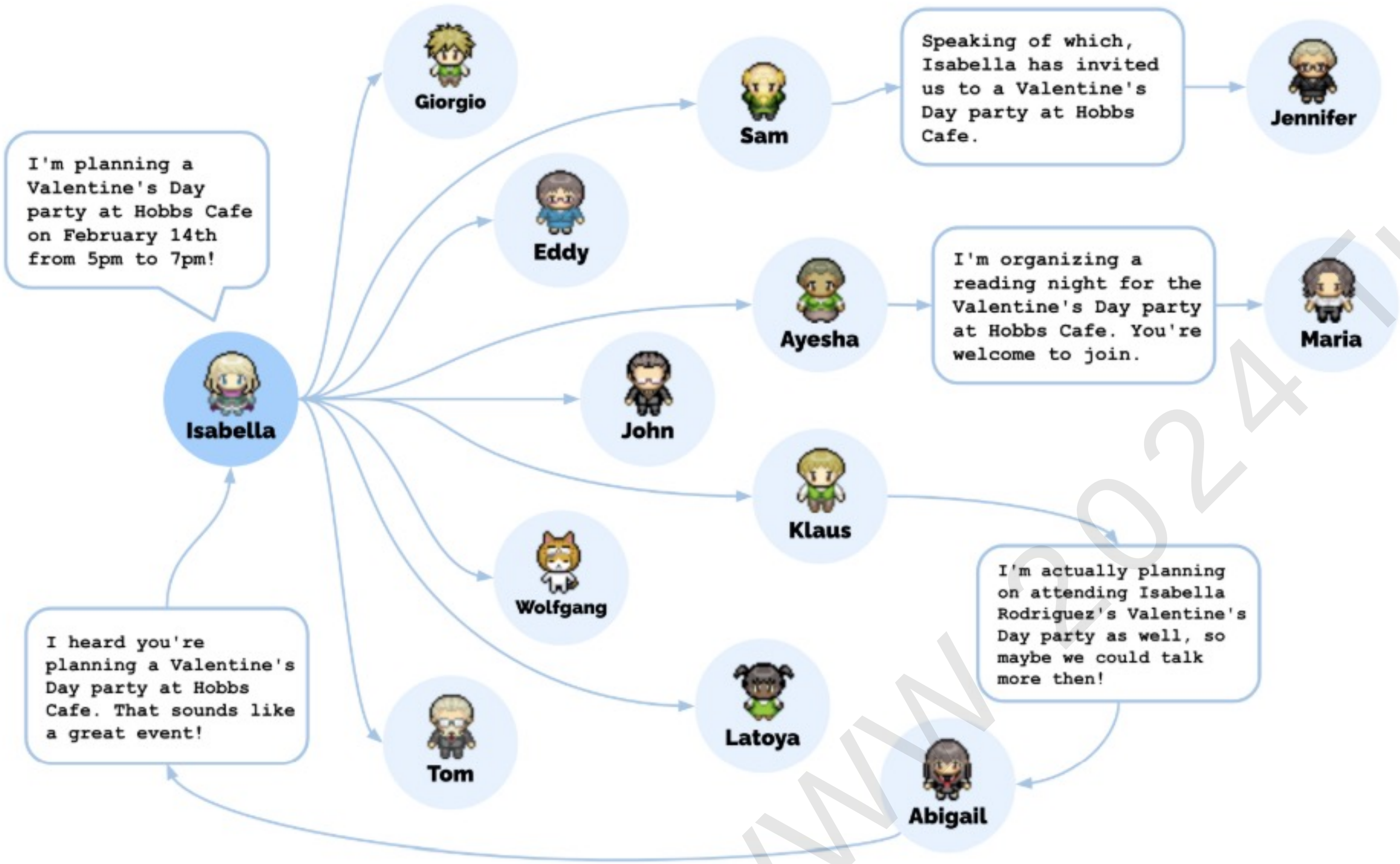
LLM-empowered Generative Agents

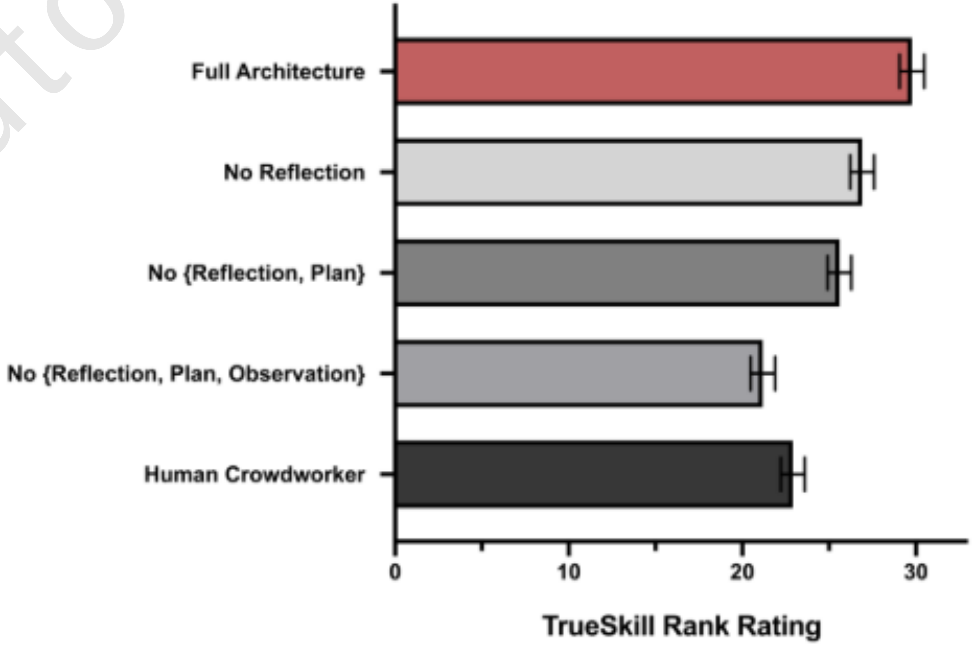


Reflect periodically summarizes low-level memories into high-level, abstract thoughts.

Park, Joon Sung, et al. "Generative agents: Interactive simulacra of human behavior." UIST. 2023.

LLM-empowered Generative Agents

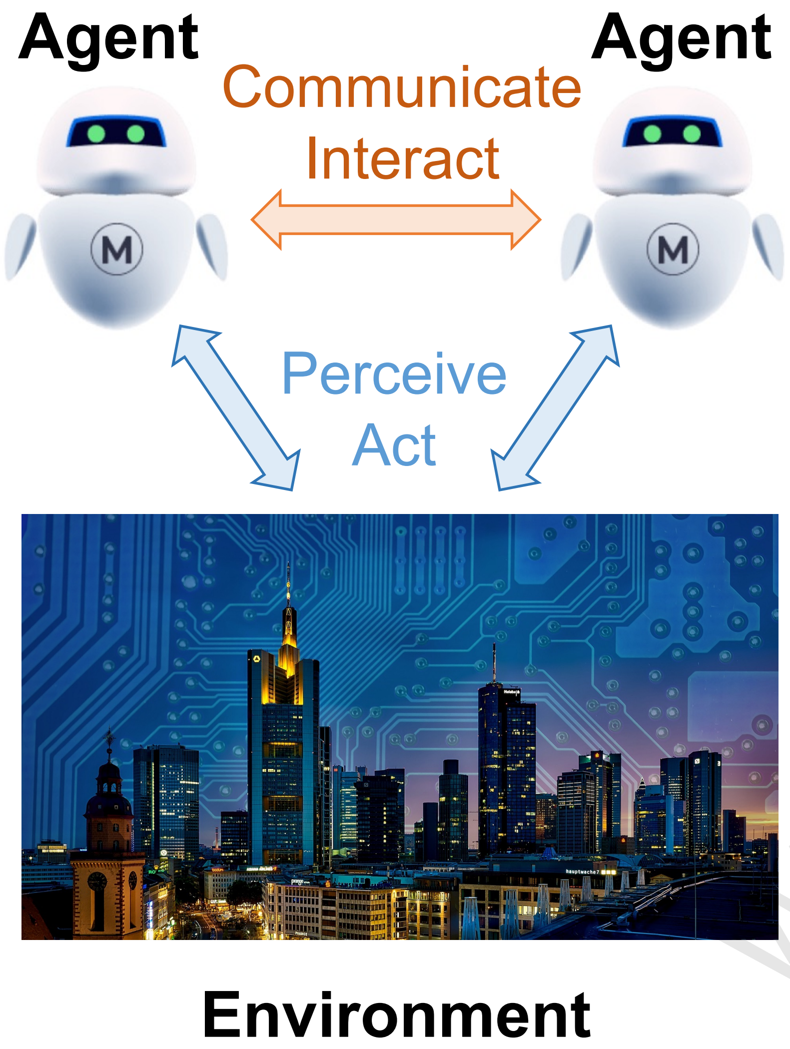


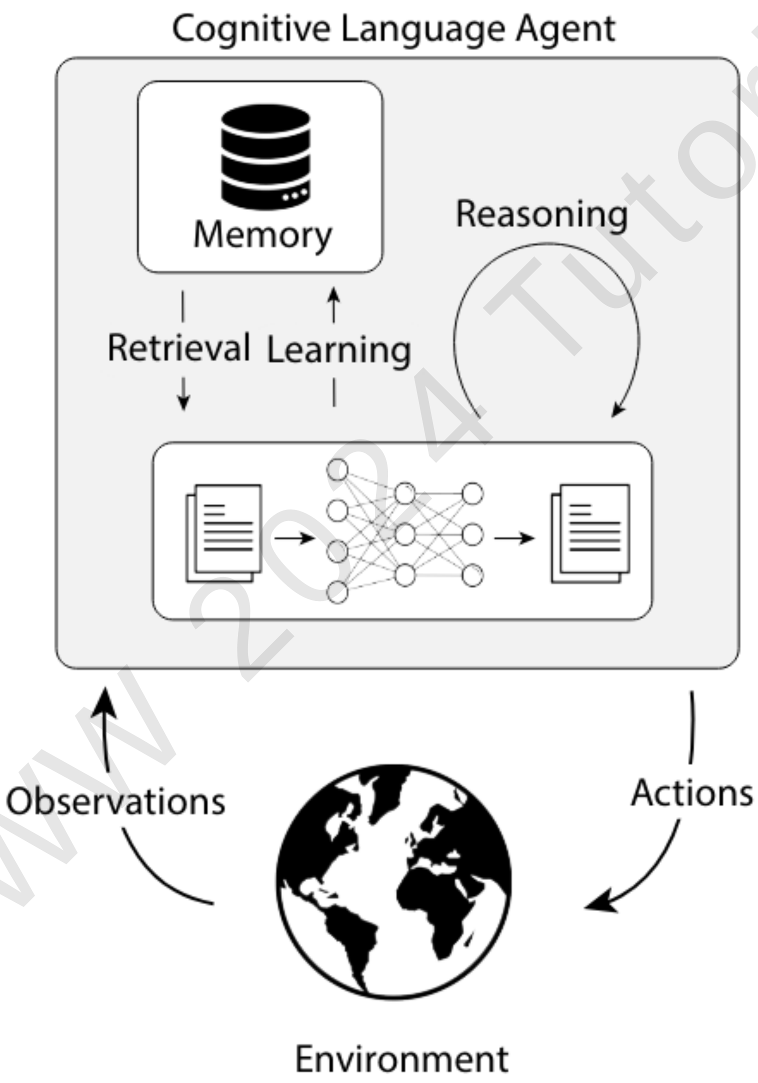


Experiments show these generative agents can: 1) simulate human activities, such as a Valentine's part; 2) produces more believable behavior than human crowdworkers.

Park, Joon Sung, et al. "Generative agents: Interactive simulacra of human behavior." UIST. 2023.

What is a LLM Agent?



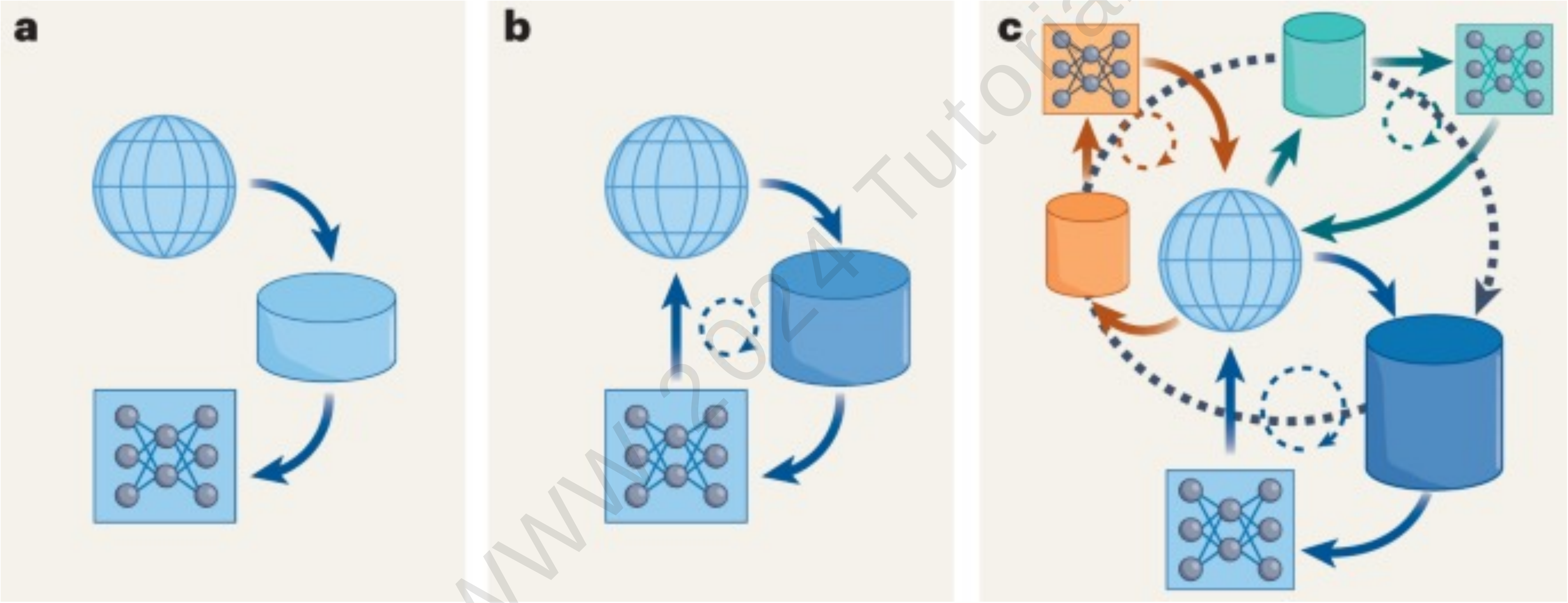


1. Reflection
2. Tool use
3. Planning
4. Multi-agent collaboration

@ Andrew Ng

Sumers, Theodore R., et al. "Cognitive architectures for language agents." arXiv preprint arXiv:2309.02427 (2023).

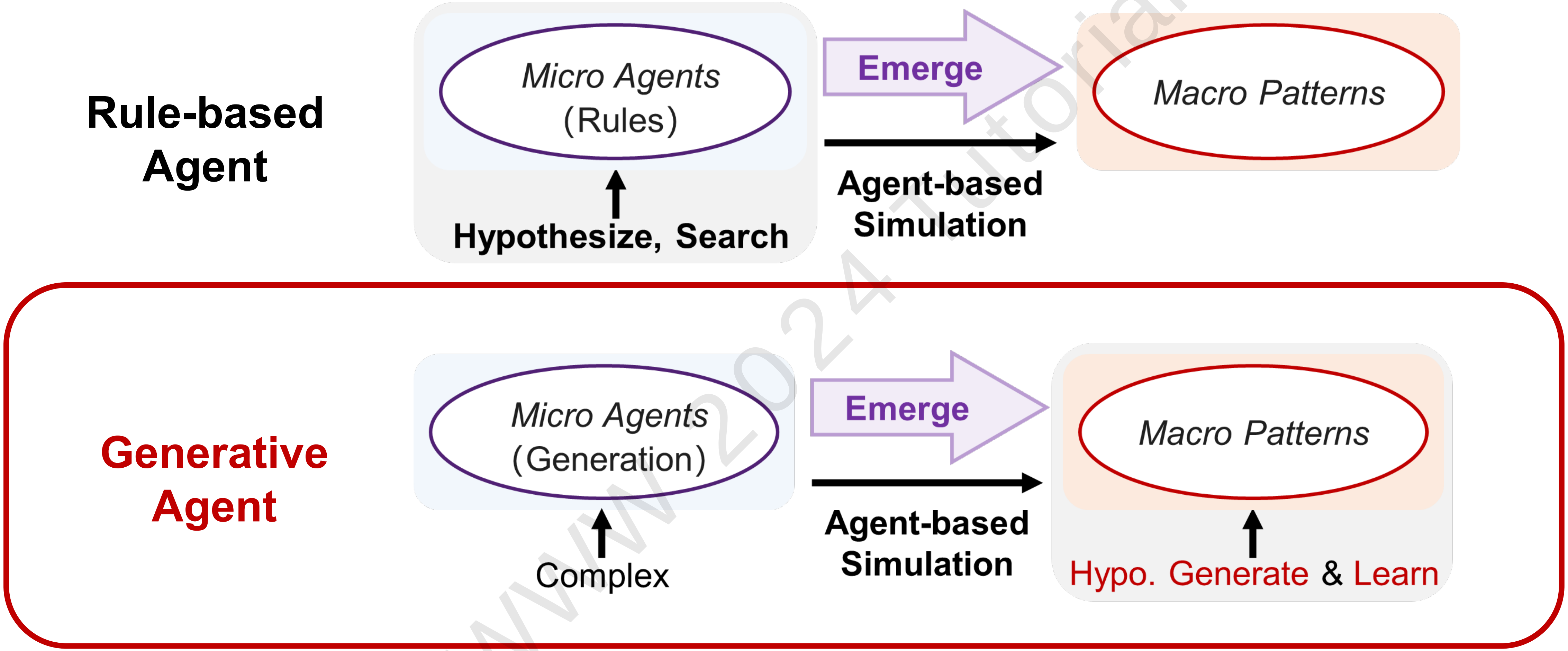
LLM Agents Could Be Novel Data Sources



Web AI **Reinforced AI** **Interacting Agents**

Duéñez-Guzmán, Edgar A., et al. "A social path to human-like artificial intelligence." Nature Machine Intelligence 5.11 (2023): 1181-1188.

Paradigm Shift



- ☐ **Data Generation:** travel survey, business site selection, policy evaluation, etc.
- ☐ **Learning Environment:** reinforced with feedback, *e.g.*, city navigation, scheduling. Providing a benchmark environment.

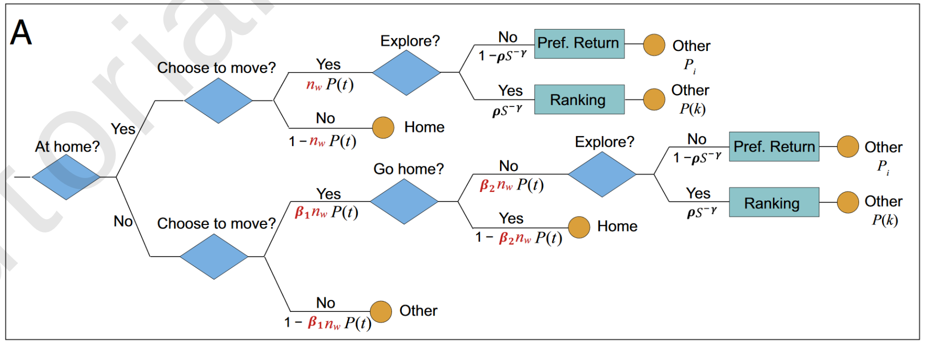
Challenges

- **Interact with Complex Urban Environment**
  - Millions of point-of-interests (places) in a city
- **Multi-modal Urban Experiences**
  - Real-world urban experiences are beyond text
- **Cost of Scaling Up**
  - LLM simulation is quite expensive

Interact with Urban Places (Mobility Behavior)

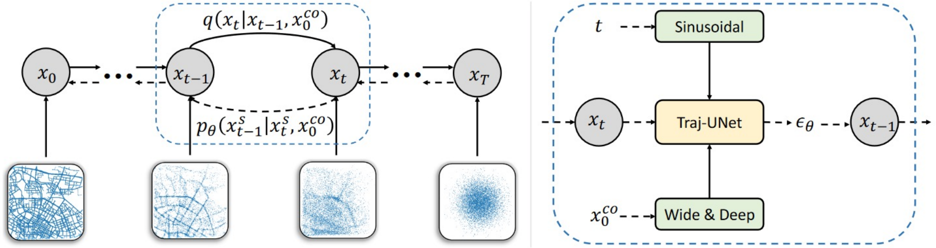
Parameterized simple rules-based:

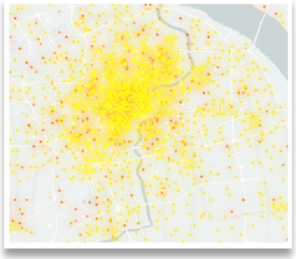
- Use **stochastic processes** to model
- decision making processes: TimeGeo



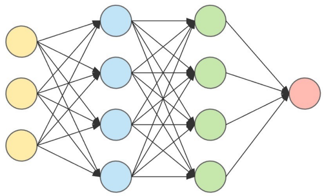
Deep Generative Models-based:

- **GAN-based**： MoveSim
- **VAE-based**： Volunteer
- **Diffusion-based**： DiffTraj

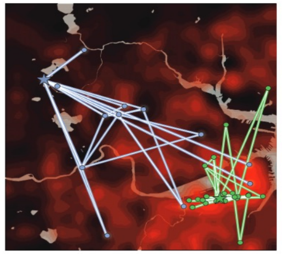




Fitting

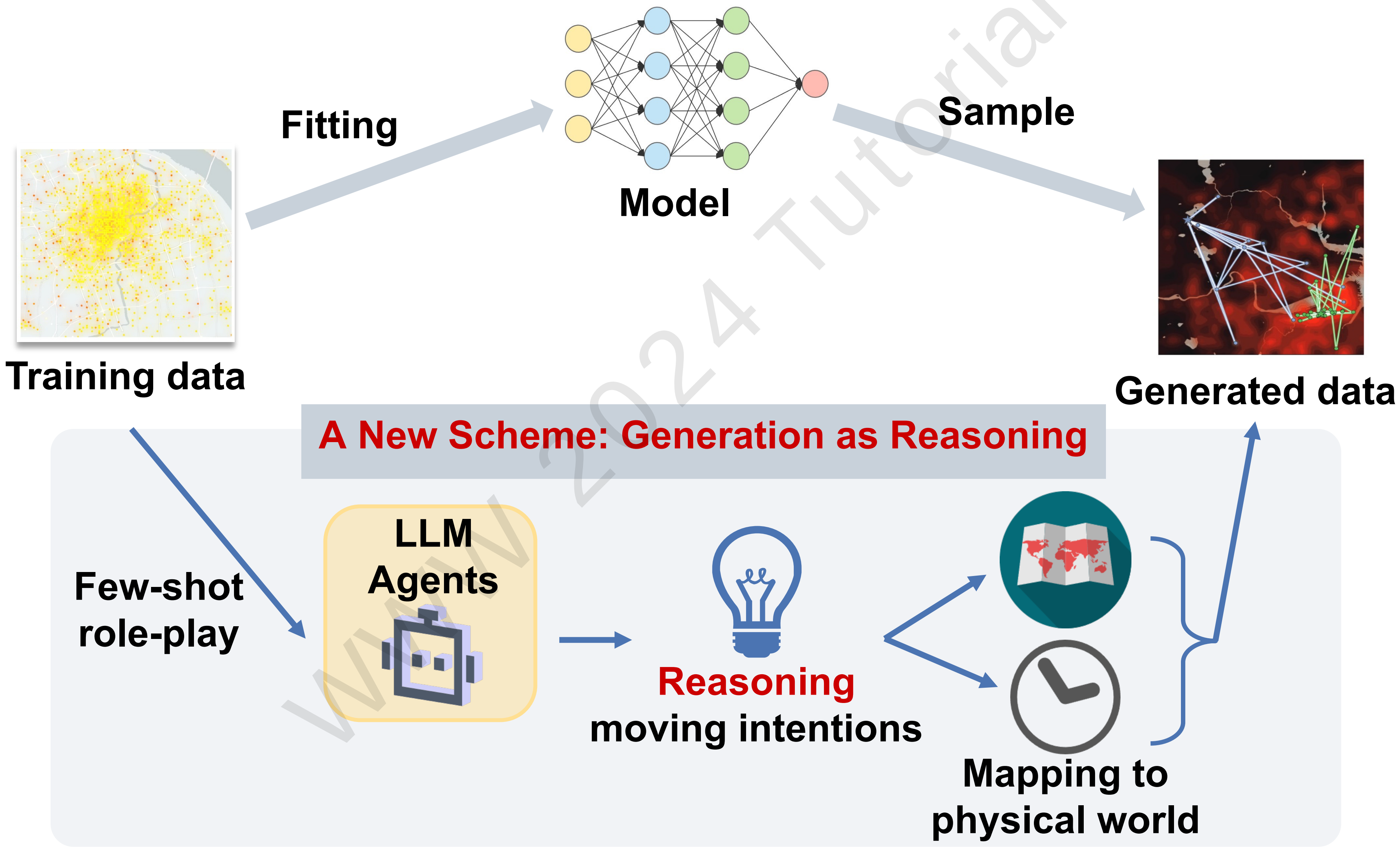


Sample



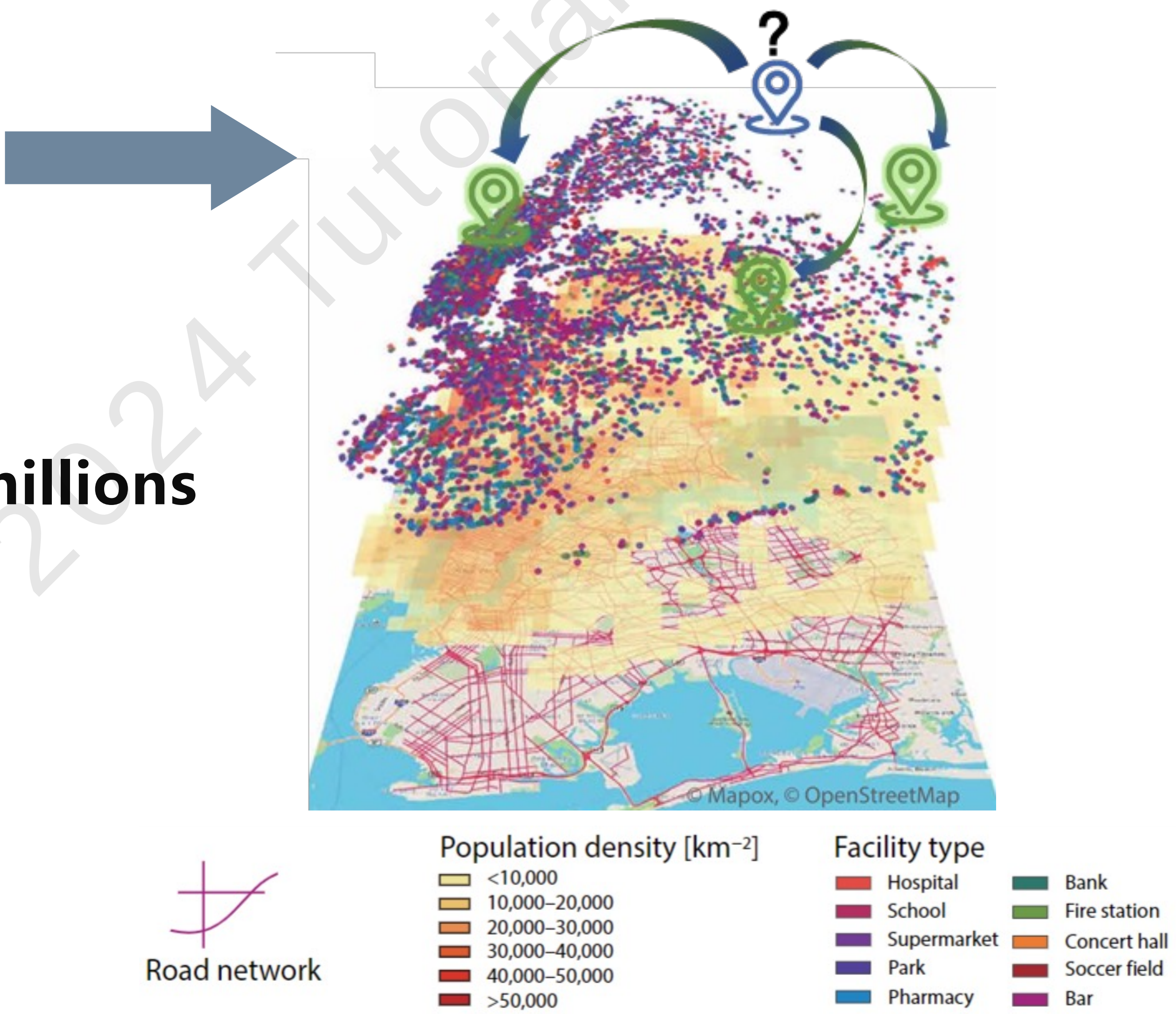
Massive training data **Model** **Generated data**

Mobility Generation as Reasoning



From Reasoning to Behaviour

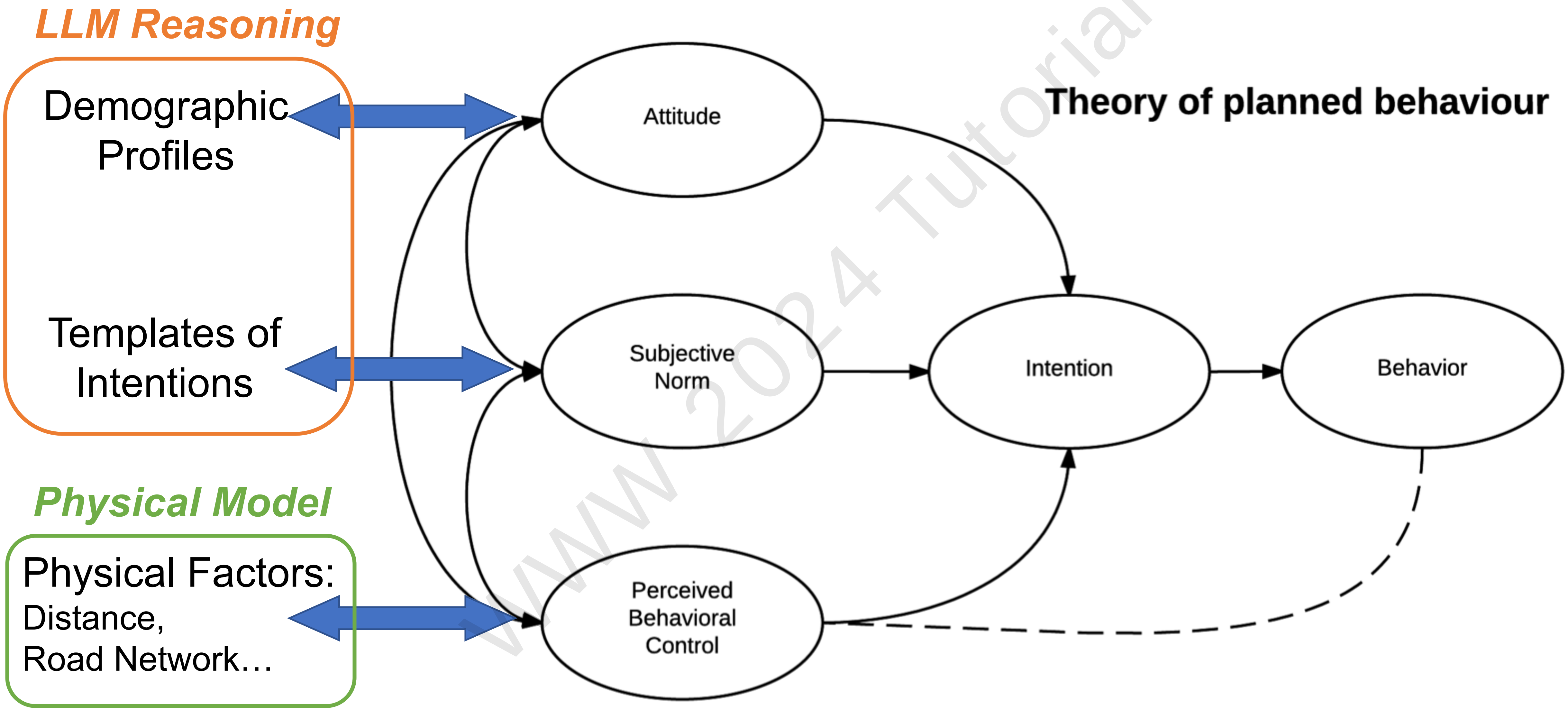
LLM Agents



Difficult to choose from millions of urban places:

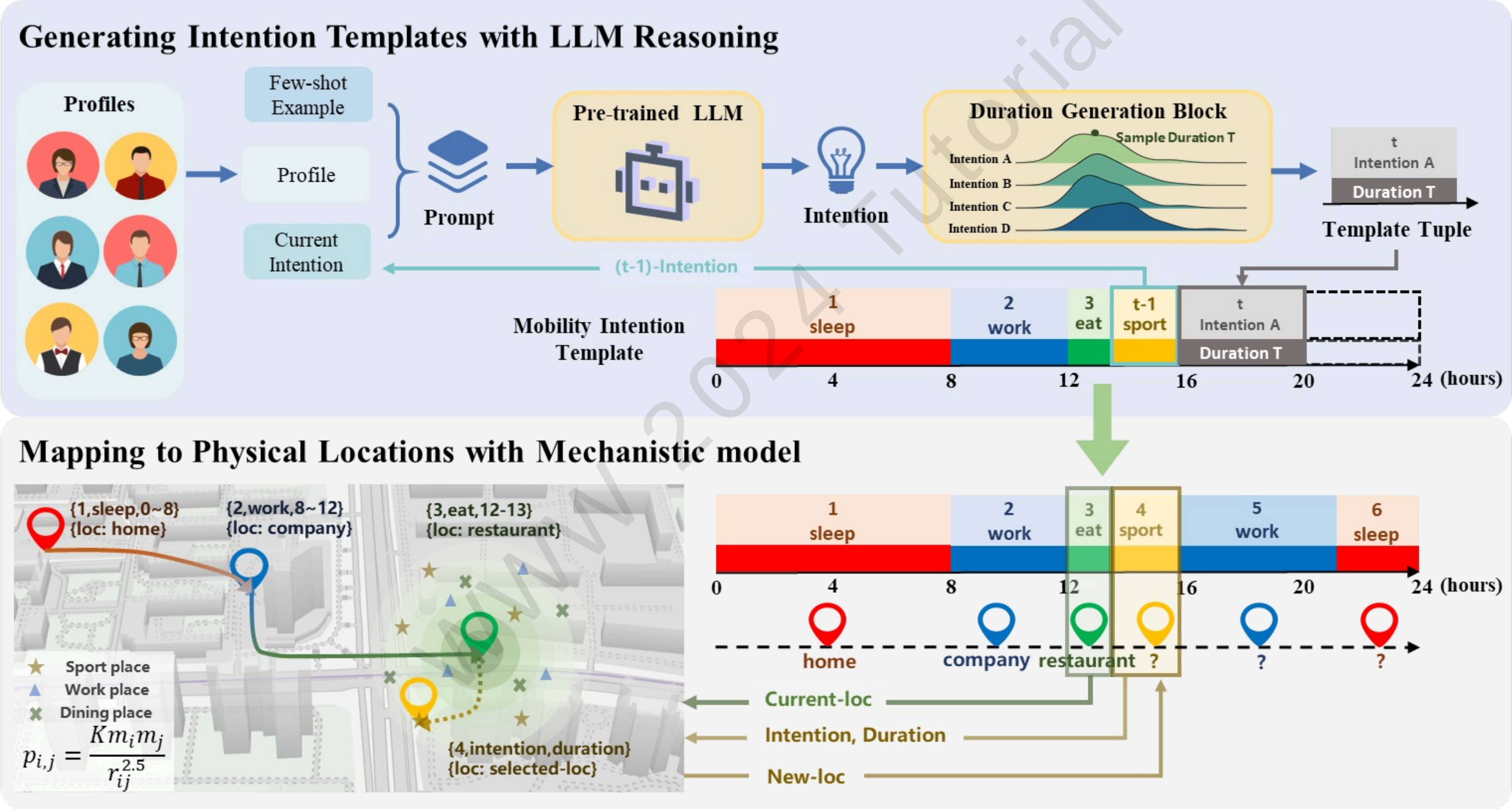
- Limited context lengths
- Not good geospatial data
- Expensive inference cost

Theory of Planned Behaviour



Armitage, Christopher J., and Mark Conner. "Efficacy of the theory of planned behaviour: A meta-analytic review." British journal of social psychology 40.4 (2001): 471-499.

Grounding LLM Reasoning with Physical Mobility Models



Generating Templates of Mobility Intentions

Few-shot Role-play:

- Demographic profiles: (Gender/education/ consumption level/occupation)
- manually annotated in-context learning examples of intention

Example construction

Demographic Profiles
Intent sequences (human labeled)
Reasoning process behind each decision (human labeled)

Chain Of Thought Reasoning :

× Generate a complete sequence of intents at once

✓ Reasoning intent sequence step by step: → chain of intention and context

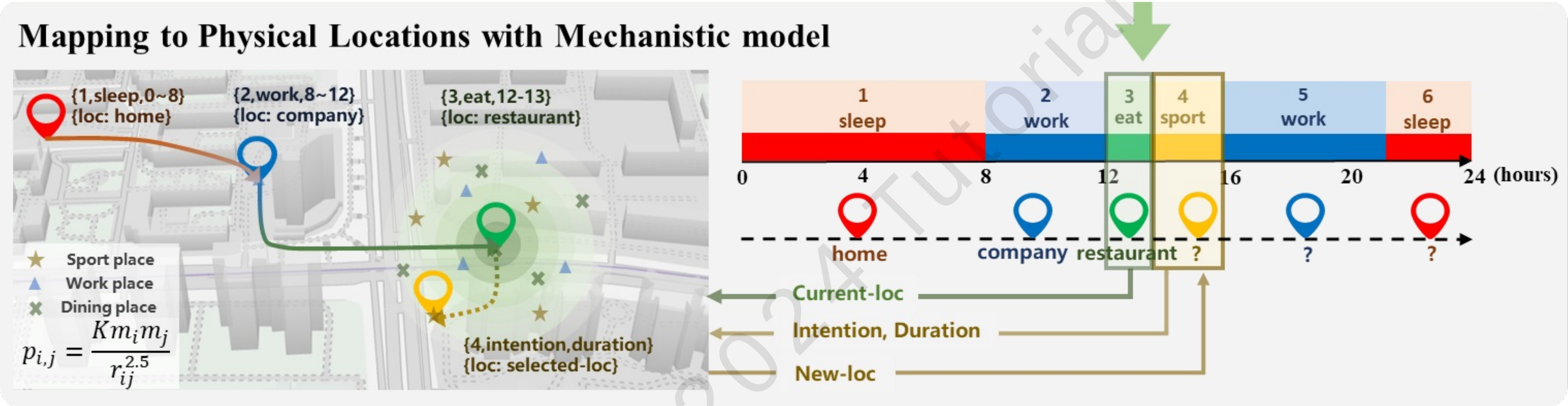
This is the daily schedule of a programmer who works at an IT company.

…………

["eat", "(12:35, 13:01)"], (Reason: Already noon. Time for lunch.)

["go to work", "(13:15, 22:07)"], (Reason:After lunch, afternoon and evening are the main working hours. Overtime is needed to make up for the lack of working time.) …………

Mapping Intentions to Physical Behaviour



- Match intents to physical locations by **gravity model**
  - Transfer probability between two regions: $P_{ij} = Km_i m_j f(r_{ij})$
  - Granularity of zoning: concentric rings centred on the previous location
  - m_i, m_j: POI density in the ring， $f(r_{ij}) = r^{-2.5}$
  - Maximum search distance : 10km

Experiments – Data Quality

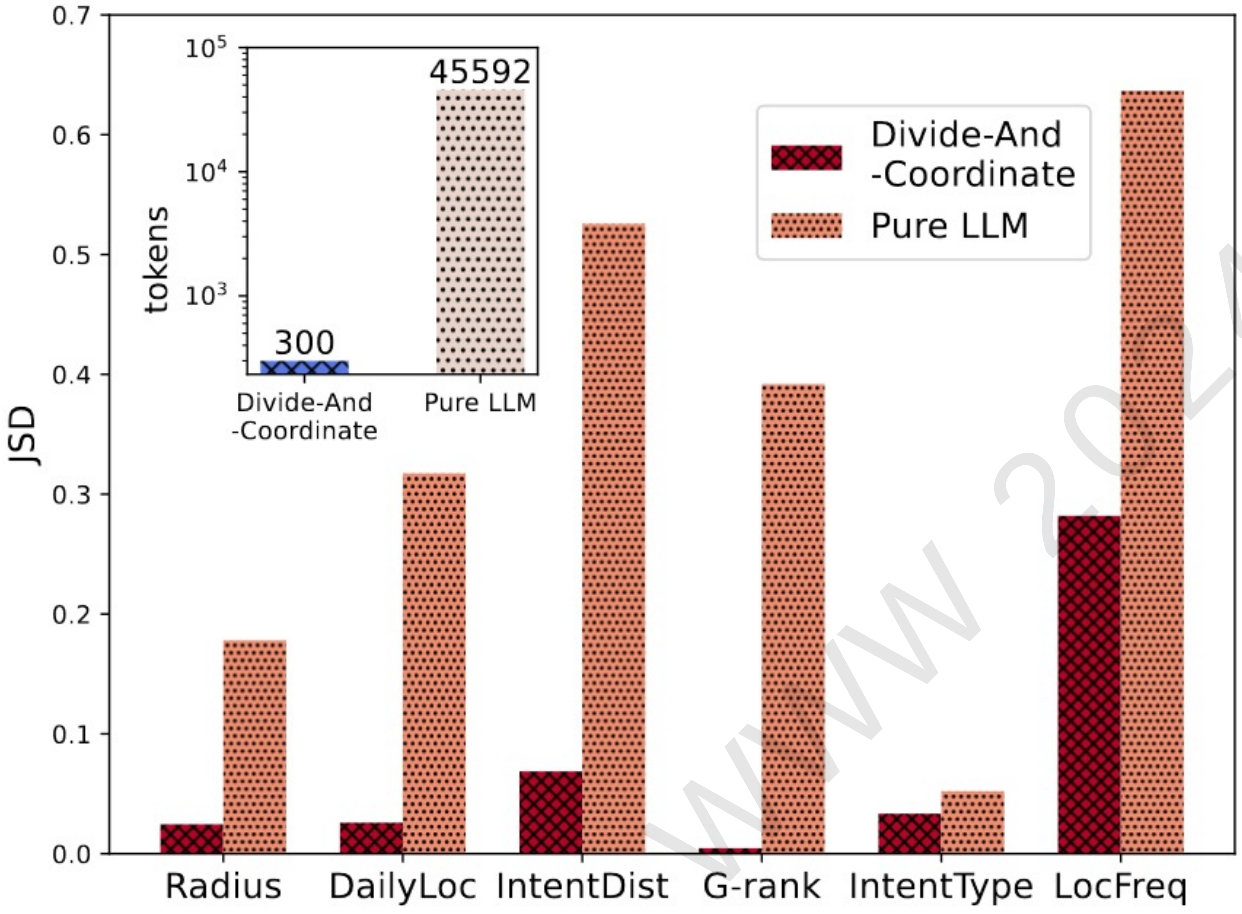
Model	Statistical				Semantic		Aggregated	
	Radius ↓	DailyLoc ↓	IntentDist ↓	G-rank ↓	IntentAcc ↑	IntentType ↓	LocFreq ↓	ODSim↓
TimeGeo	0.2592	0.2513	0.2040	<u>0.0176</u>	<u>0.4639</u>	0.1545	0.6931	7.38E-05
MoveSim	0.2235	<u>0.0521</u>	<u>0.1010</u>	0.0244	0.0956	0.1781	0.6384	<u>5.93E-05</u>
VOLUNTEER	0.5116	0.0560	0.3296	0.0213	0.1906	0.1620	0.2956	6.00E-05
DiffTraj	<u>0.0284</u>	0.6931	0.6931	0.0286	0.4035	<u>0.0804</u>	<u>0.2872</u>	6.38E-05
MobiGeaR(Ours)	**0.0245**	**0.0259**	**0.0158**	**0.0046**	**0.7526**	**0.0334**	**0.2820**	**5.45E-05**

- **Baselines**：
  - Mechanism models：TimeGeo
  - Deep generative models：
    - GAN-based：MoveSim
    - VAE-based：Volunteer
    - Diffusion-based：DiffTraj

- **Evaluation dimensions**：
  - Classic statistical indicators
    - time、distance、frequency
  - Group Aggregation Authenticity
    - OD matrix error
    - frequency of visits to all grid points
  - **Semantic Accuracy**

> Improve all performance metrics by **44%** on average.
> Intent accuracy was significantly improved by **62.23%**.
> Training data was reduced from **100k** to **200**.

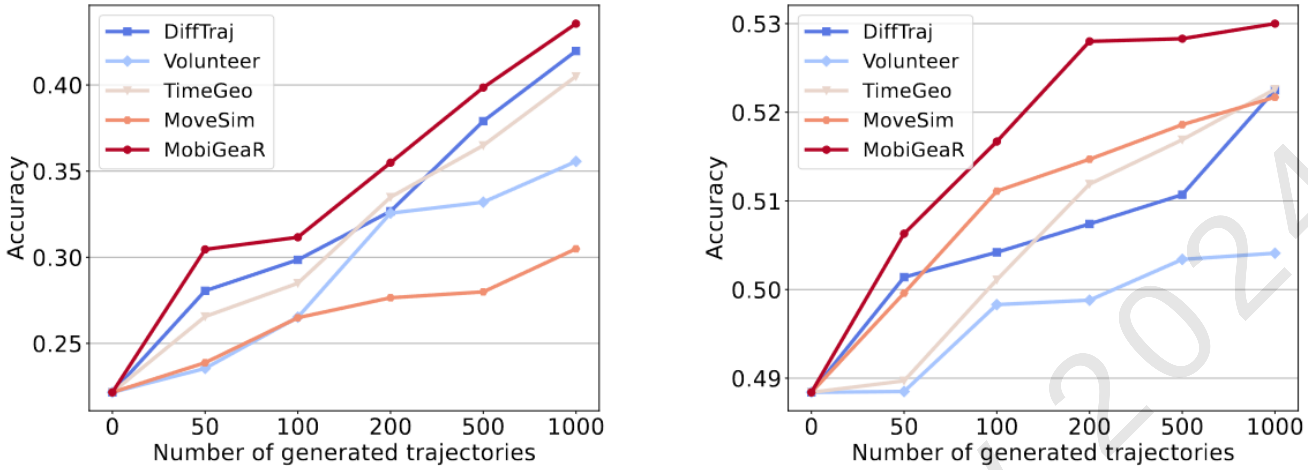
Experiments – Token Cost



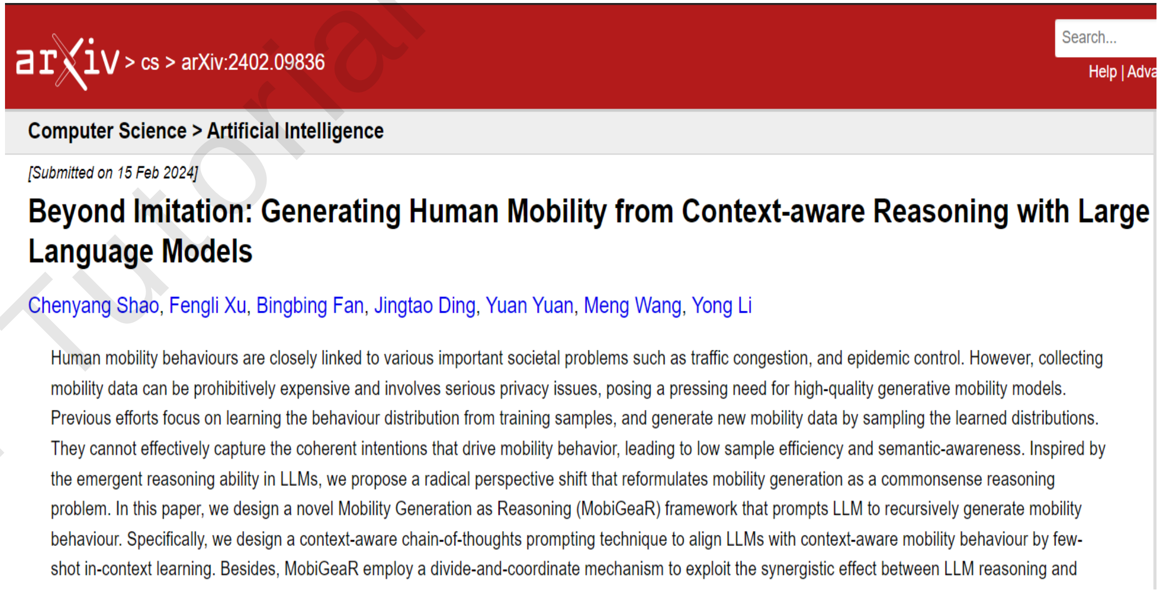
- ☐ **Token cost: 0.6% of pure LLM**
  - Gravity model saves lots of POI selection token
  - Multiple trajectories can be generated from the same intent template (20)
- ☐ Achieve large performance gains on several data quality metrics.

Experiments – Downstream Applications

- Enhance Mobility Prediction Tasks:



(a) Location Prediction Accuracy (b) Intent Prediction Accuracy



arXiv > cs > arXiv:2402.09836

Search...

Help | Adva

Computer Science > Artificial Intelligence

[Submitted on 15 Feb 2024]

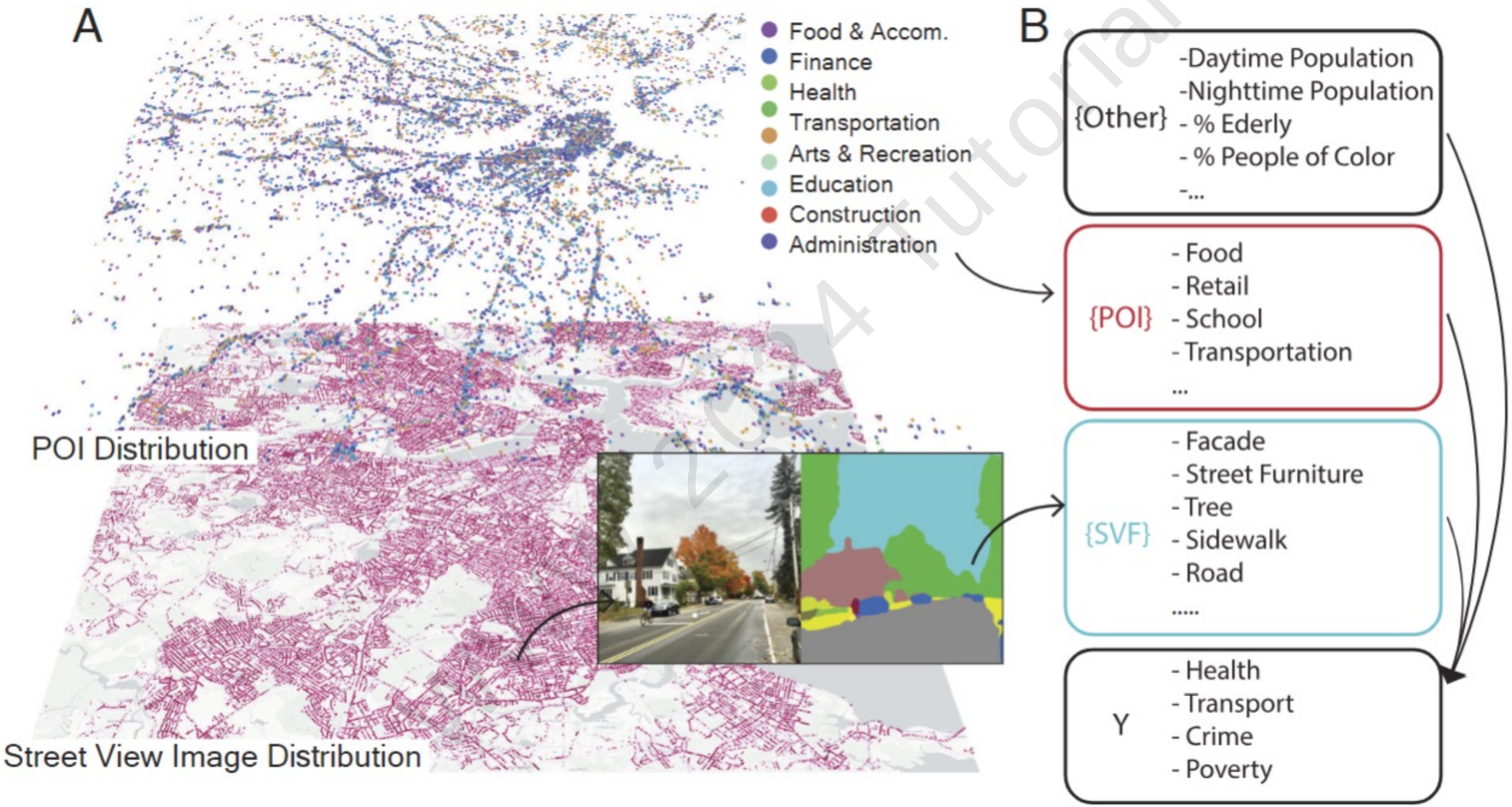
Beyond Imitation: Generating Human Mobility from Context-aware Reasoning with Large Language Models

Chenyang Shao, Fengli Xu, Bingbing Fan, Jingtao Ding, Yuan Yuan, Meng Wang, Yong Li

Human mobility behaviours are closely linked to various important societal problems such as traffic congestion, and epidemic control. However, collecting mobility data can be prohibitively expensive and involves serious privacy issues, posing a pressing need for high-quality generative mobility models. Previous efforts focus on learning the behaviour distribution from training samples, and generate new mobility data by sampling the learned distributions. They cannot effectively capture the coherent intentions that drive mobility behavior, leading to low sample efficiency and semantic-awareness. Inspired by the emergent reasoning ability in LLMs, we propose a radical perspective shift that reformulates mobility generation as a commonsense reasoning problem. In this paper, we design a novel Mobility Generation as Reasoning (MobiGeaR) framework that prompts LLM to recursively generate mobility behaviour. Specifically, we design a context-aware chain-of-thoughts prompting technique to align LLMs with context-aware mobility behaviour by few-shot in-context learning. Besides, MobiGeaR employ a divide-and-coordinate mechanism to exploit the synergistic effect between LLM reasoning and

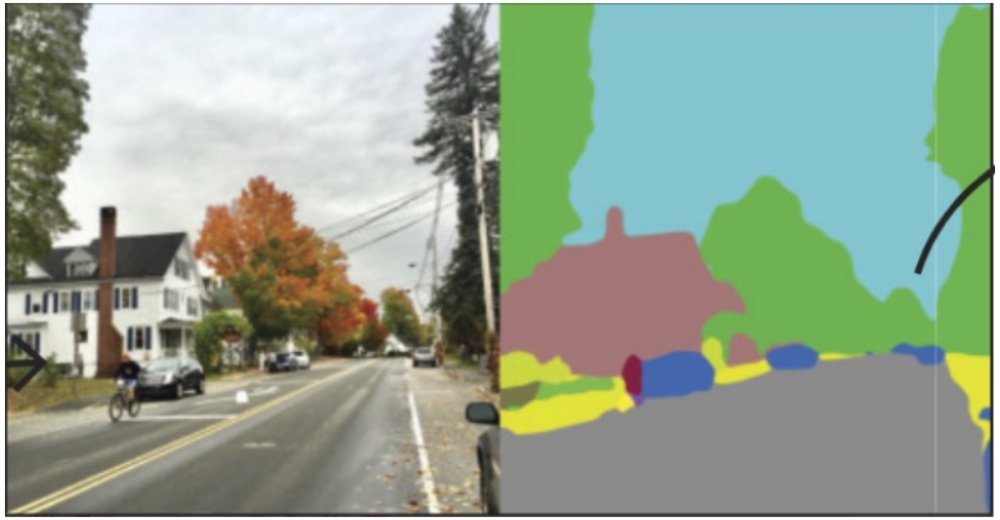
- Fix the num of real data: 100
- Num of generated data: 50/100/200/500/1000
- **33%** improvement over baselines

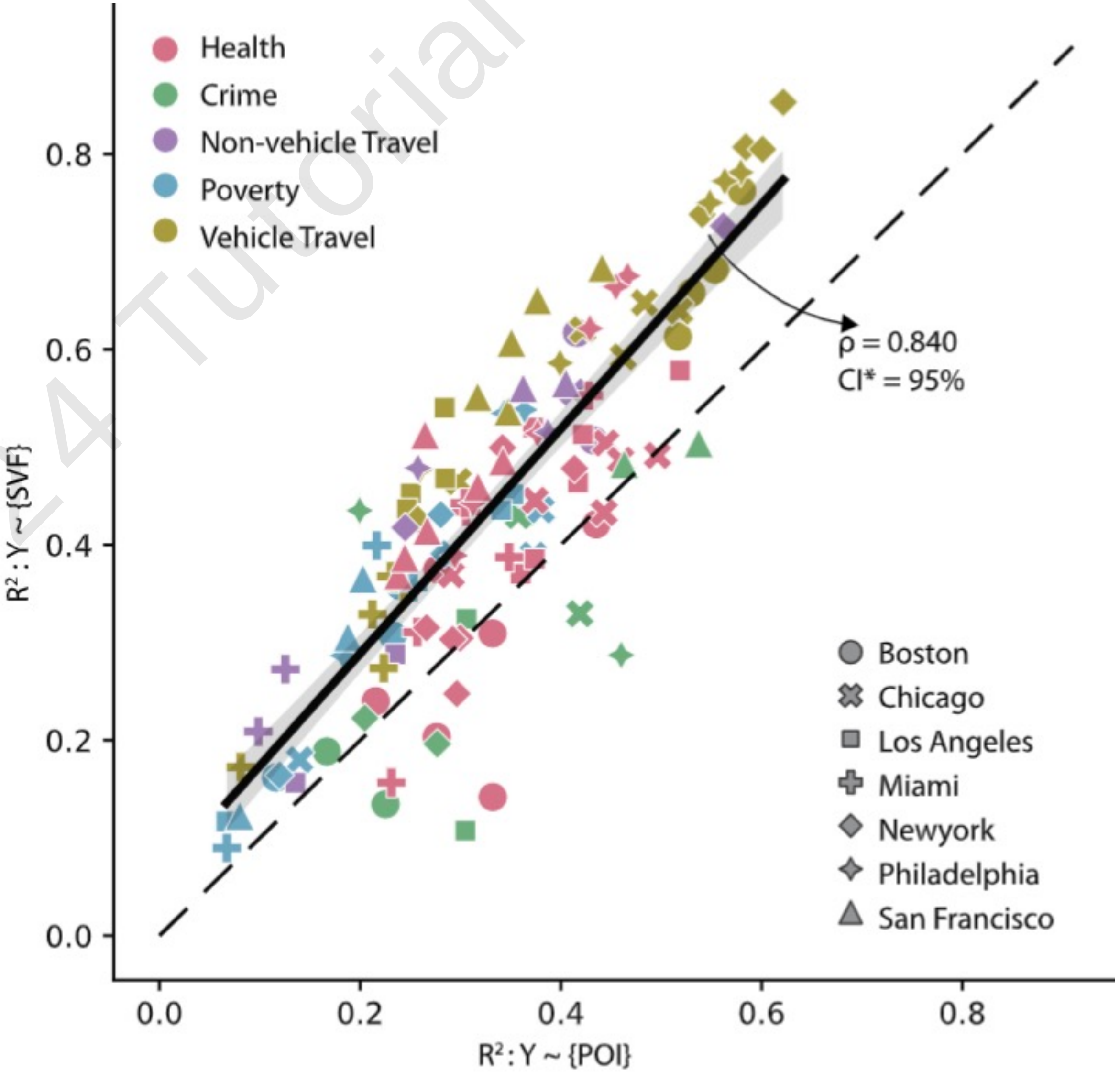
Multi-modal Urban Experiences: Street View



Fan, Zhuangyuan, et al. "Urban visual intelligence: Uncovering hidden city profiles with street view images." Proceedings of the National Academy of Sciences 120.27 (2023): e2220417120.

Multi-modal Urban Experiences: Street View

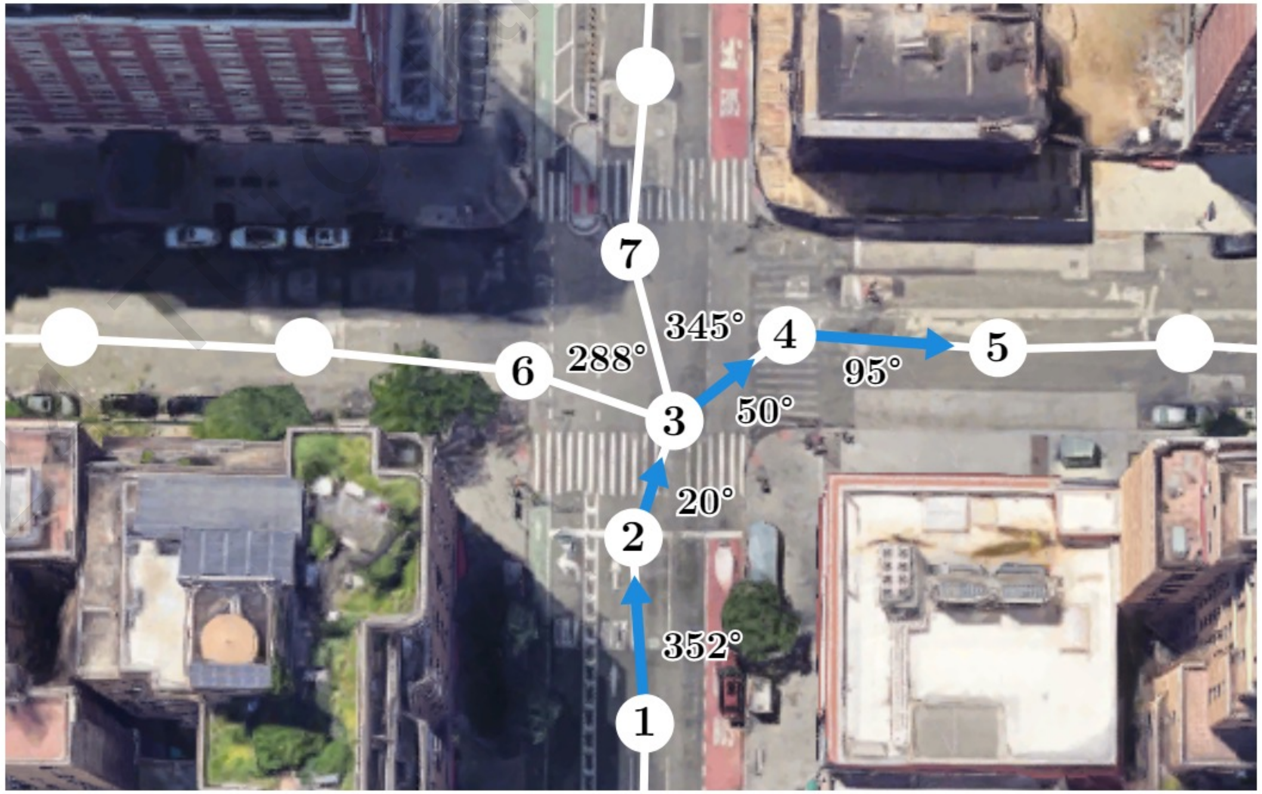




Fan, Zhuangyuan, et al. "Urban visual intelligence: Uncovering hidden city profiles with street view images." Proceedings of the National Academy of Sciences 120.27 (2023): e2220417120.

Can LLM Agents Pick Up Visual Cues?

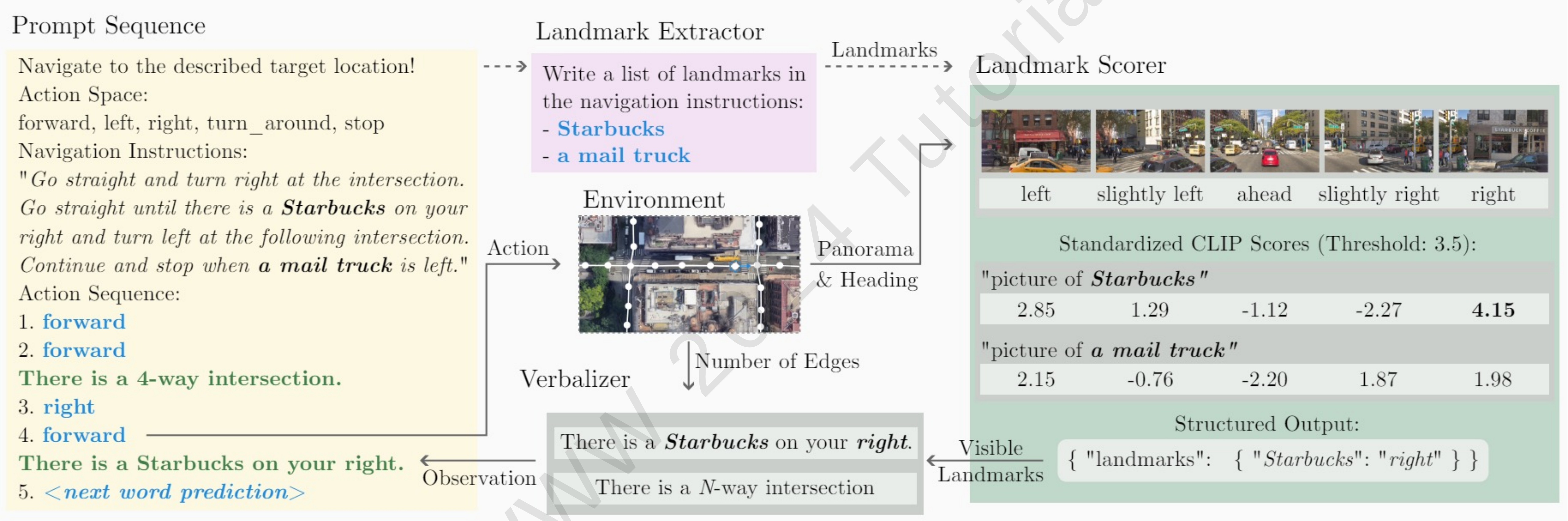
Navigate to the described target location!
Action Space: forward, left, right, turn_around, stop
Navigation Instructions:
"*Orientate yourself such that* ***a blue bench*** *is on your right, go to the end of the block and make a right. Follow the* ***park*** *on your left and make a right at the intersection. Pass the* ***black fire hydrant*** *on your right and stop when you get to a* ***gray door on the brown building***."
Action Sequence:
There is a blue bench on your left.
1. turn_around
There is a blue bench on your right.
2. forward
There is a 3-way intersection.
3. right
4. forward
There is a park on your left.
5. forward
There is a park on your left.
6. forward
There is a 4-way intersection.
7. *<next word prediction>*



Following Language Instructions to Navigate with Visual Cues in Street Views.

Schumann, Raphael, et al. "Velma: Verbalization embodiment of llm agents for vision and language navigation in street view." *AAAI* ,2024.

CLIP as a Multi-modal Perception Module



Using CLIP as a module to enable the multi-modal capabilities of LLM agents.

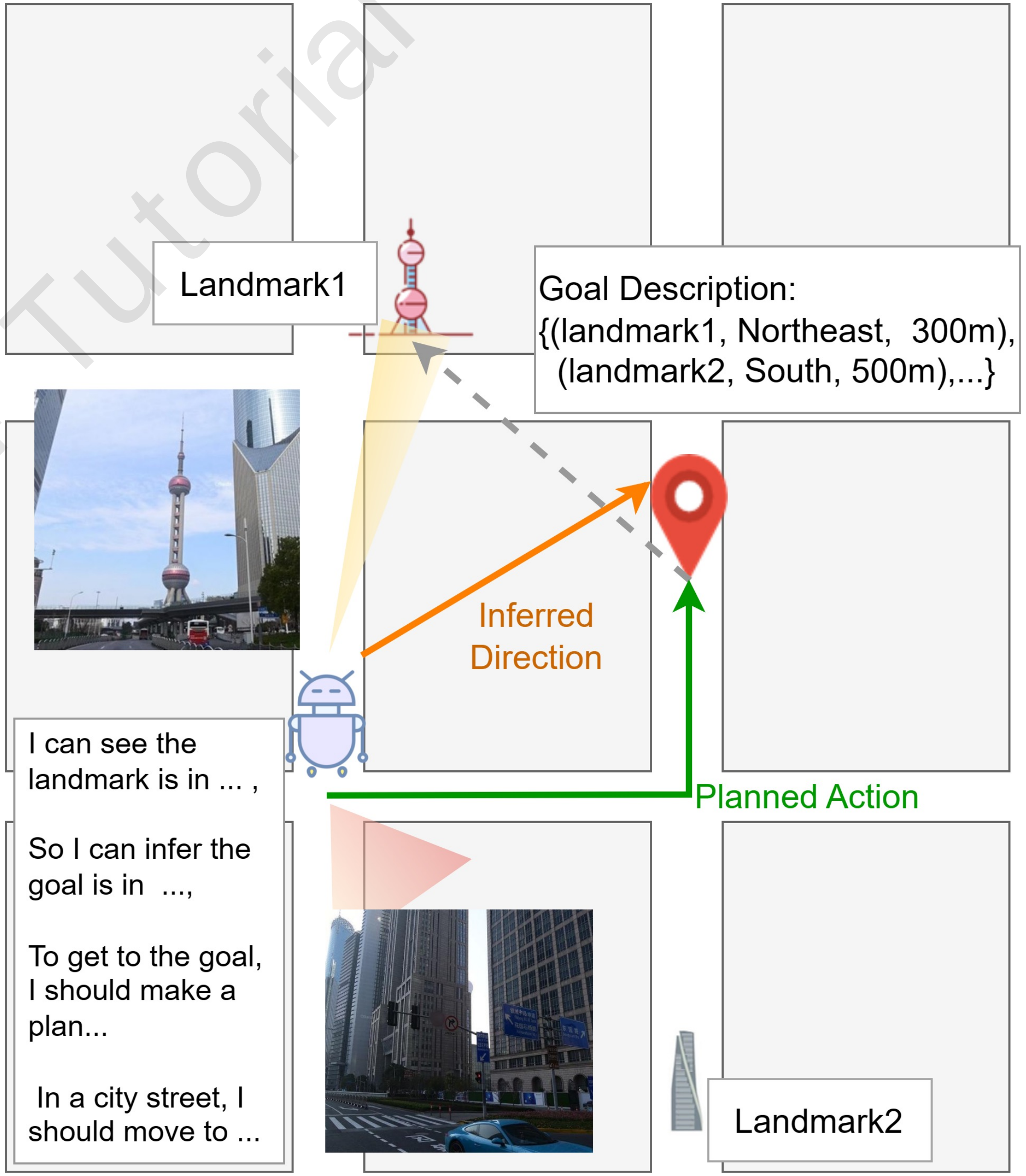
Schumann, Raphael, et al. "Velma: Verbalization embodiment of llm agents for vision and language navigation in street view." *AAAI* ,2024.

Goal-directed Navigation

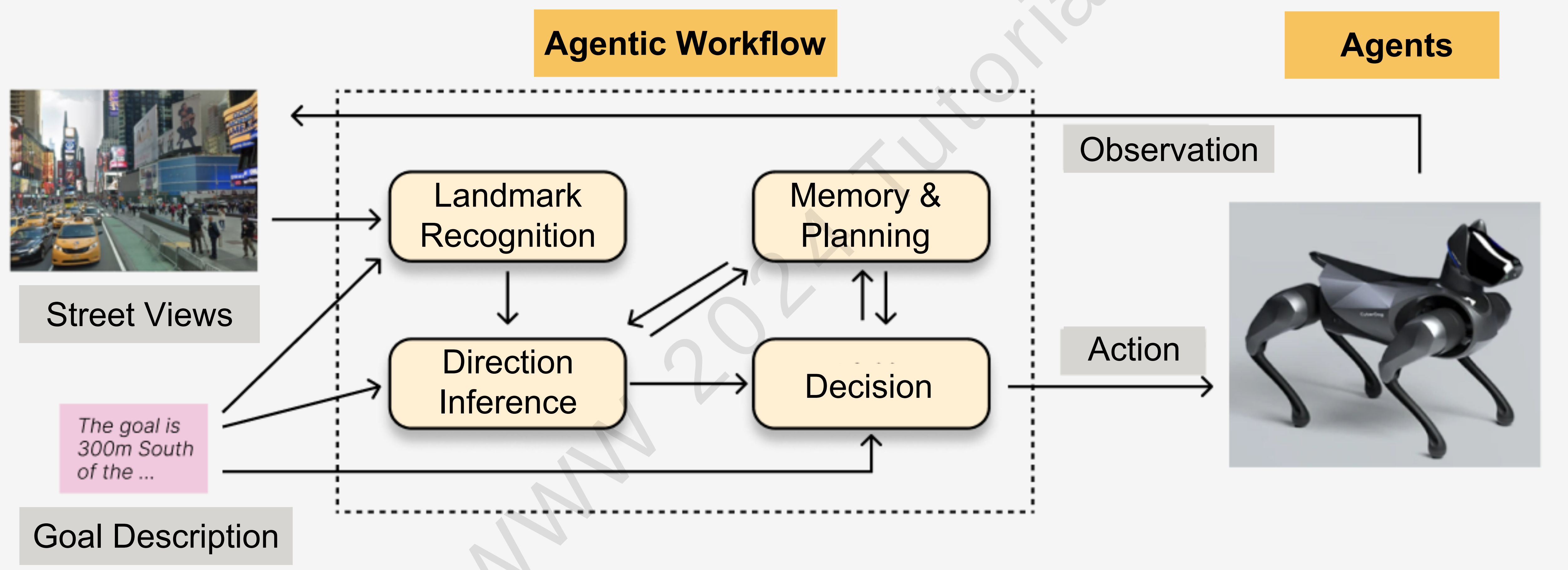
Can LLM Agents Navigate without Instructions?

Need to make its own decision:

- Spatial reasoning
- Memory
- Planning

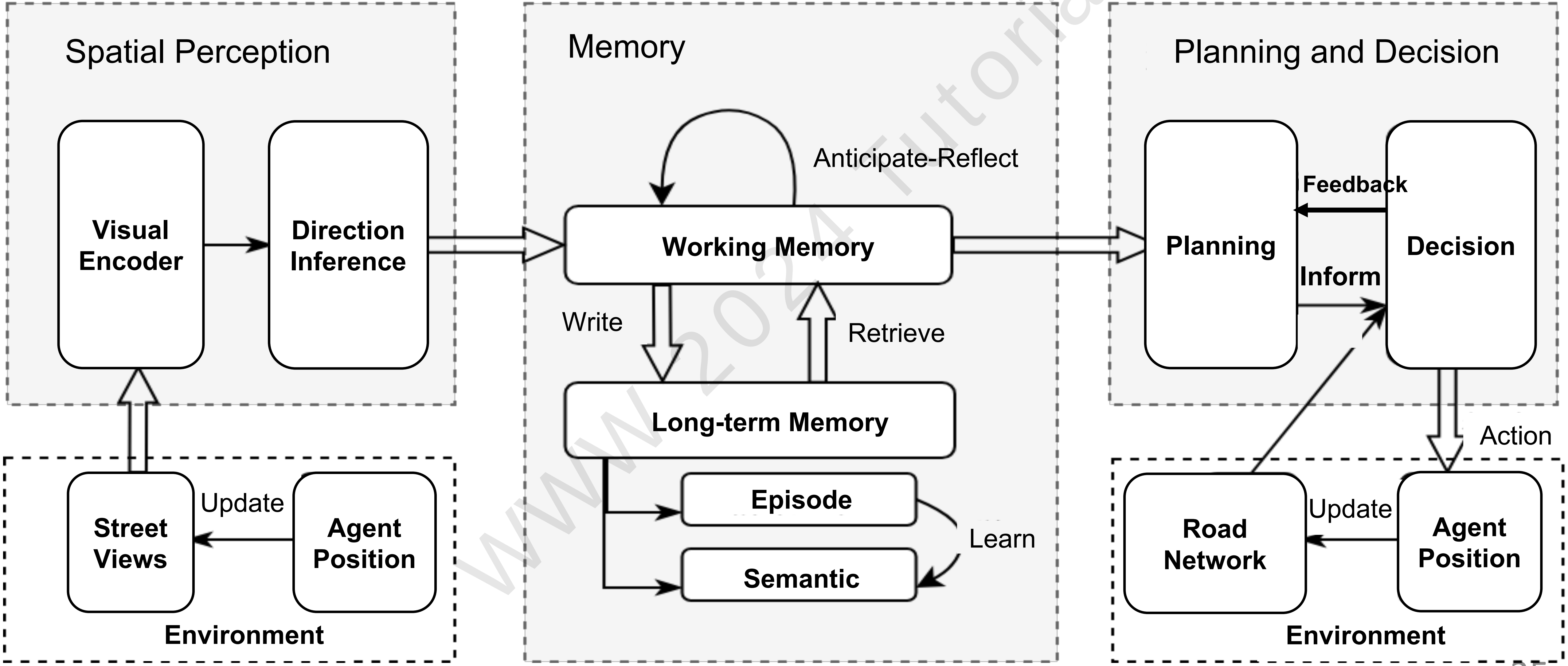


LLM Agents for Multi-modal Goal-directed Navigation

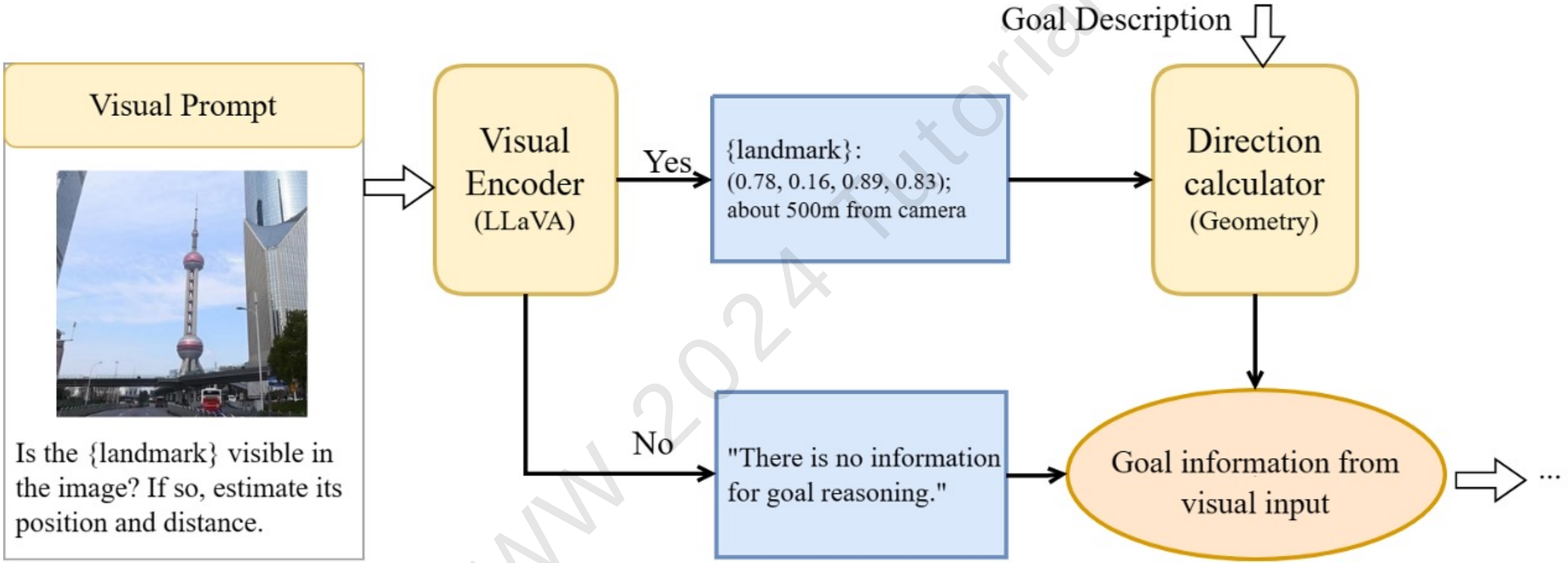


Multi-modal LLMs + Agentic Workflow

Workflows for Multi-modal Goal-directed Navigation



Finetuning Multi-modal LLMs for Landmark Recognition



Use image-conversation data to fine-tune a multimodal base model with the ability to recognize landmark:

Which landmark? Which direction? How far away?

Finetuning LLaVA

We finetune LLaVA-1.5-7B base model with 30k conversation data for 3 hours (LoRA mode on 1 x A100)

	Accuracy	Precision	Recall	F1-Score	IoU
Base	0.1873	0.0576	0.9347	0.1072	0.6432
Finetuned	**0.9980**	**0.9868**	**0.9695**	**0.9779**	**0.9152**

Q: Is the Oriental Pearl Tower visible in the image?

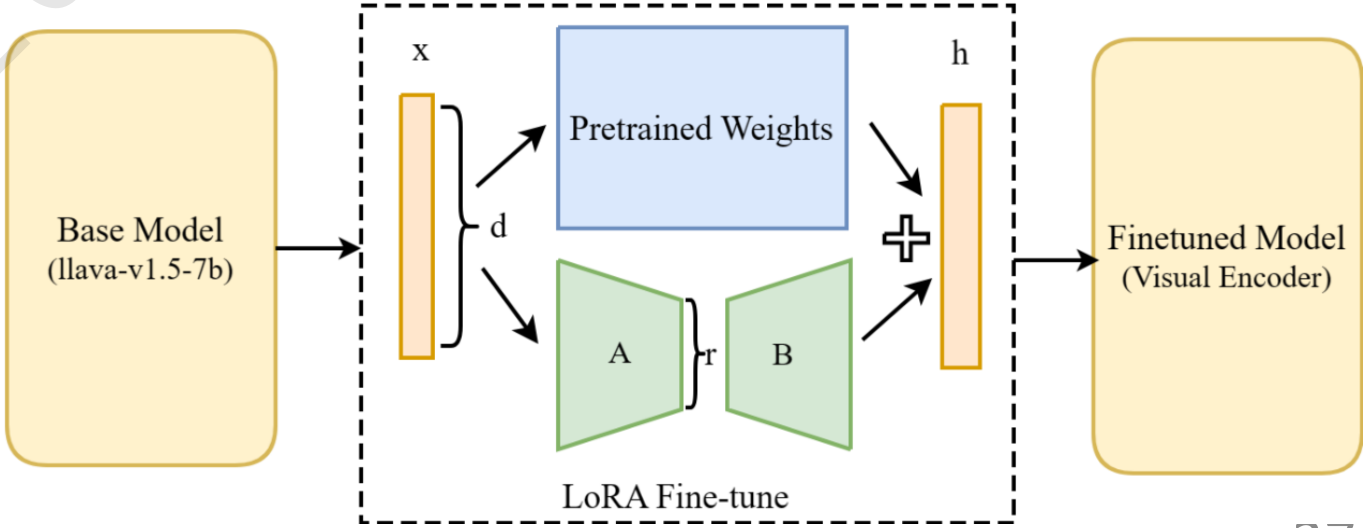
A: Yes. The Oriental Pearl Tower is visible in the image, and it's bounding box is (0.78, 0.16, 0.89, 0.83).

Q: How far is that place actually from the camera?

A: The Oriental Pearl Tower is about 650 meters away from the camera.

Finetune Setup

Env	Parameter
CPU	Intel(R) Xeon(R) Platinum 8358P
GPU	NVIDIA GeForce RTX A100 80G
OS	Ubuntu 22.04.1
Compiler	Python 3.10.13



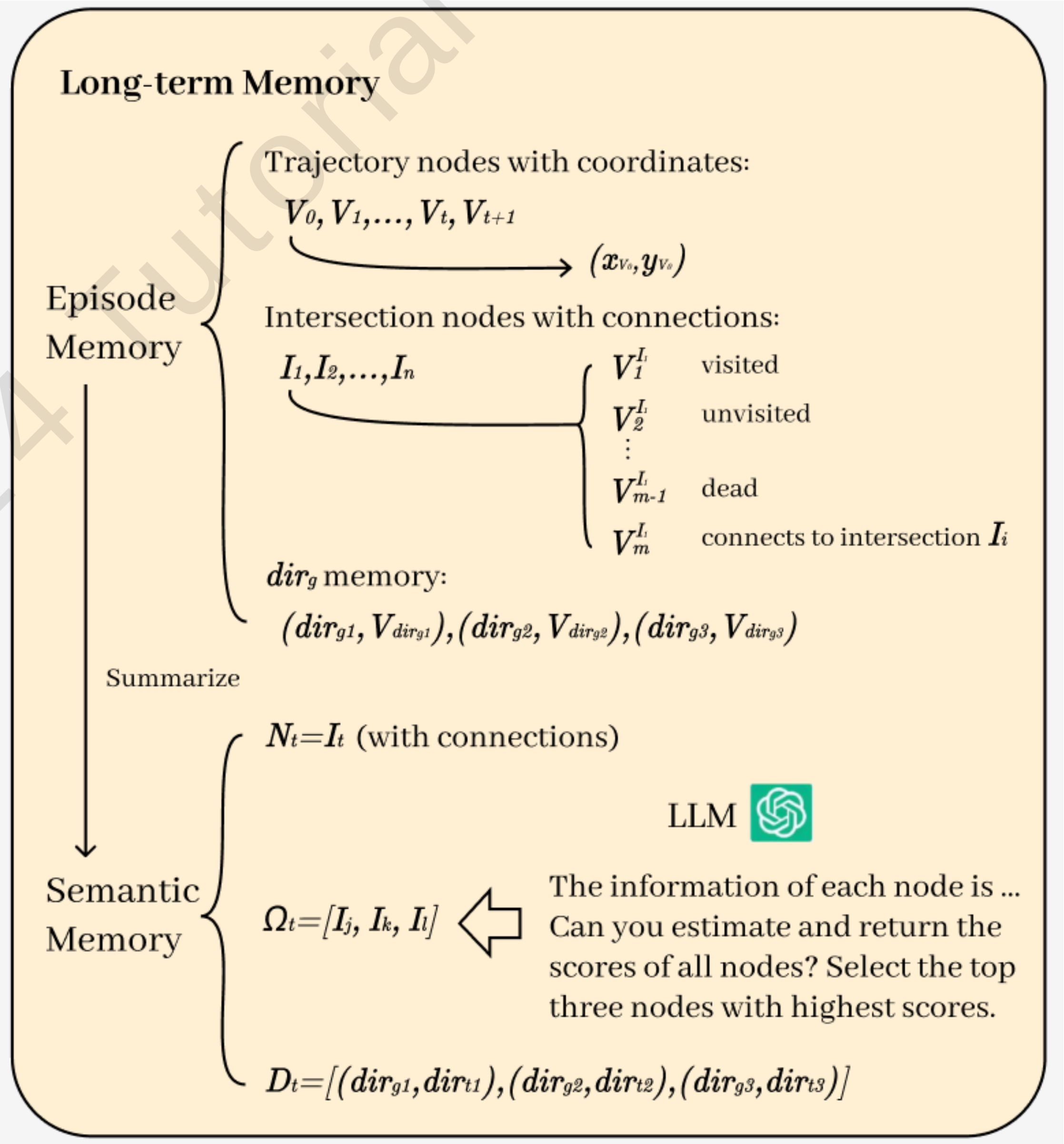
Memory Mechanism

Episodic Memory

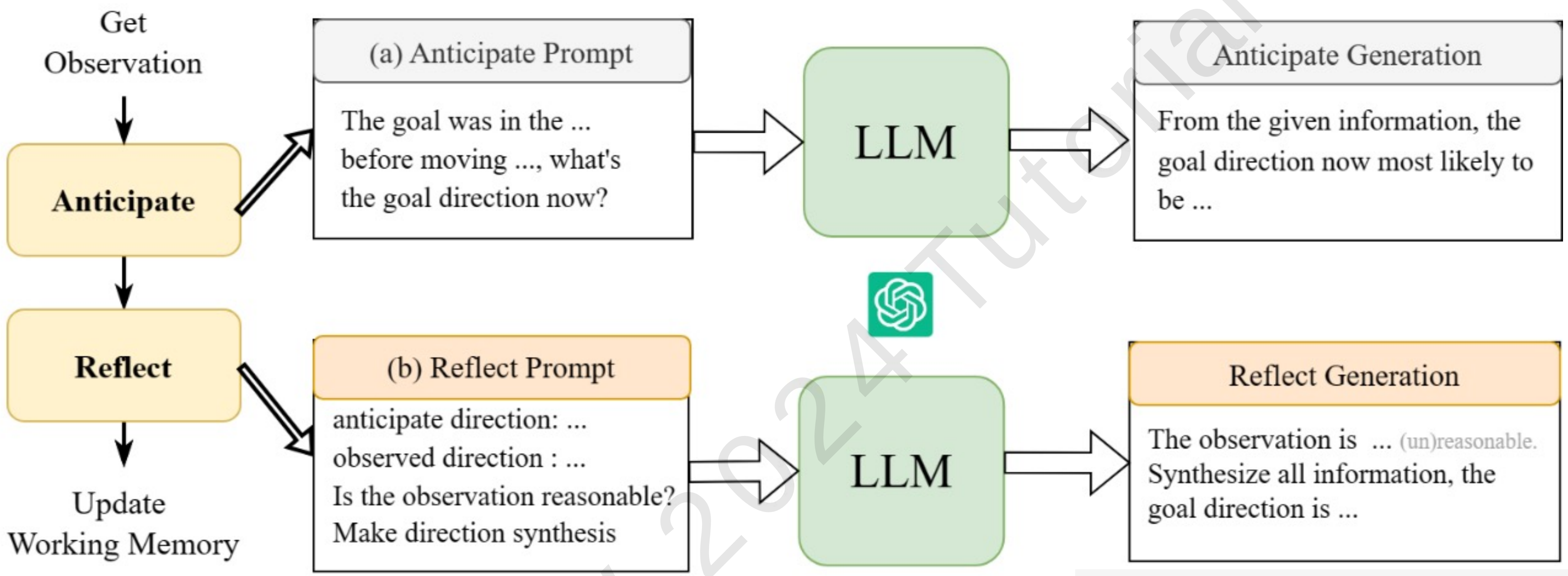
- Trajectory Nodes
  - ➢ Represented by coordinates
  - ➢ E.g. (0,0) – East → (1,0)
- Intersection Nodes
  - ➢ Visite Information
- Direction Memory
  - ➢ The three $\boldsymbol{dir\_i}$ memories and their corresponding nodes $\boldsymbol{V\_(dir\_i)}$

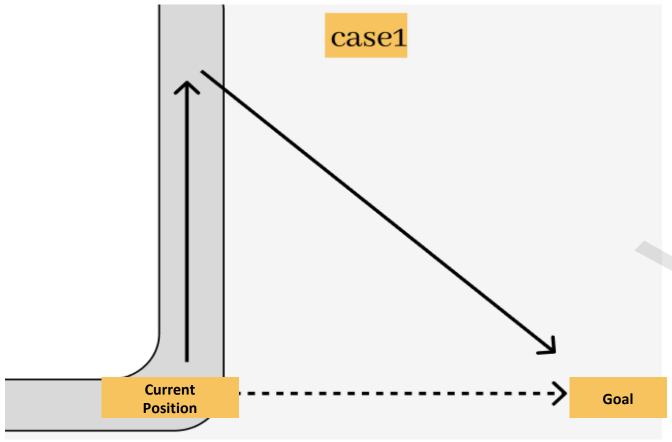
Semantic Memory

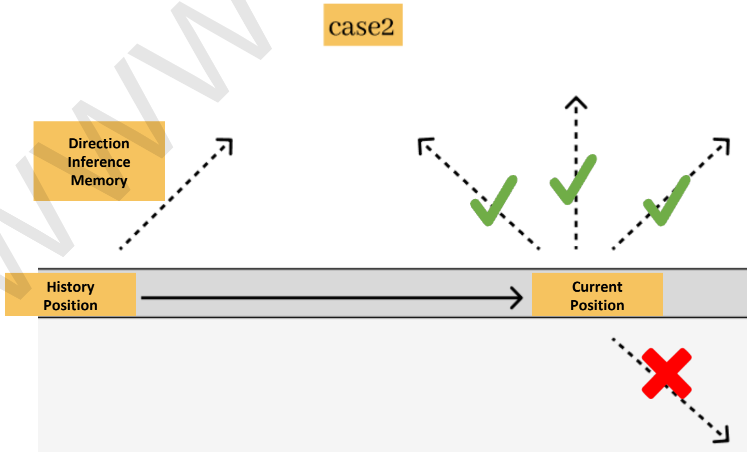
- A high-level summarizing description of movement history

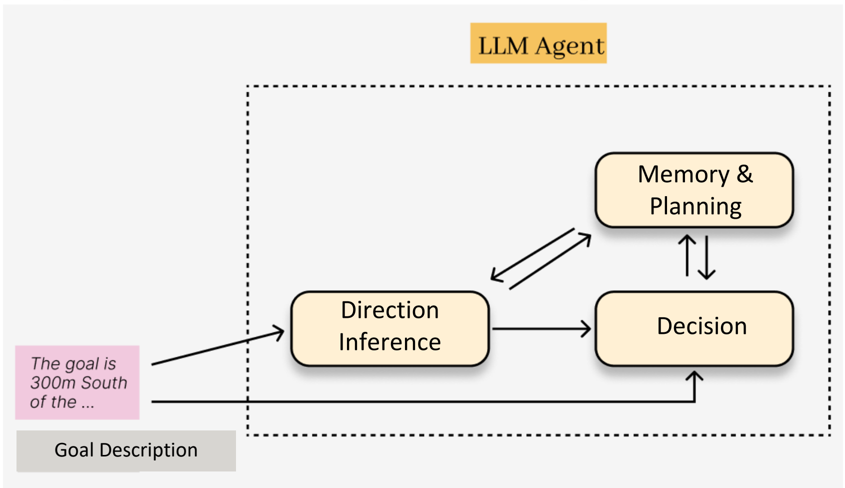


Anticipate & Reflect





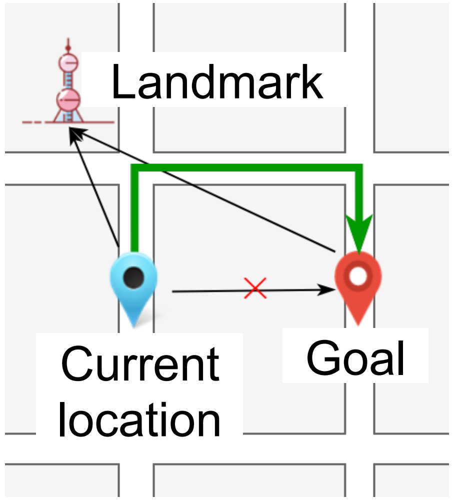


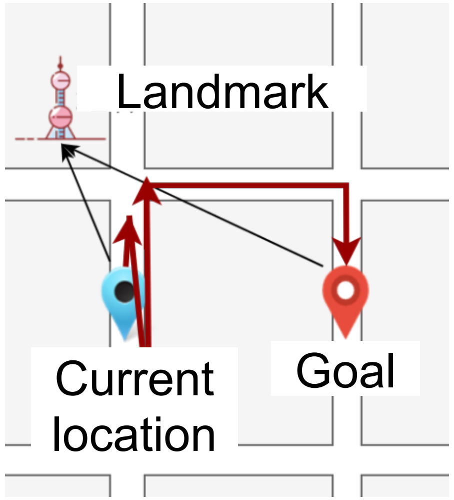


Experiments

Methods	Beijing			Shanghai		
	Success (%)	Steps	SPL	Success (%)	Steps	SPL
Random	0.0632	24.97	0.0168	0.0632	24.90	0.0220
RL method [*]	0.2105	20.80	0.1979	0.2000	21.06	0.1884
Heuristics	0.2421	20.37	0.1698	0.2632	19.81	0.1884
Ours	**0.4421**	**15.25**	**0.3384**	**0.4211**	**15.80**	**0.3154**
Ours (w/o Finetuned LLaVA)	0.1263	23.05	0.1038	0.1579	22.27	0.1298
Ours (w/o anticipate-reflection)	0.3263	18.07	0.2668	0.3158	18.26	0.2723
Ours (w/o Planning)	0.3895	16.50	0.3138	0.3789	16.78	0.3033

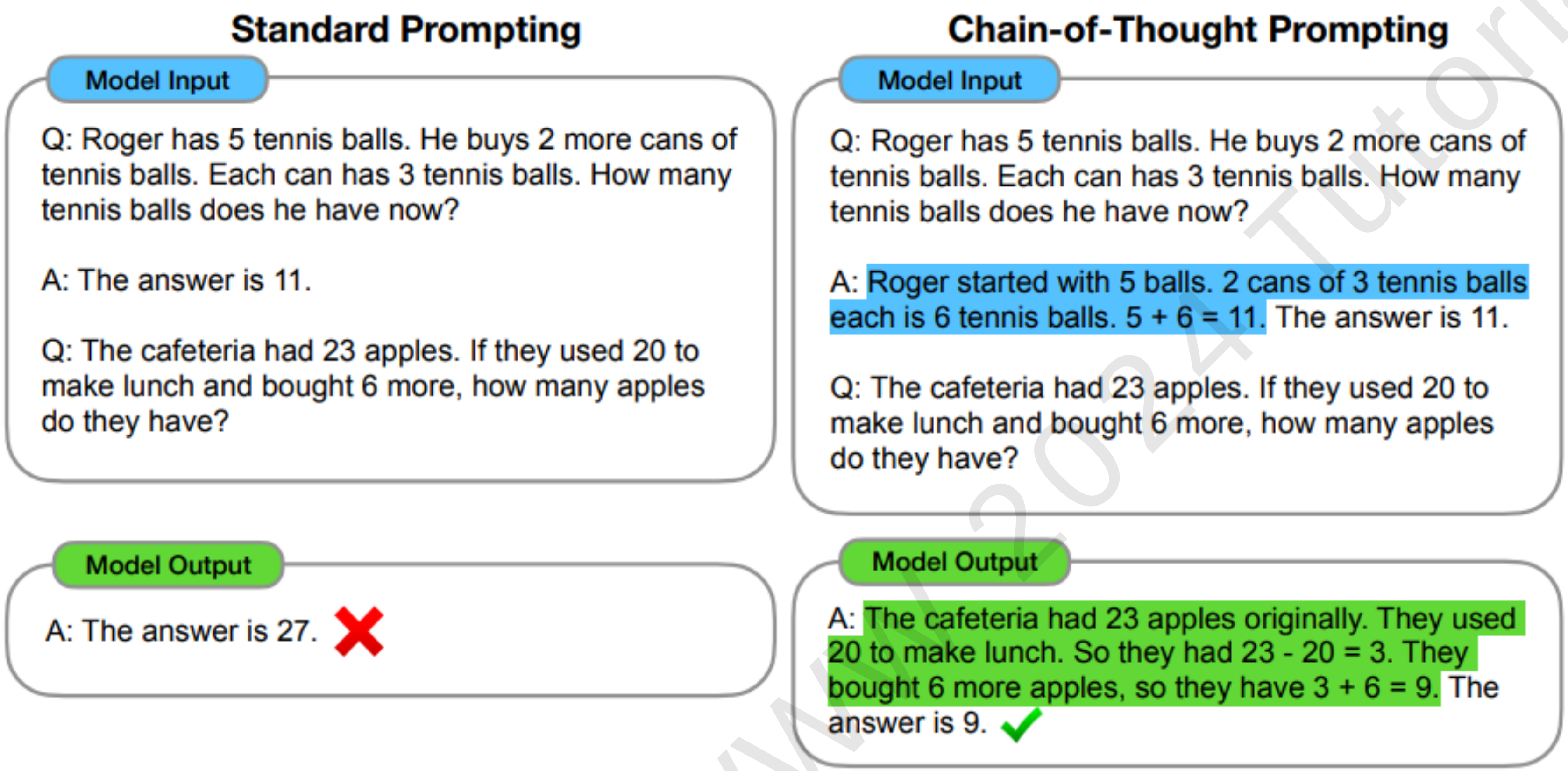
- Outperform RL methods[*] with thousands of training trajectories
- Ablation study show the effectiveness of each module
- Produce more consistent navigation behavior



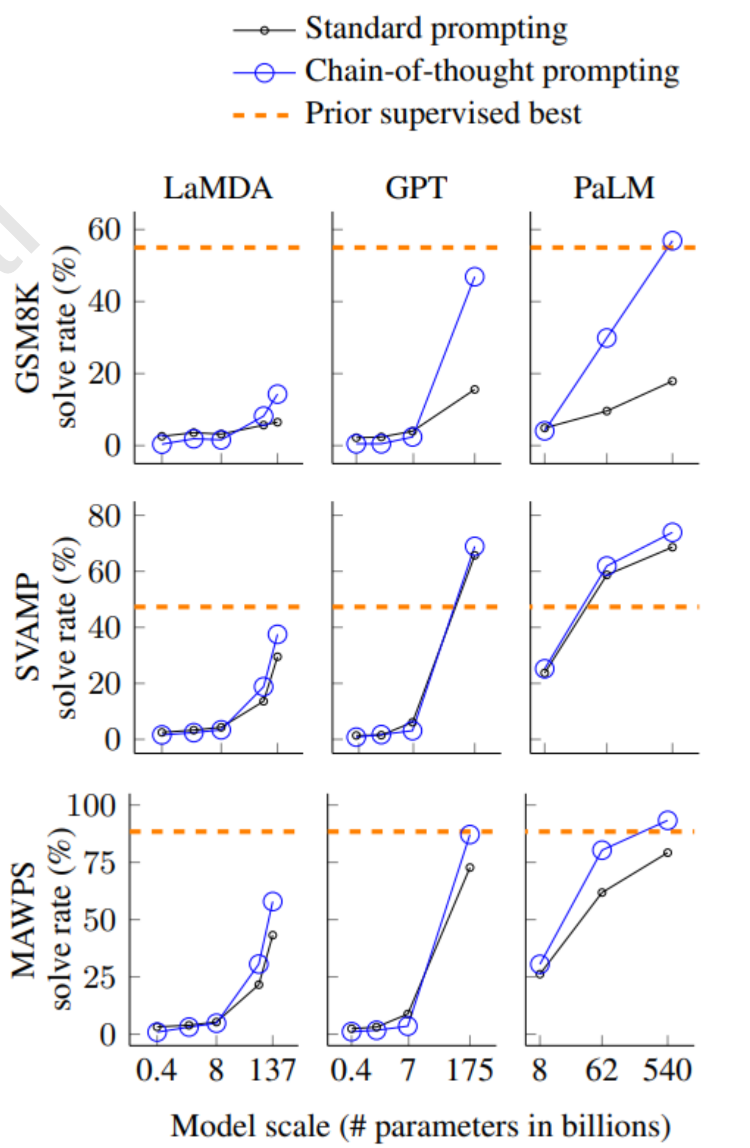


[*] Mirowski, Piotr, et al. "Learning to navigate in cities without a map." *Advances in neural information processing systems* 31 (2018).

Commonsense Reasoning in LLMs



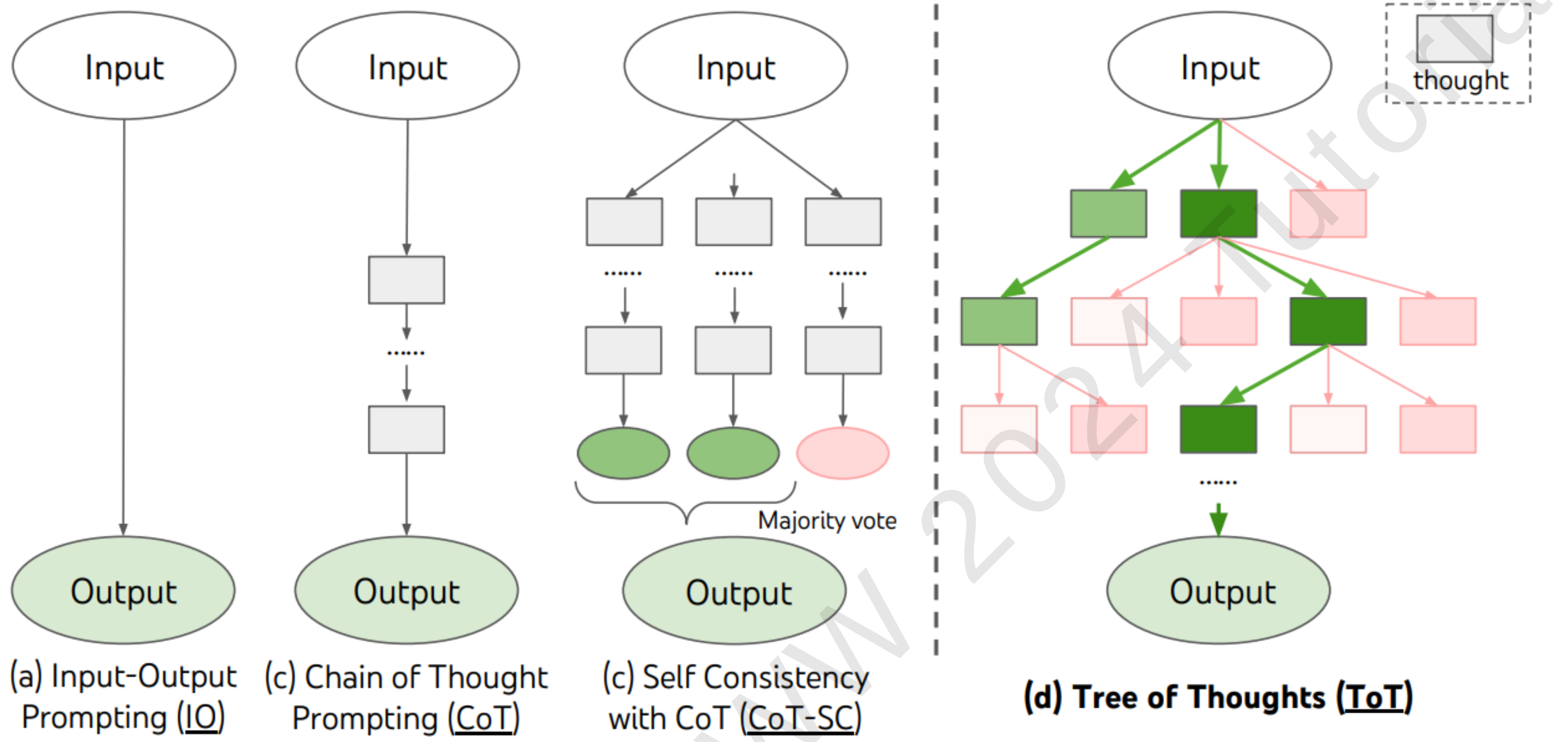
Chain-of-thoughts



Unleashing the scaling law in reasoning problems

Wei, Jason, et al. "Chain-of-thought prompting elicits reasoning in large language models." Advances in neural information processing systems 35 (2022): 24824-24837.

Token Cost



CoT -> ToT:

➢ Accuracy 4%->74%

➢ Token cost X100!

(Game of 24)

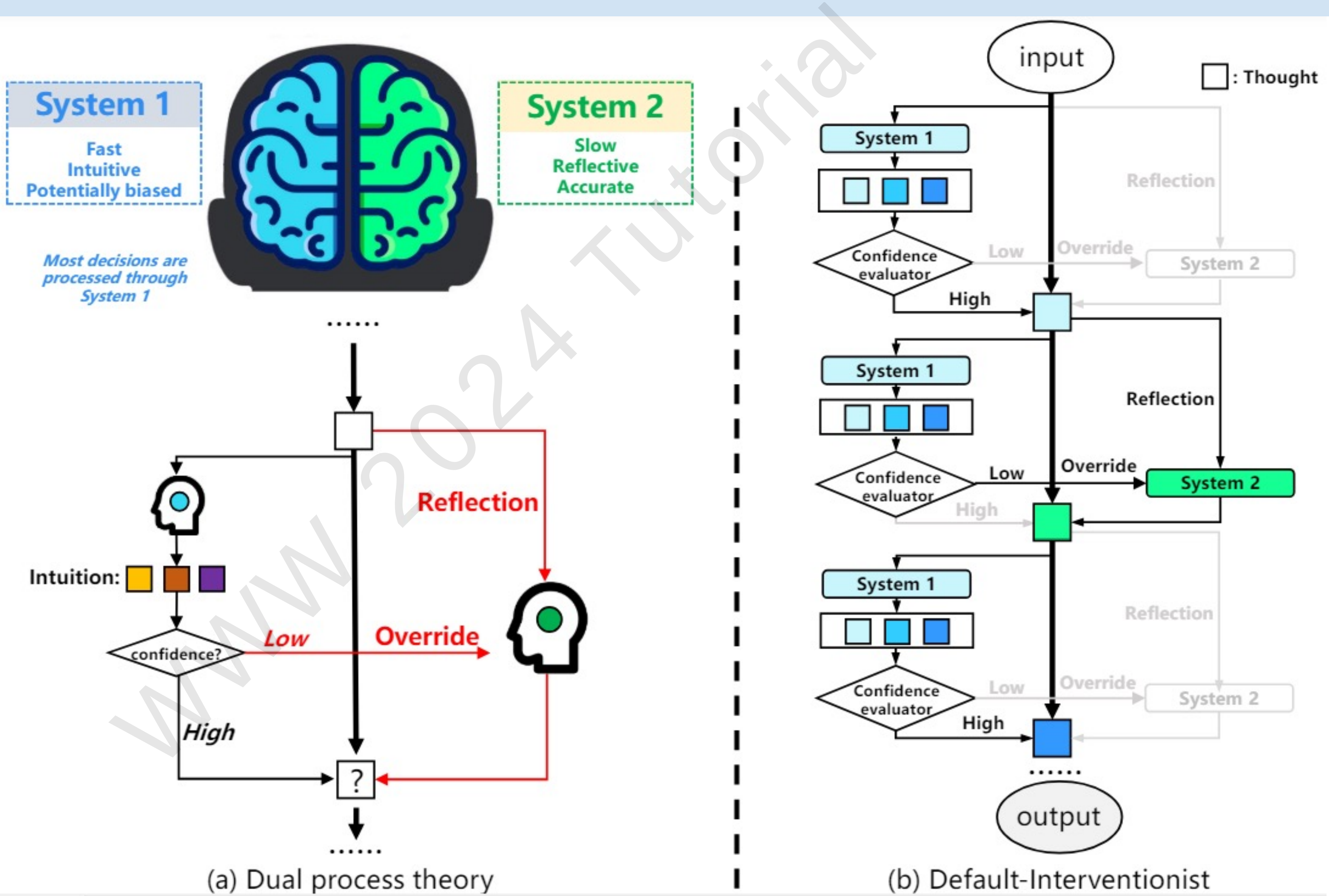
The current reasoning frameworks are getting increasingly accurate but also more costly.

Yao, Shunyu, et al. "Tree of thoughts: Deliberate problem solving with large language models." Advances in Neural Information Processing Systems 36 (2024).

Dual Process Theory

Dual Process in Human Cognition

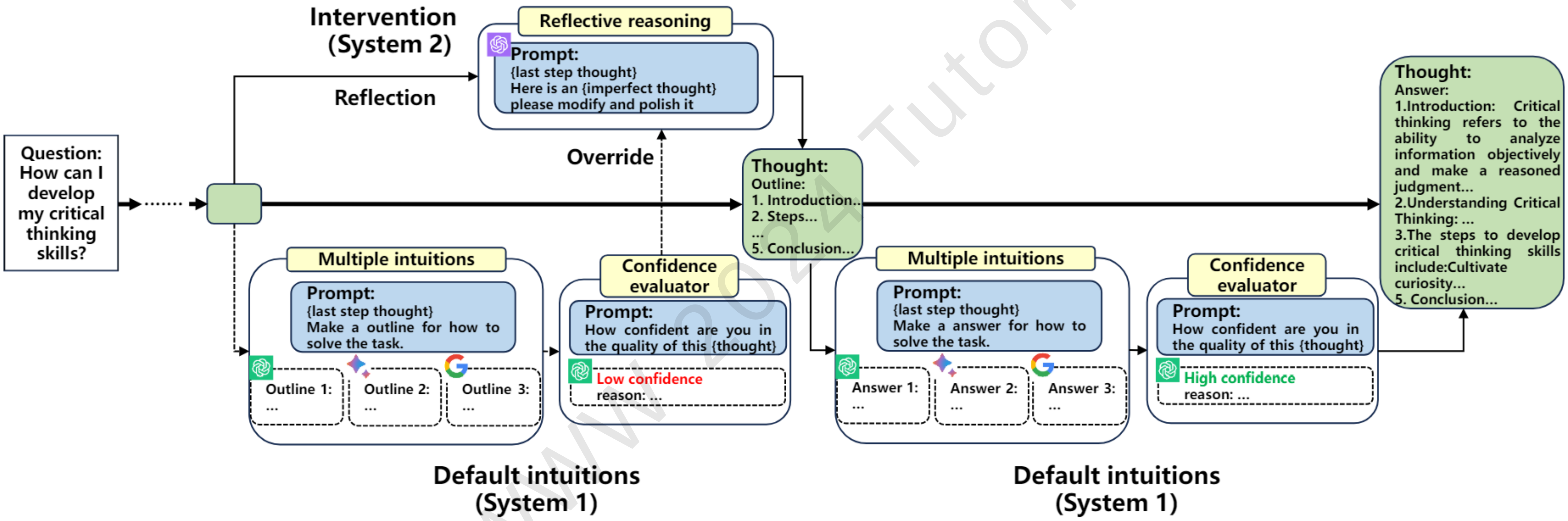
Synergy of Large & Smaller Models



(a) Dual process theory

(b) Default-Interventionist

Default-Interventionist Framework (DefInt)



Using Smaller LMs for default reasoning, when necessary trigger the accurate but effortful reflection of Larger LMs.

Default-Interventionist Framework (DefInt)

Intuition form Smaller LMs

Confidence Evaluator

Reflection of Larger LMs

Algorithm 2 The whole framework with DefInt

Input: Reasoning problem p, required reasoning steps N, period of System 2 T, evaluator LLM $f_E(\cdot)$

```
t = 0 //  Current reasoning step
S = {p} //  Set of current thoughts
while t ≤ N do
    t = t + 1
    if t%(T + 1)! = 0 then
        H = System 1(S) //  Regular System 1
        p, a = f_E(H) //  Evaluate and summarize
        if not p then
            z_t = a //  Accept intuitive thoughts
        end
        else
            S = S + a //  Wrap proposed intuitive thoughts
            z_t = System 2(S) //  Intervene
        end
    end
    else
        z_t = System 2(S) //  Regular System 2
    end
    S = {z_t} //  Update thought state
end
return S
```

Model	Input /1K tokens	Output /1K tokens
GPT-3.5	\$0.0015	\$0.002
GPT-4	\$0.03	\$0.06
PaLM2	0	0
Gemini	0	0

Denote the intervention rate as r, when $(1-r)(C_I + C_J) + r(C_I + C_J + C_R) < C_R^{full}$ is satisfied, DefInt is expected to effectively save token cost and the corresponding requirement of r is (taking $K_{in} = 1$ in most cases):

$$\boxed{r} < 1 - \frac{5C_{Ii} + 4C_{Io} + 3C_{Ji} + C_{Jo}}{C_{Ri} + C_{Ro}}. \quad (9)$$

Large margin of profitable intervention rates

Experiment

Math & Logic

Methods	Accuracy	Diversity	Token cost ($)	TFLOPS
CoT(best of 1)	4%	1.0	0.87	5.22E+04
CoT(best of 5)	14%	1.1	1.73	1.04E+05
CoT(best of 25)	32%	1.3	6.68	4.01E+05
CoT(best of 100)	58%	1.6	18.70	1.49E+06
Self-refine	20%	1.2	24.83	1.51E+06
ToT	64%	2.1	23.37	1.40E+06
SPP	12%	1.2	29.97	1.80E+06
MAD+judge	22%	1.3	28.09	1.69E+06
DefInt	**78%**	**2.5**	11.93	1.00E+06

Game of 24

Methods	Accuracy	Token cost ($)	TFLOPS
CoT(best of 1)	65.8%	6.72	4.03E+05
CoT(best of 5)	67.1%	27.26	1.64E+06
Self-refine	60.6%	33.37	2.00E+06
ToT	66.1%	38.66	2.32E+06
SPP	68.3%	20.68	1.24E+06
MAD+judge	66.8%	45.00	2.70E+06
DefInt	**72.0%**	9.63	6.41E+05

Logic Grid Puzzle

Creative

Methods	Accuracy	Diversity	Token cost ($)	TFLOPS
CoT(best of 1)	67.1%	3.8	3.37	2.02E+05
Self-refine	78.2%	4.9	17.79	1.07E+06
ToT	76.8%	4.4	27.32	1.64E+06
SPP	79.9%	5.8	10.94	6.56E+05
MAD+judge	77.4%	6.1	17.00	1.02E+06
DefInt	**83.4%**	**6.3**	2.75	1.84E+05

Creative Writing

Open-ended

Methods	FairEval	Diversity	Token cost ($)	TFLOPS
DefInt		**5.2**	4.77	3.03E+05
v.s. CoT(best of 1)	66.5%	3.1	2.27	1.36E+05
v.s. CoT(best of 5)	62.7%	3.3	8.72	5.23E+05
v.s. Self-refine	56.2%	4.2	15.27	9.16E+05
v.s. ToT	61.3%	3.3	19.44	1.17E+06
v.s. SPP	67.6%	3.8	8.33	5.00E+05
v.s. MAD+judge	55.0%	4.9	17.00	1.02E+06

Open-ended Question Answering

Methods	FairEval	Diversity	Token cost ($)	TFLOPS
DefInt		**6.1**	6.66	4.16E+05
v.s. CoT(best of 1)	71.9%	4.2	4.84	2.91E+05
v.s. CoT(best of 5)	68.1%	4.2	20.11	1.21E+06
v.s. Self-refine	61.2%	5.3	31.86	1.91E+06
v.s. ToT	67.3%	4.7	58.52	3.51E+06
v.s. SPP	74.5%	4.6	26.69	1.60E+06
v.s. MAD+judge	63.6%	5.6	36.59	2.20E+06

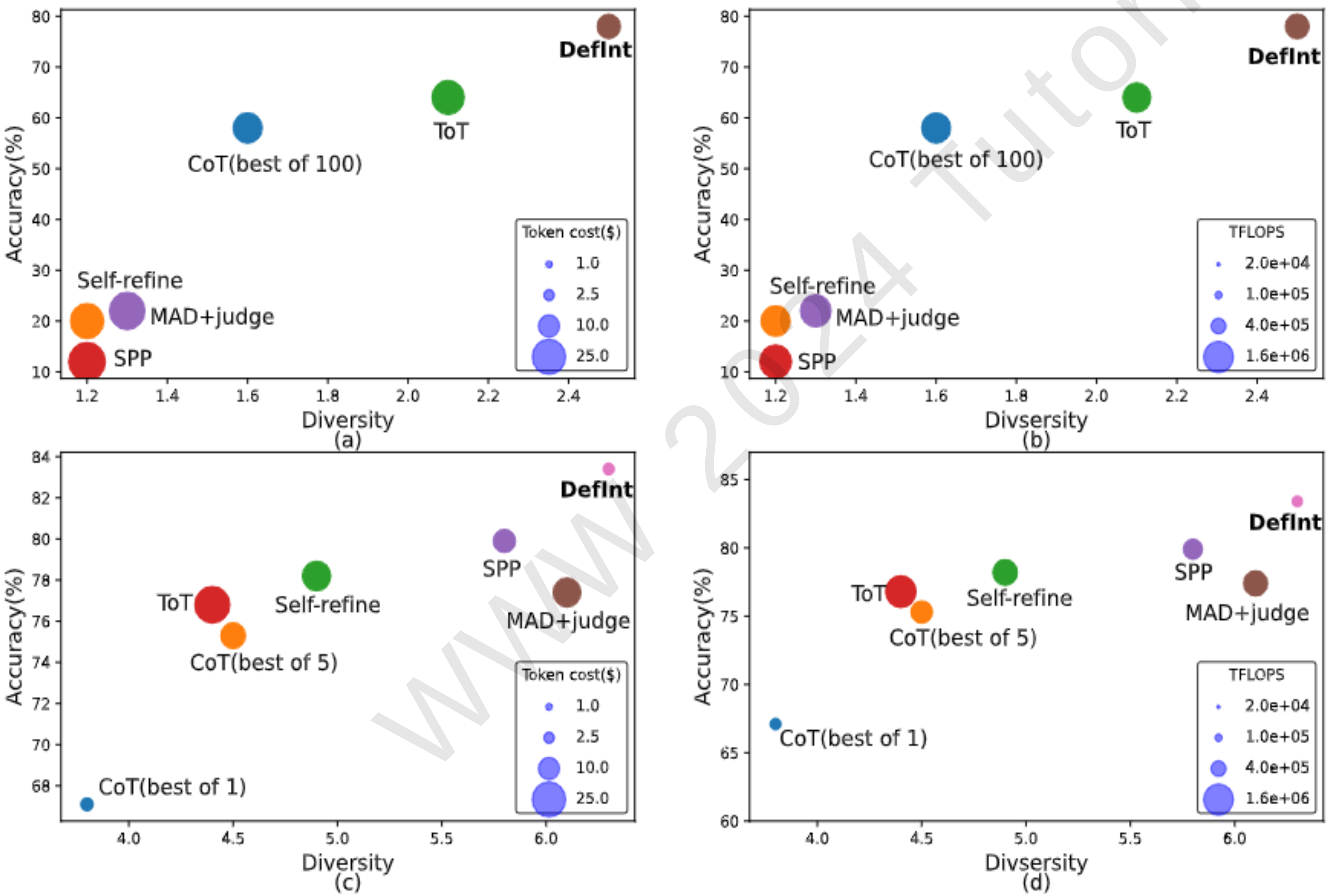
Constrained Generation

Experiment

Dollar Cost　　　　Estimated TFLOPS

Game of 24

Creative Writing



➢ Reduce token cost by **75%**

➢ Achieving **SOTA** performance

Empirical Intervention Rate

Synergy options	Game of 24	Logic Grid Puzzle	Creative Writing	OpenQA	Constrained Generation
3 GPT-3.5 + GPT-4	18%	12%	24%	15%	13%
3 PaLM2 + GPT-4	26%	23%	29%	19%	15%
3 Gemini + GPT-4	22%	13%	30%	14%	18%
GPT-3.5/PaLM2/Gemini + GPT-4	16%	8%	21%	11%	15%

The empirical intervention rate
is between 10%~30%

Even in complex reasoning tasks, large amount of intermediary steps can be offloaded to smaller LMs

More Configurations

GSM8K：

Methods	Accuracy	Token cost ($)	TFLOPS
DefInt (Sys1: GPT-3.5+Palm2+Gemini, Sys2: GPT-4)	**94.4%**	13.79	8.76E+05
DefInt (Sys1: Mistral-7B+LLaMA-13B+Yi-34B, Sys2: GPT-4)	93.3%	15.05	9.03E+05

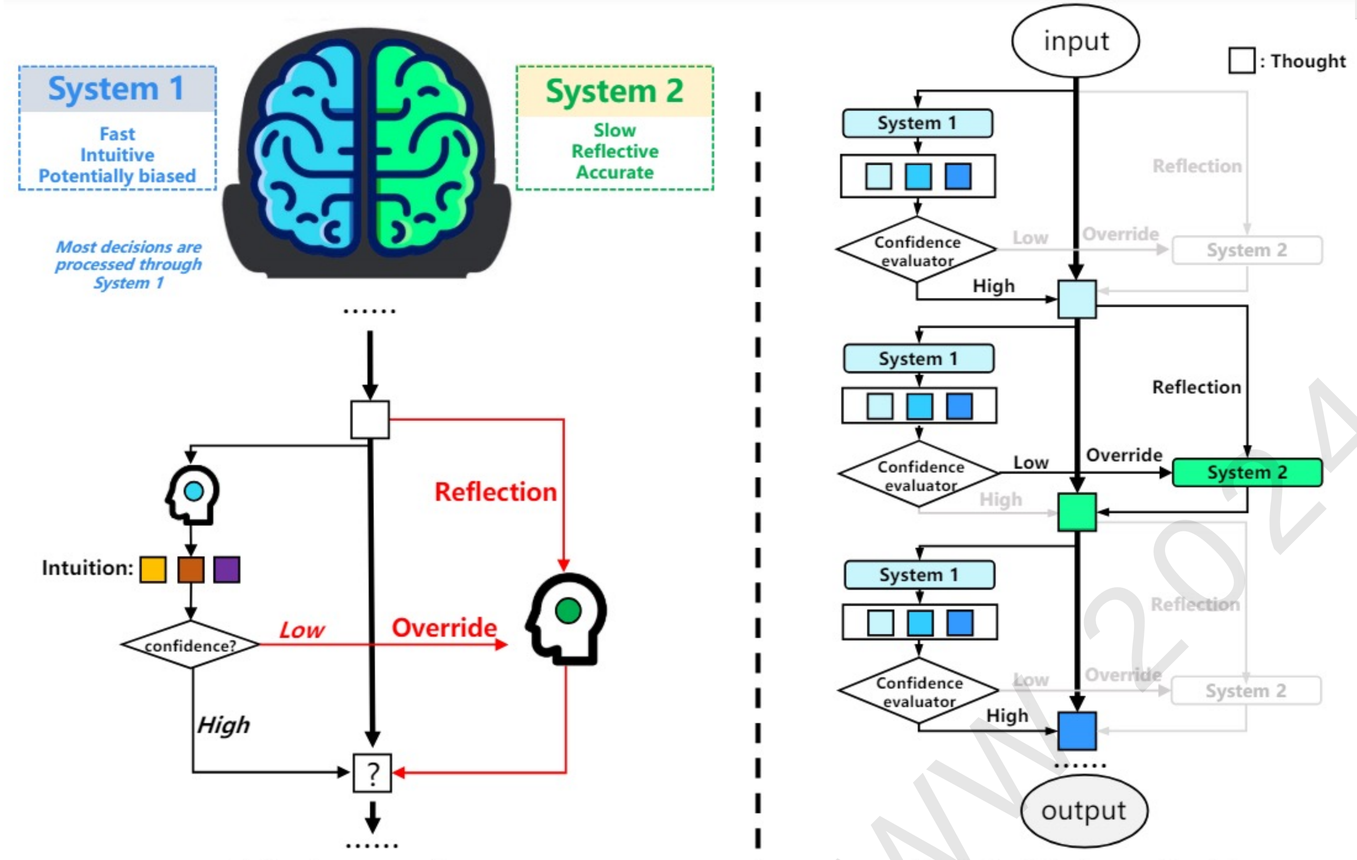
Game of 24：

Methods	Accuracy	Solution diversity	Token cost ($)	TFLOPS
DefInt (Sys1: GPT-3.5+Palm2+Gemini, Sys2: GPT-4)	**78%**	**2.5**	11.93	1.00E+06
DefInt (Sys1: Mistral-7B+LLaMA-13B+Yi-34B, Sys2: GPT-4)	74%	2.3	9.06	5.44E+05

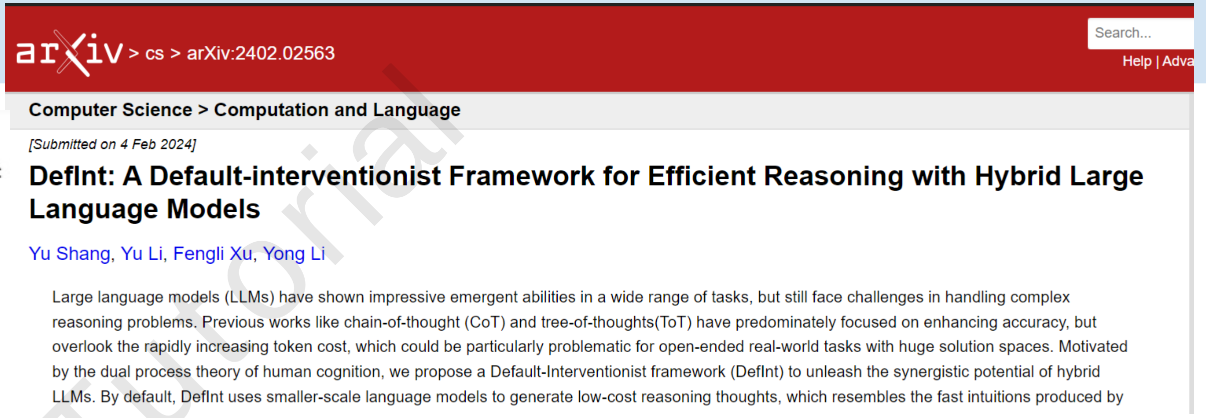
Creative Writing：

Methods	Accuracy	Solution diversity	Token cost ($)	TFLOPS
DefInt (Sys1: GPT-3.5+Palm2+Gemini, Sys2: GPT-4)	**83.4%**	**6.3**	2.75	1.84E+05
DefInt (Sys1: Mistral-7B+LLaMA-13B+Yi-34B, Sys2: GPT-4)	82.6%	6.1	2.70	1.62E+05

Summary



(a) Dual process theory
(b) Default-Interventionist


arXiv > cs > arXiv:2402.02563

Search...
Help | Adva

Computer Science > Computation and Language

[Submitted on 4 Feb 2024]

DefInt: A Default-interventionist Framework for Efficient Reasoning with Hybrid Large Language Models

Yu Shang, Yu Li, Fengli Xu, Yong Li

Large language models (LLMs) have shown impressive emergent abilities in a wide range of tasks, but still face challenges in handling complex reasoning problems. Previous works like chain-of-thought (CoT) and tree-of-thoughts(ToT) have predominately focused on enhancing accuracy, but overlook the rapidly increasing token cost, which could be particularly problematic for open-ended real-world tasks with huge solution spaces. Motivated by the dual process theory of human cognition, we propose a Default-Interventionist framework (DefInt) to unleash the synergistic potential of hybrid LLMs. By default, DefInt uses smaller-scale language models to generate low-cost reasoning thoughts, which resembles the fast intuitions produced by

DefInt: A Default-interventionist Framework for Efficient Reasoning with Hybrid Large Language Models

Large room for exploiting the synergy between small & large LMs

Building a Platform for City Simulation

Simple, *Finite* Simulation

V.S.

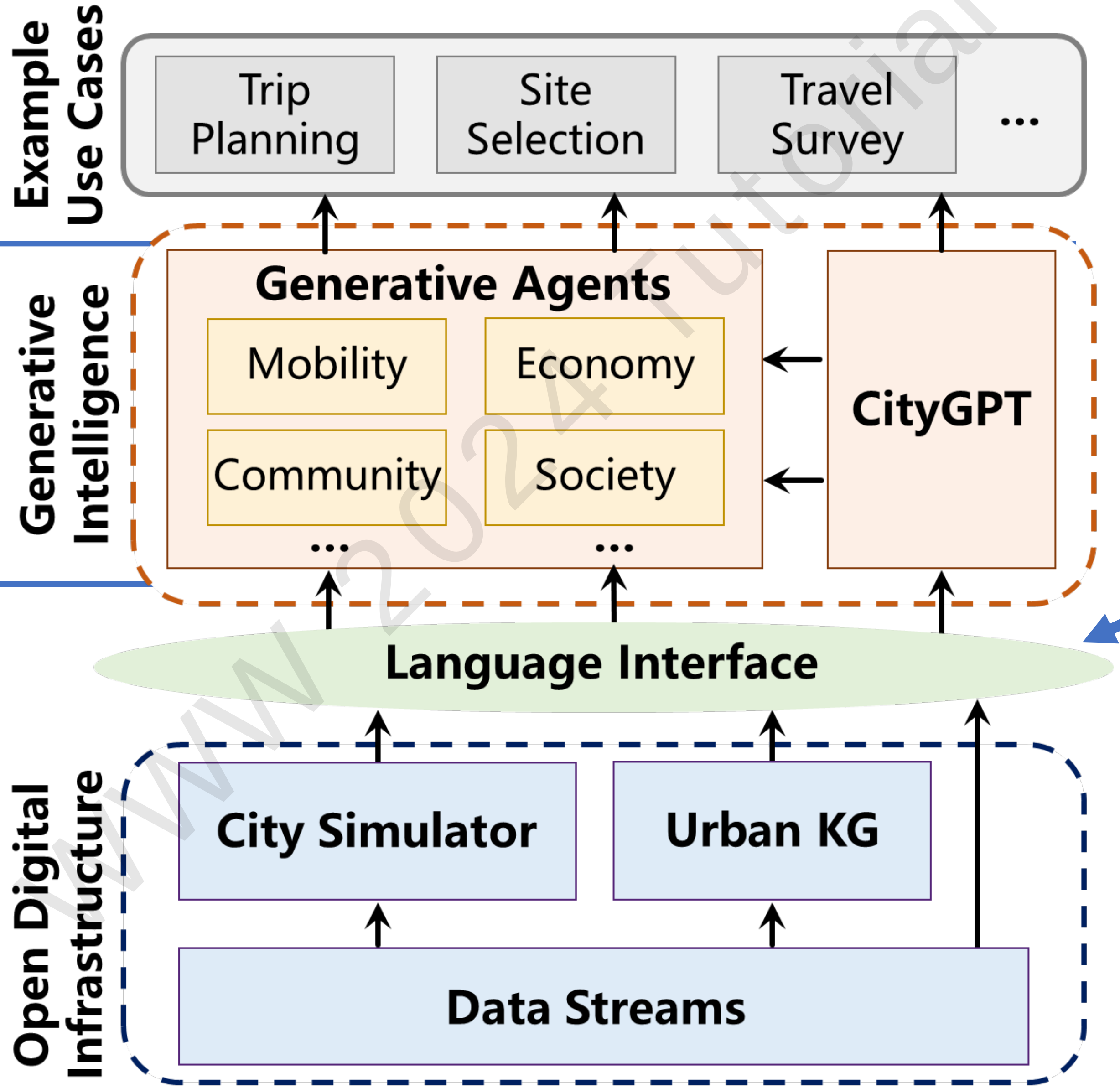
Complex, *Open-ended* Urban System

Building a Platform for City Simulation

Human Friendly

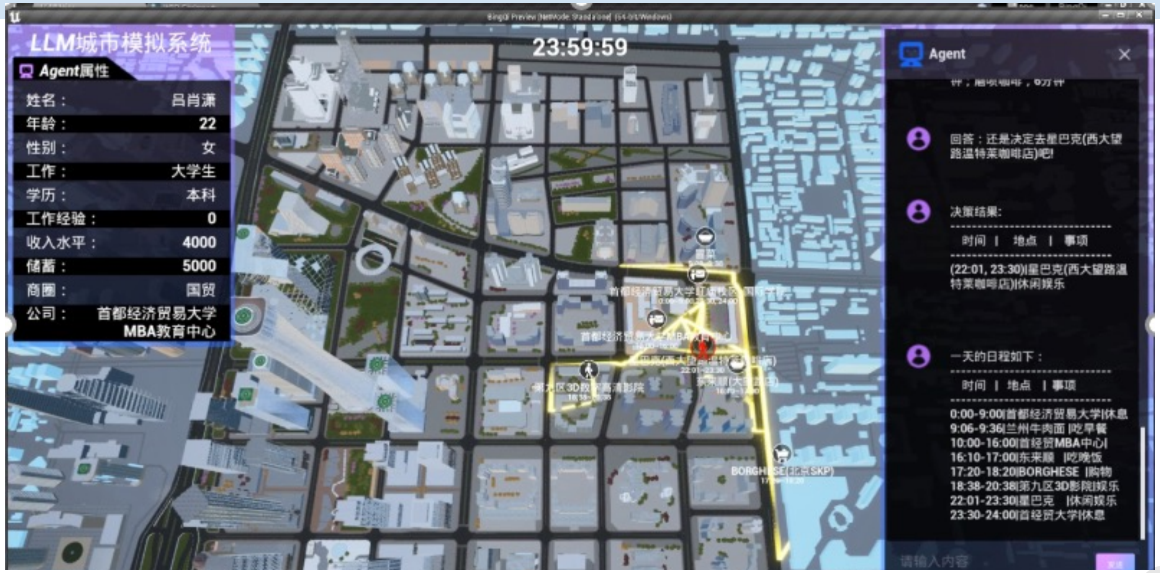
Complex, Deliberated, Coherent.

Reliable, Simple, Efficient.

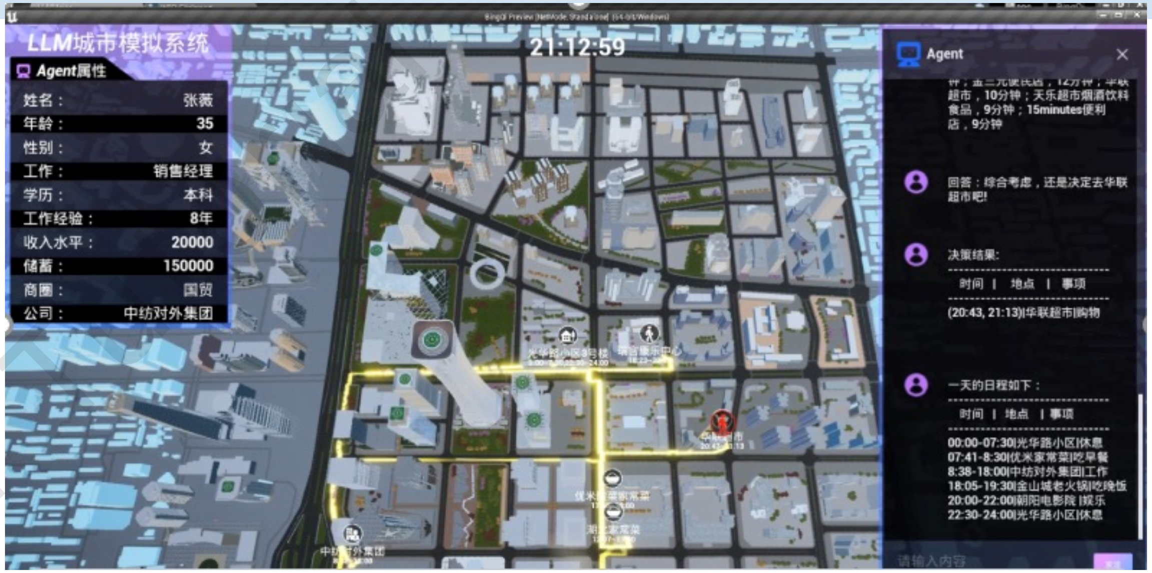


Embodied, Standardized, Intuitive.

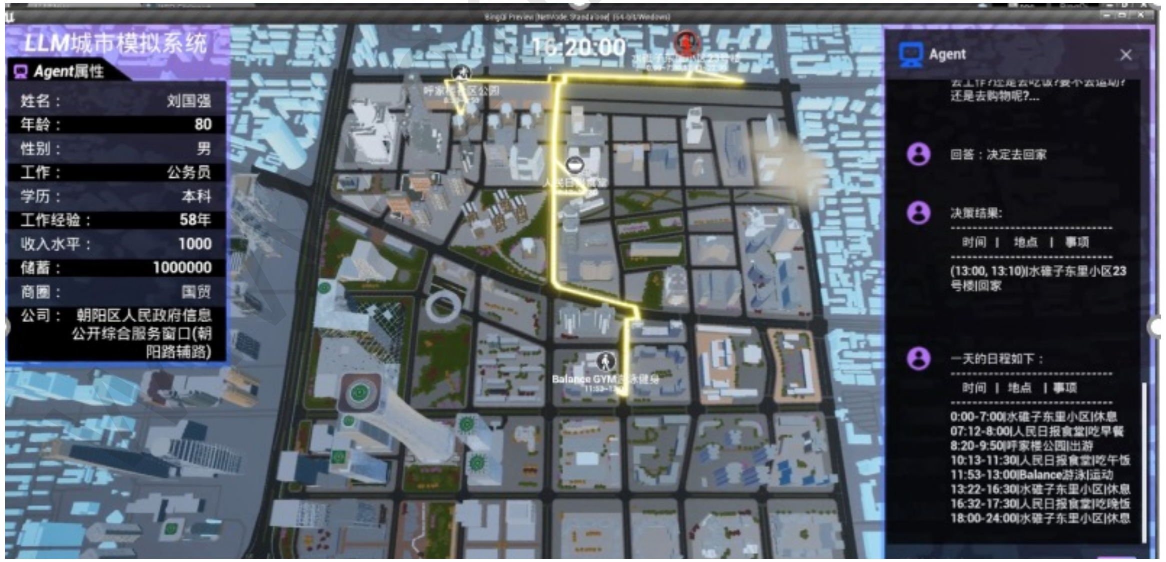
Building a Platform for City Simulation



College Student



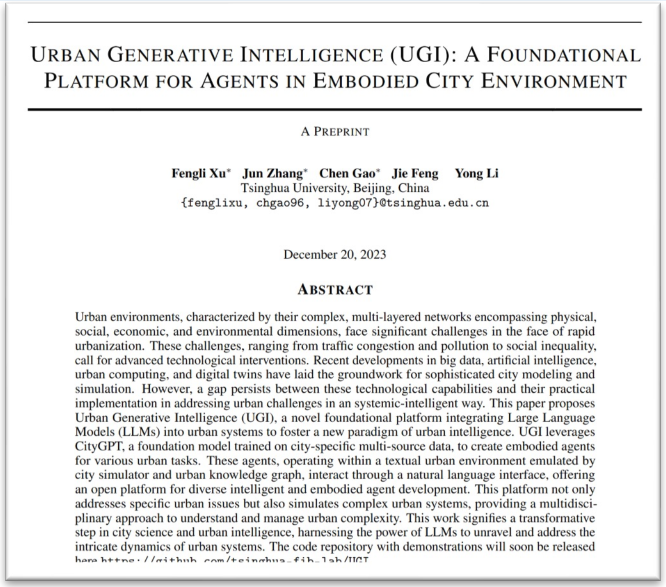
White Collar



Retired Person

A 2D Web Portal

https://opencity.fiblab.net/



URBAN GENERATIVE INTELLIGENCE (UGI): A FOUNDATIONAL PLATFORM FOR AGENTS IN EMBODIED CITY ENVIRONMENT

A PREPRINT

Fengli Xu* Jun Zhang* Chen Gao* Jie Feng Yong Li
Tsinghua University, Beijing, China
{fenglixu, chgao96, liyong07}@tsinghua.edu.cn

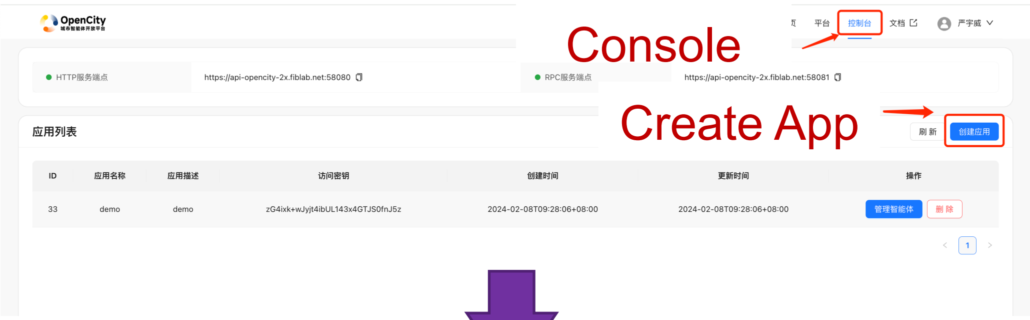
December 20, 2023

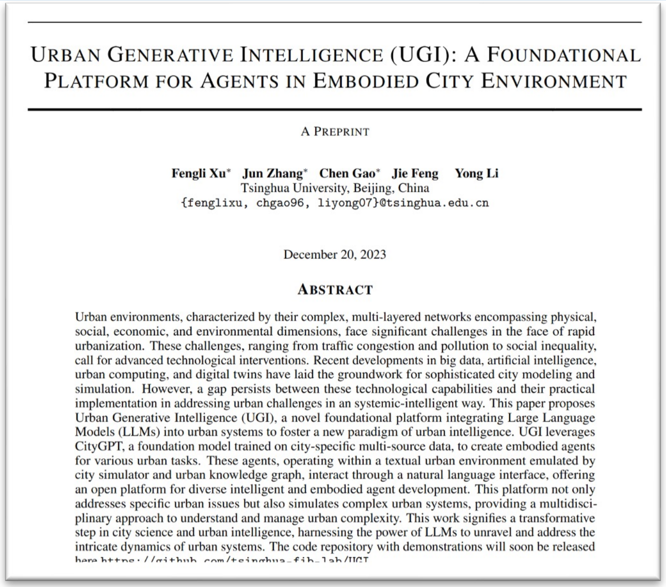
ABSTRACT

Urban environments, characterized by their complex, multi-layered networks encompassing physical, social, economic, and environmental dimensions, face significant challenges in the face of rapid urbanization. These challenges, ranging from traffic congestion and pollution to social inequality, call for advanced technological interventions. Recent developments in big data, artificial intelligence, urban computing, and digital twins have laid the groundwork for sophisticated city modeling and simulation. However, a gap persists between these technological capabilities and their practical implementation in addressing urban challenges in an systemic-intelligent way. This paper proposes Urban Generative Intelligence (UGI), a novel foundational platform integrating Large Language Models (LLMs) into urban systems to foster a new paradigm of urban intelligence. UGI leverages CityGPT, a foundation model trained on city-specific multi-source data, to create embodied agents for various urban tasks. These agents, operating within a textual urban environment emulated by city simulator and urban knowledge graph, interact through a natural language interface, offering an open platform for diverse intelligent and embodied agent development. This platform not only addresses specific urban issues but also simulates complex urban systems, providing a multidisciplinary approach to understand and manage urban complexity. This work signifies a transformative step in city science and urban intelligence, harnessing the power of LLMs to unravel and address the intricate dynamics of urban systems. The code repository with demonstrations will soon be released here https://github.com/tsinghua-fib-lab/UGI

Urban Generative Intelligence

A 2D Web Portal





URBAN GENERATIVE INTELLIGENCE (UGI): A FOUNDATIONAL PLATFORM FOR AGENTS IN EMBODIED CITY ENVIRONMENT

A PREPRINT

Fengli Xu* Jun Zhang* Chen Gao* Jie Feng Yong Li
Tsinghua University, Beijing, China
{fenglixu, chgao96, liyong07}@tsinghua.edu.cn

December 20, 2023

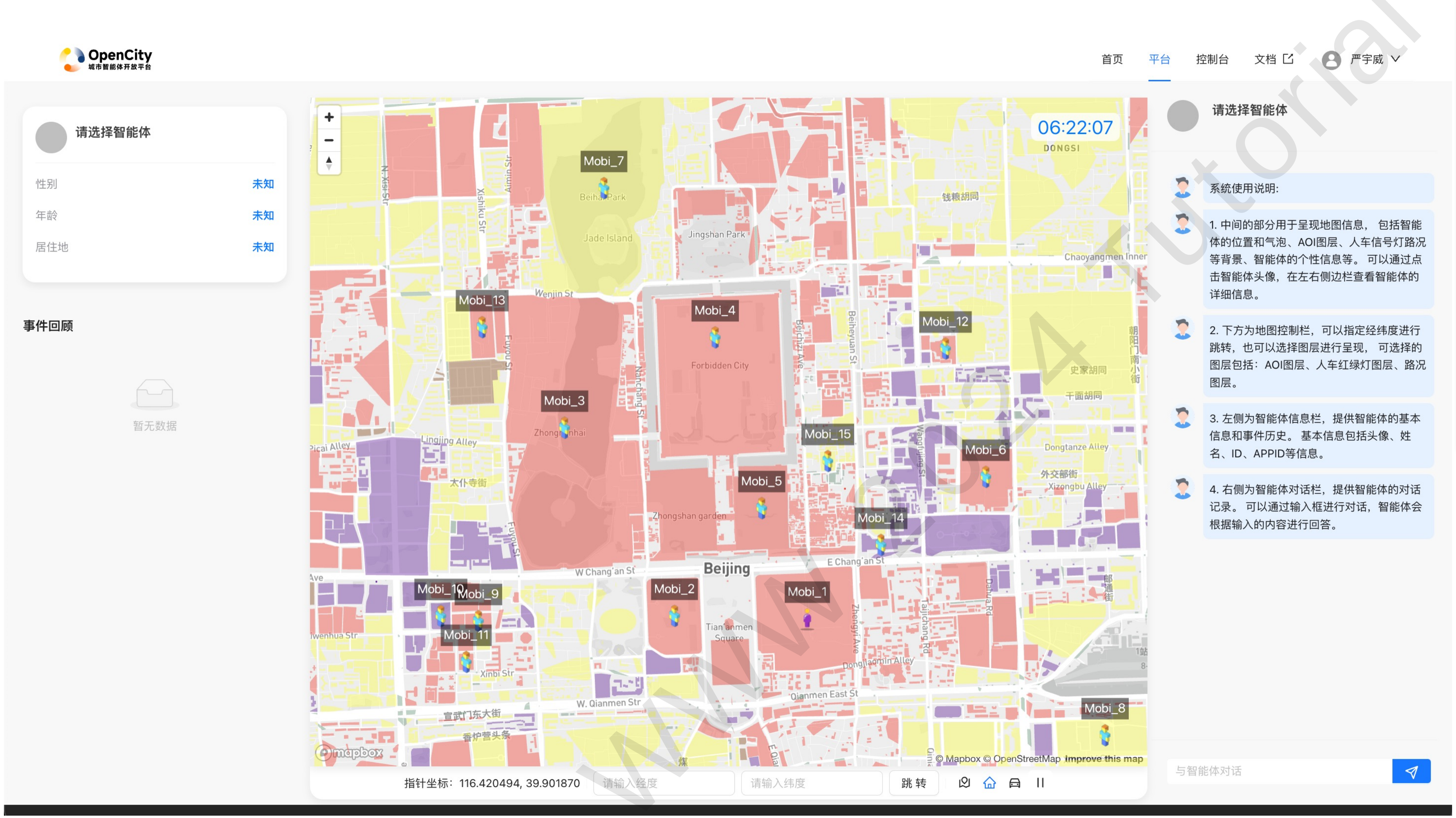
ABSTRACT

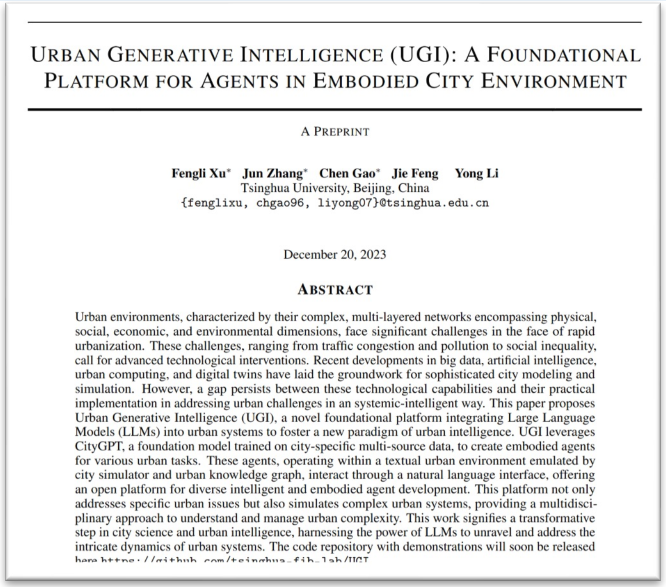
Urban environments, characterized by their complex, multi-layered networks encompassing physical, social, economic, and environmental dimensions, face significant challenges in the face of rapid urbanization. These challenges, ranging from traffic congestion and pollution to social inequality, call for advanced technological interventions. Recent developments in big data, artificial intelligence, urban computing, and digital twins have laid the groundwork for sophisticated city modeling and simulation. However, a gap persists between these technological capabilities and their practical implementation in addressing urban challenges in an systemic-intelligent way. This paper proposes Urban Generative Intelligence (UGI), a novel foundational platform integrating Large Language Models (LLMs) into urban systems to foster a new paradigm of urban intelligence. UGI leverages CityGPT, a foundation model trained on city-specific multi-source data, to create embodied agents for various urban tasks. These agents, operating within a textual urban environment emulated by city simulator and urban knowledge graph, interact through a natural language interface, offering an open platform for diverse intelligent and embodied agent development. This platform not only addresses specific urban issues but also simulates complex urban systems, providing a multidisciplinary approach to understand and manage urban complexity. This work signifies a transformative step in city science and urban intelligence, harnessing the power of LLMs to unravel and address the intricate dynamics of urban systems. The code repository with demonstrations will soon be released here https://github.com/tsinghua-fib-lab/UGI

Urban Generative Intelligence

https://opencity.fiblab.net/

A 2D Web Portal





URBAN GENERATIVE INTELLIGENCE (UGI): A FOUNDATIONAL PLATFORM FOR AGENTS IN EMBODIED CITY ENVIRONMENT

A PREPRINT

Fengli Xu* Jun Zhang* Chen Gao* Jie Feng Yong Li
Tsinghua University, Beijing, China
{fenglixu, chgao96, liyong07}@tsinghua.edu.cn

December 20, 2023

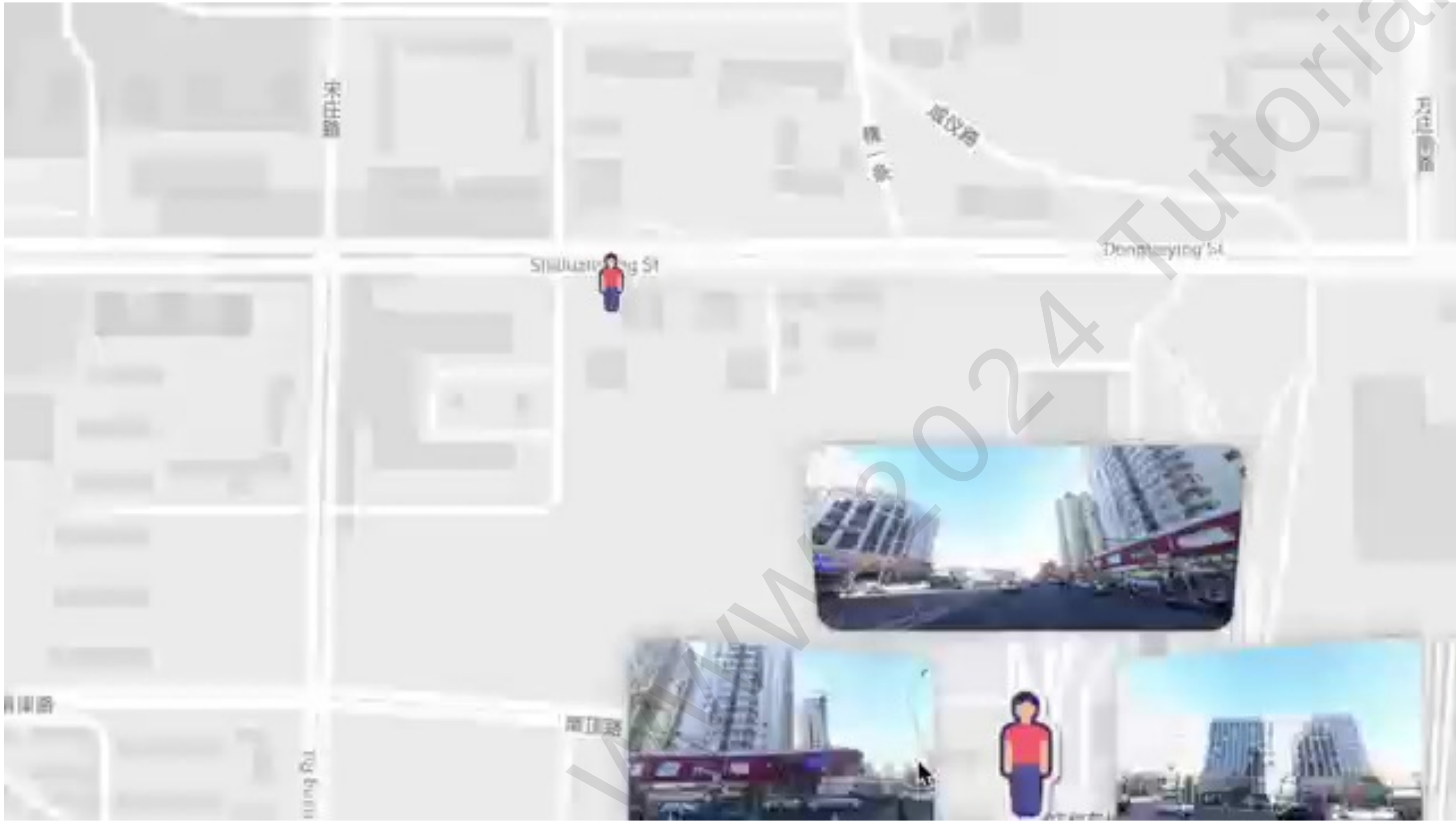
ABSTRACT

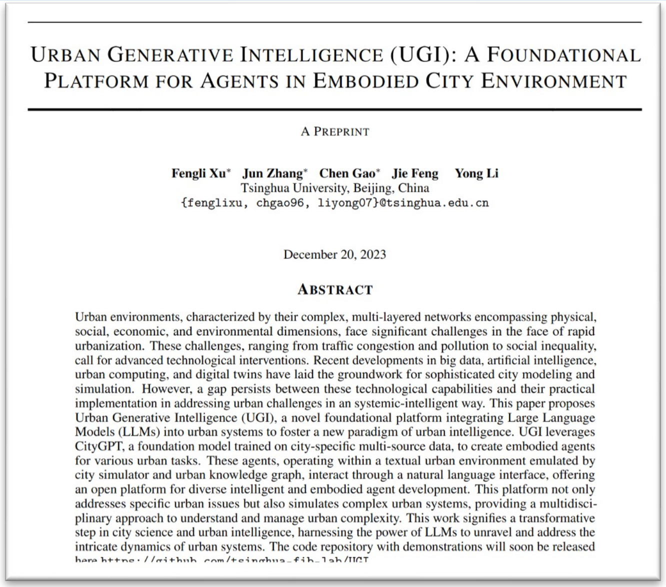
Urban environments, characterized by their complex, multi-layered networks encompassing physical, social, economic, and environmental dimensions, face significant challenges in the face of rapid urbanization. These challenges, ranging from traffic congestion and pollution to social inequality, call for advanced technological interventions. Recent developments in big data, artificial intelligence, urban computing, and digital twins have laid the groundwork for sophisticated city modeling and simulation. However, a gap persists between these technological capabilities and their practical implementation in addressing urban challenges in an systemic-intelligent way. This paper proposes Urban Generative Intelligence (UGI), a novel foundational platform integrating Large Language Models (LLMs) into urban systems to foster a new paradigm of urban intelligence. UGI leverages CityGPT, a foundation model trained on city-specific multi-source data, to create embodied agents for various urban tasks. These agents, operating within a textual urban environment emulated by city simulator and urban knowledge graph, interact through a natural language interface, offering an open platform for diverse intelligent and embodied agent development. This platform not only addresses specific urban issues but also simulates complex urban systems, providing a multidisciplinary approach to understand and manage urban complexity. This work signifies a transformative step in city science and urban intelligence, harnessing the power of LLMs to unravel and address the intricate dynamics of urban systems. The code repository with demonstrations will soon be released here https://github.com/tsinghua-fib-lab/UGI

Urban Generative Intelligence

https://opencity.fiblab.net/

A 2D Web Portal





URBAN GENERATIVE INTELLIGENCE (UGI): A FOUNDATIONAL PLATFORM FOR AGENTS IN EMBODIED CITY ENVIRONMENT

A PREPRINT

Fengli Xu* Jun Zhang* Chen Gao* Jie Feng Yong Li
Tsinghua University, Beijing, China
{fenglixu, chgao96, liyong07}@tsinghua.edu.cn

December 20, 2023

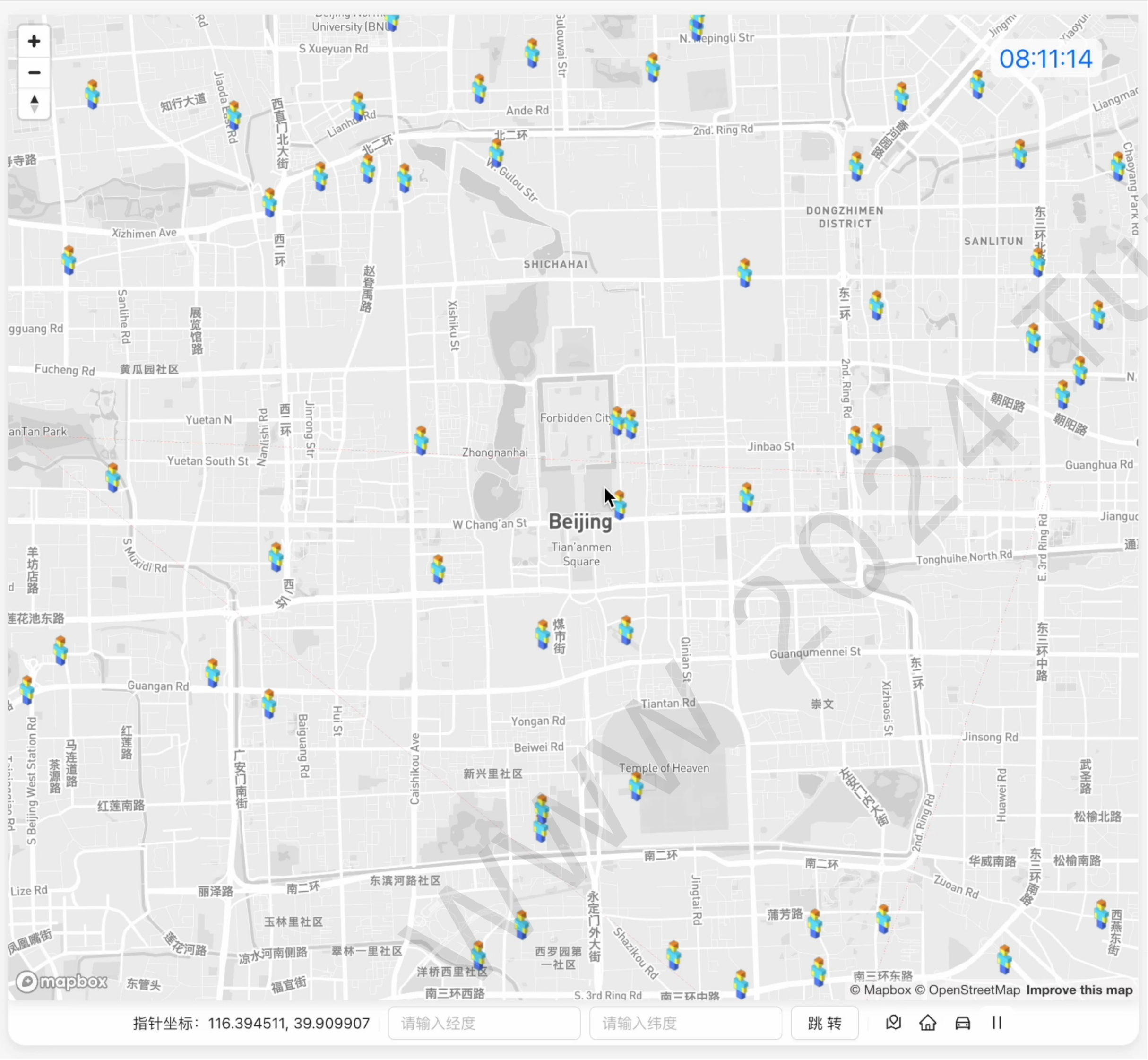
ABSTRACT

Urban environments, characterized by their complex, multi-layered networks encompassing physical, social, economic, and environmental dimensions, face significant challenges in the face of rapid urbanization. These challenges, ranging from traffic congestion and pollution to social inequality, call for advanced technological interventions. Recent developments in big data, artificial intelligence, urban computing, and digital twins have laid the groundwork for sophisticated city modeling and simulation. However, a gap persists between these technological capabilities and their practical implementation in addressing urban challenges in an systemic-intelligent way. This paper proposes Urban Generative Intelligence (UGI), a novel foundational platform integrating Large Language Models (LLMs) into urban systems to foster a new paradigm of urban intelligence. UGI leverages CityGPT, a foundation model trained on city-specific multi-source data, to create embodied agents for various urban tasks. These agents, operating within a textual urban environment emulated by city simulator and urban knowledge graph, interact through a natural language interface, offering an open platform for diverse intelligent and embodied agent development. This platform not only addresses specific urban issues but also simulates complex urban systems, providing a multidisciplinary approach to understand and manage urban complexity. This work signifies a transformative step in city science and urban intelligence, harnessing the power of LLMs to unravel and address the intricate dynamics of urban systems. The code repository with demonstrations will soon be released

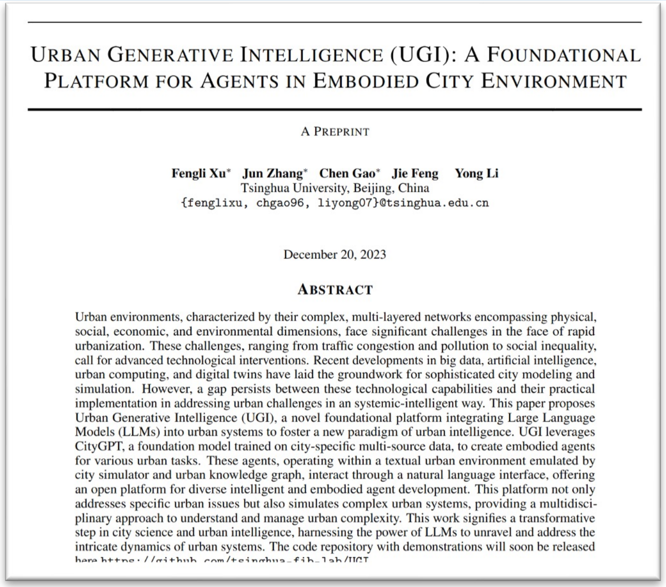
Urban Generative Intelligence

https://opencity.fiblab.net/

A 2D Web Portal



https://opencity.fiblab.net/


URBAN GENERATIVE INTELLIGENCE (UGI): A FOUNDATIONAL PLATFORM FOR AGENTS IN EMBODIED CITY ENVIRONMENT

A PREPRINT

Fengli Xu* Jun Zhang* Chen Gao* Jie Feng Yong Li
Tsinghua University, Beijing, China
{fenglixu, chgao96, liyong07}@tsinghua.edu.cn

December 20, 2023

ABSTRACT

Urban environments, characterized by their complex, multi-layered networks encompassing physical, social, economic, and environmental dimensions, face significant challenges in the face of rapid urbanization. These challenges, ranging from traffic congestion and pollution to social inequality, call for advanced technological interventions. Recent developments in big data, artificial intelligence, urban computing, and digital twins have laid the groundwork for sophisticated city modeling and simulation. However, a gap persists between these technological capabilities and their practical implementation in addressing urban challenges in an systemic-intelligent way. This paper proposes Urban Generative Intelligence (UGI), a novel foundational platform integrating Large Language Models (LLMs) into urban systems to foster a new paradigm of urban intelligence. UGI leverages CityGPT, a foundation model trained on city-specific multi-source data, to create embodied agents for various urban tasks. These agents, operating within a textual urban environment emulated by city simulator and urban knowledge graph, interact through a natural language interface, offering an open platform for diverse intelligent and embodied agent development. This platform not only addresses specific urban issues but also simulates complex urban systems, providing a multidisciplinary approach to understand and manage urban complexity. This work signifies a transformative step in city science and urban intelligence, harnessing the power of LLMs to unravel and address the intricate dynamics of urban systems. The code repository with demonstrations will soon be released here https://github.com/tsinghua-fib-lab/UGI

Urban Generative Intelligence

Open Discussions

Materials of this tutorial

Large Language Models Empowered Agent-based Modeling and Simulation: A Survey and Perspectives

Chen Gao Xiaochong Lan Nian Li Yuan Yuan Jingtao Ding Zhilun Zhou
Fengli Xu Yong Li
Tsinghua University, Beijing, China
{chgao96, fenglixu, liyong07}@tsinghua.edu.cn

Abstract

Agent-based modeling and simulat
complex systems, offering insights
diverse agents. Integrating large
and simulation presents a promisi
This paper surveys the landscap
based modeling and simulation, e
directions. In this survey, since th
the background of agent-based mo
empowered agents. We then disc
models to agent-based simulatio
environment perception, human ali
importantly, we provide a compr
language model-empowered age
scenarios, which can be divided i
hybrid, covering simulation of bo
since this area is new and quickl
promising future directions.

Chen Gao, et al. **Large Language Models Empowered Agent-based Modeling and Simulation: A Survey and Perspectives,** arXiv, 2023

WWW 2024 Tutorial 报告: 基于大模型智能体的社会模拟仿真

数据科学与智能实验室 2024-05-12 13:41

在即将于新加坡召开的 WWW 2024 会议上，我们将举办主题为《基于大模
真》（Simulating Human Society with LLM Agents: City, Social M
System）的 Tutorial 报告，欢迎参加！