SCHOOL OF COMPUTING & DATA SCIENCE
The University of Hong Kong

香港大學自然語言處理實驗室
Natural Language Processing Group, The University of Hong Kong

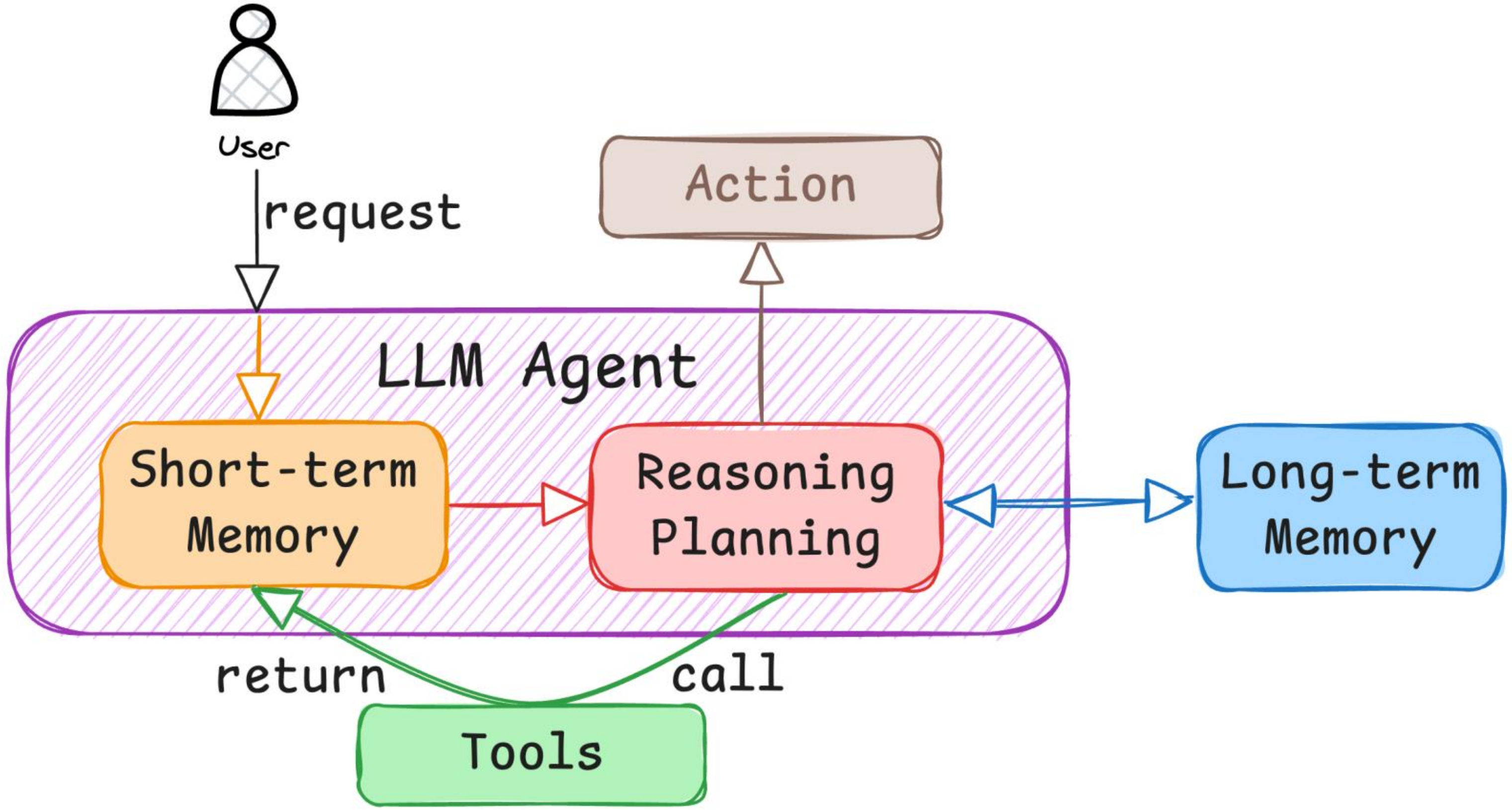
Advancing the Social Intelligence of LLM-based Agents

Xiachong Feng, Lingpeng Kong

27/03/2025

LLM-based Agent

- LLM agents are AI systems that leverage Large Language Models (LLMs), tools, and memory to perform tasks, make decisions, and interact with users or other systems autonomously.



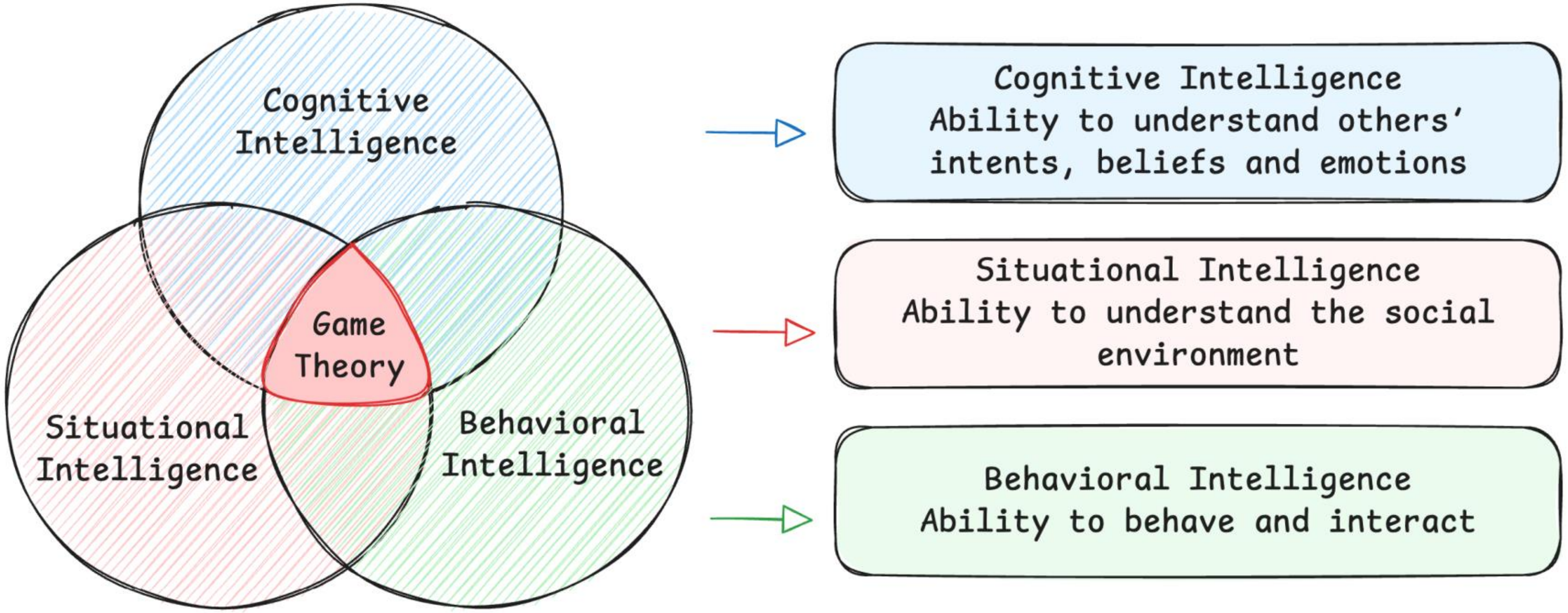
Human-AI Symbiotic Society

- The progress of LLMs brings the realization of Artificial General Intelligence (AGI) within reach paving the way for a future where human-AI interaction, collaboration, and coexistence shape a shared, symbiotic society.

Generated by DALL-E

Social Intelligence

- Social intelligence is the foundation of all successful interpersonal relationships and is also a prerequisite for AGI.
- Evaluations in game-theoretic scenarios require social agents to understand the game scenario, infer opponents’ actions, and adopt appropriate responses, representing an advanced form of social intelligence.



Cognitive Intelligence
Ability to understand others‘ intents, beliefs and emotions

Situational Intelligence
Ability to understand the social environment

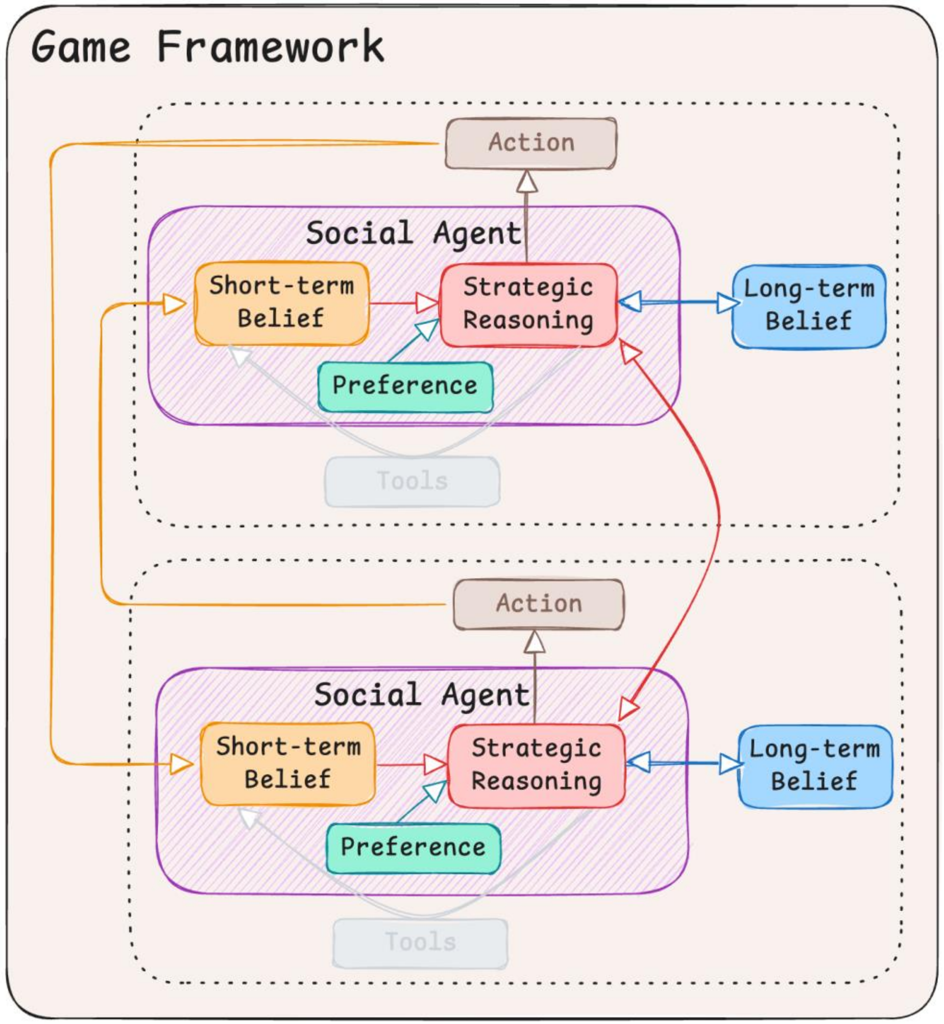
Behavioral Intelligence
Ability to behave and interact

Social Agent

- Preference refers to an individual's subjective inclination toward certain things, reflecting personal tastes, values, or choices in decision-making.

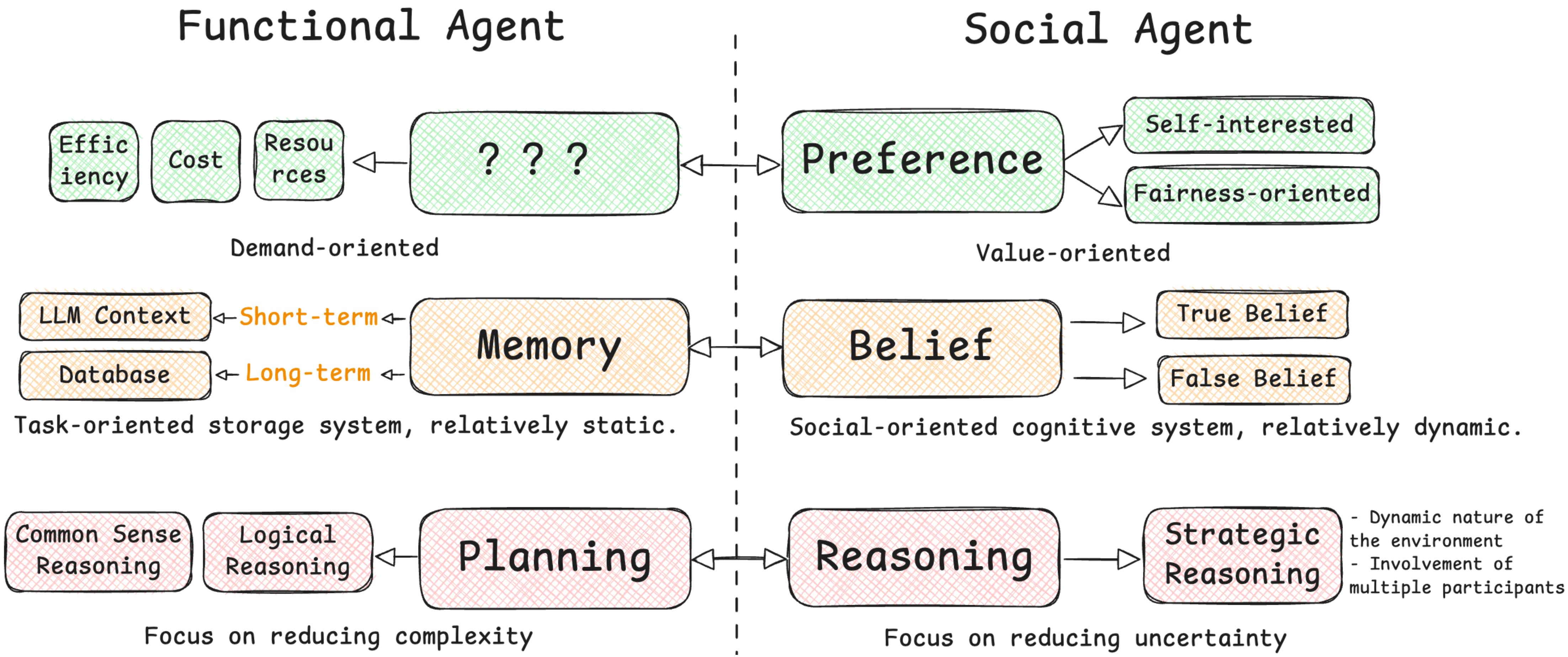
- Beliefs represent an agent's informational (or mental) state about the world, encompassing its understanding of itself and other agents, and consist of the facts or knowledge the agent considers true

- Reasoning refers to the process of inferring actions based on one's preferences and beliefs, as well as the historical information of other agents.



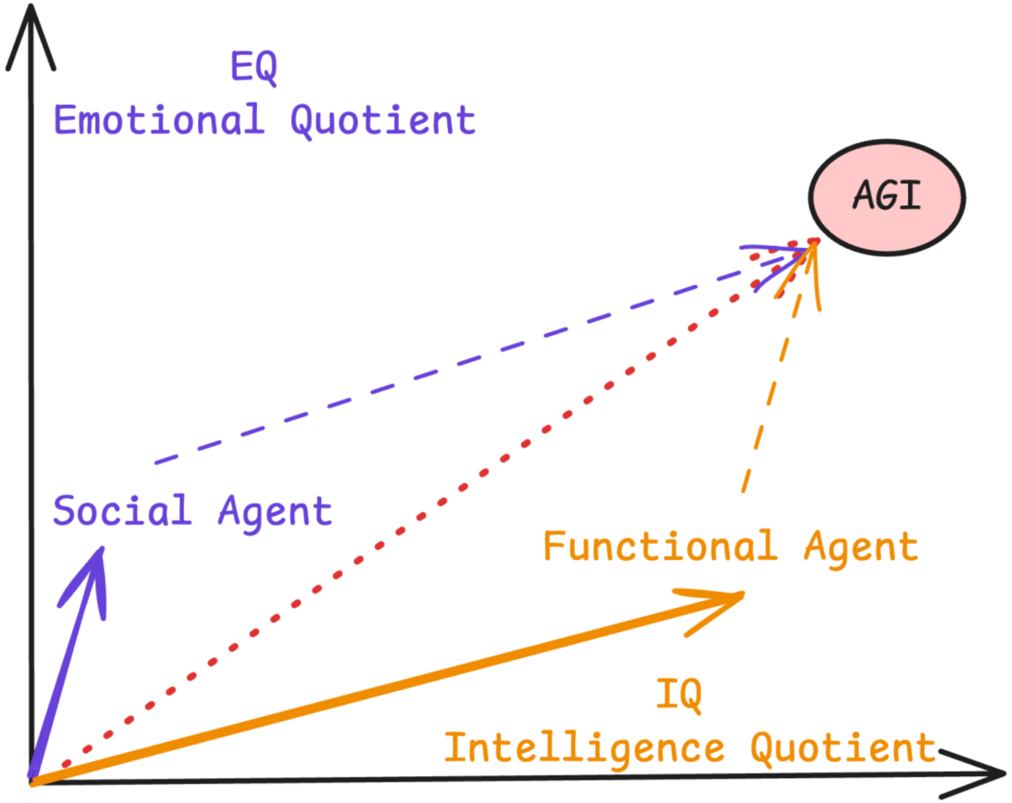
A Survey on Large Language Model-Based Social Agents in Game-Theoretic Scenarios

Functional Agent vs Social Agent

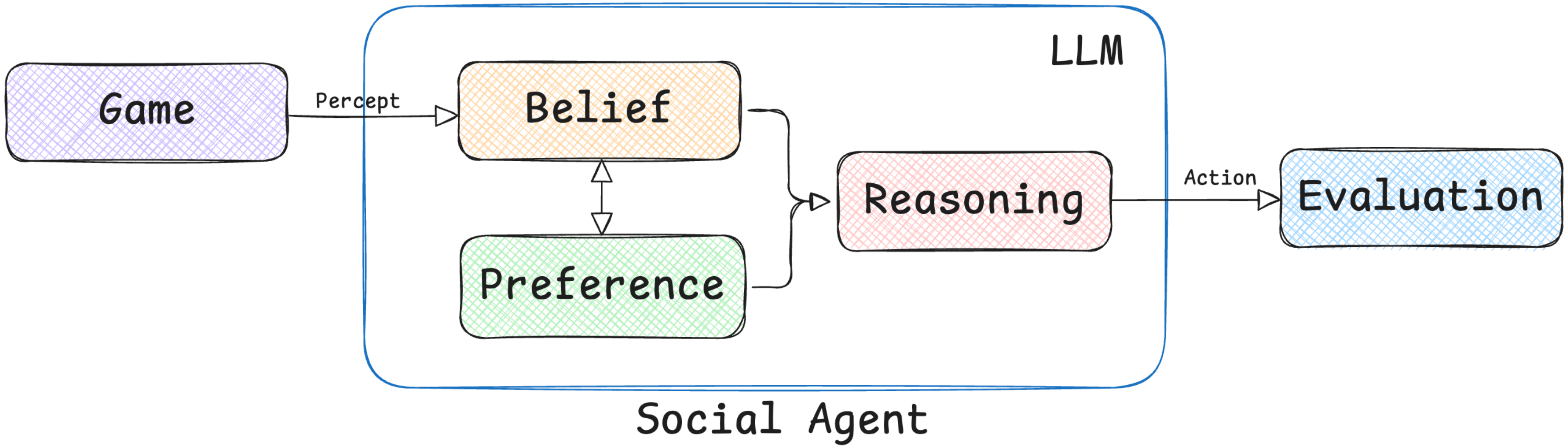


Functional Agent and Social Agent

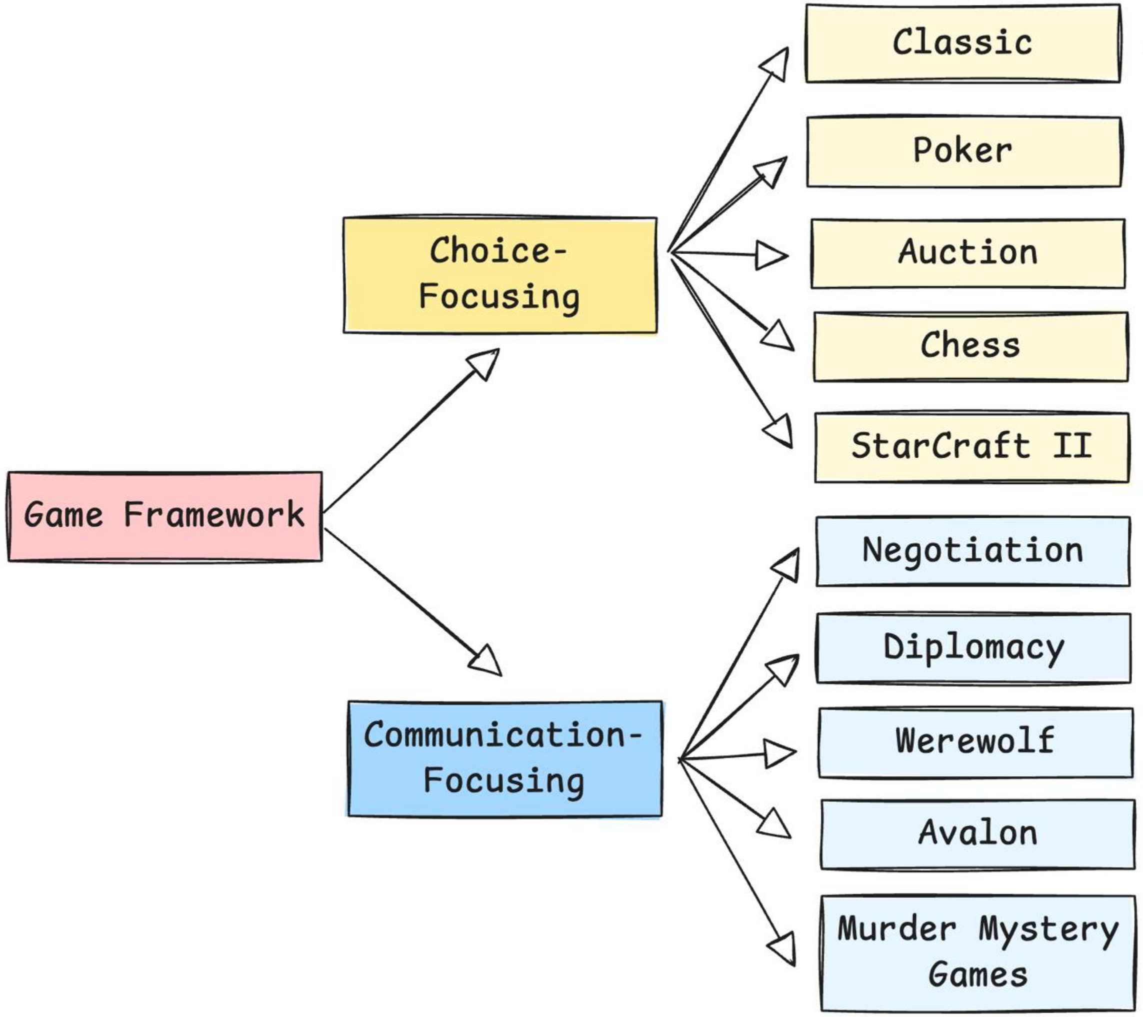
- The general artificial intelligence of the future should be a superintelligent agent that integrates both exceptionally high IQ and EQ.

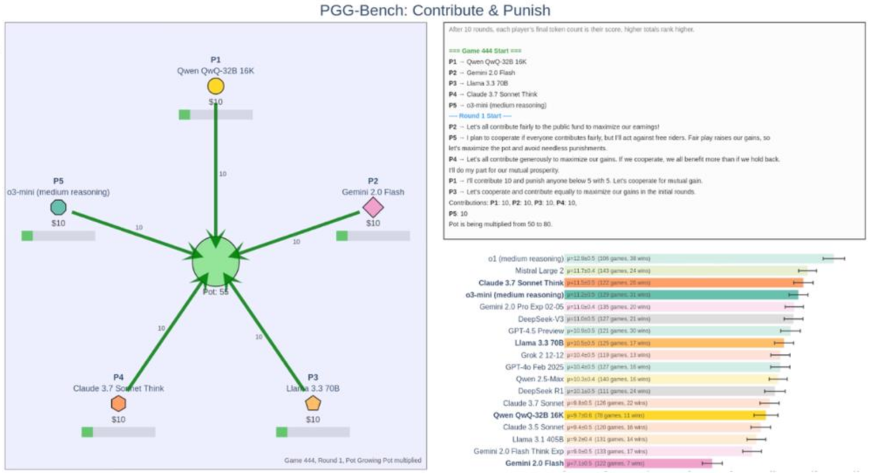


Key Questions in Social Agent



Game Framework





After 10 rounds, each player's final token count is their score; higher totals rank higher.

=== Game 444 Start ===

P1 → Qwen QwQ-32B 16K

P2 → Gemini 2.0 Flash

P3 → Llama 3.3 70B

P4 → Claude 3.7 Sonnet Think

P5 → o3-mini (medium reasoning)

—— Round 1 Start ——

P2 → Let's all contribute fairly to the public fund to maximize our earnings!

P5 → I plan to cooperate if everyone contributes fairly, but I'll act against free riders. Fair play raises our gains, so let's maximize the pot and avoid needless punishments.

P4 → Let's all contribute generously to maximize our gains. If we cooperate, we all benefit more than if we hold back. I'll do my part for our mutual prosperity.

P1 → I'll contribute 10 and punish anyone below 5 with 5. Let's cooperate for mutual gain.

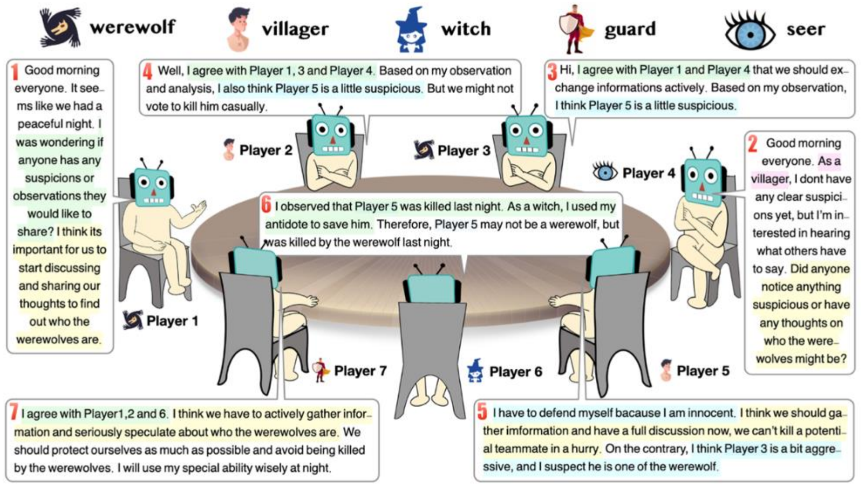
P3 → Let's cooperate and contribute equally to maximize our gains in the initial rounds.

Contributions: **P1**: 10, **P2**: 10, **P3**: 10, **P4**: 10,

P5: 10

Pot is being multiplied from 50 to 80.

o1 (medium reasoning) μ=12.9±0.5 (106 games, 38 wins)
Mistral Large 2 μ=11.7±0.4 (143 games, 24 wins)
Claude 3.7 Sonnet Think μ=11.5±0.5 (122 games, 26 wins)
o3-mini (medium reasoning) μ=11.2±0.5 (129 games, 31 wins)
Gemini 2.0 Pro Exp 02-05 μ=11.0±0.4 (135 games, 20 wins)
DeepSeek-V3 μ=11.0±0.5 (127 games, 21 wins)
GPT-4.5 Preview μ=10.9±0.5 (121 games, 30 wins)
Llama 3.3 70B μ=10.5±0.5 (125 games, 17 wins)
Grok 2 12-12 μ=10.4±0.5 (119 games, 13 wins)
GPT-4o Feb 2025 μ=10.4±0.5 (127 games, 16 wins)
Qwen 2.5-Max μ=10.3±0.4 (140 games, 16 wins)
DeepSeek R1 μ=10.1±0.5 (111 games, 24 wins)
Claude 3.7 Sonnet μ=9.8±0.5 (126 games, 22 wins)
Qwen QwQ-32B 16K μ=9.7±0.6 (78 games, 11 wins)
Claude 3.5 Sonnet μ=9.4±0.5 (120 games, 16 wins)
Llama 3.1 405B μ=9.2±0.4 (131 games, 14 wins)
Gemini 2.0 Flash Think Exp μ=9.0±0.5 (133 games, 17 wins)
Gemini 2.0 Flash μ=7.1±0.5 (122 games, 7 wins)


werewolf
villager
witch
guard
seer
Player 1
Player 2
Player 3
Player 4
Player 5
Player 6
Player 7

1 Good morning everyone. It seems like we had a peaceful night. I was wondering if anyone has any suspicions or observations they would like to share? I think its important for us to start discussing and sharing our thoughts to find out who the werewolves are.

4 Well, I agree with Player 1, 3 and Player 4. Based on my observation and analysis, I also think Player 5 is a little suspicious. But we might not vote to kill him casually.

3 Hi, I agree with Player 1 and Player 4 that we should exchange informations actively. Based on my observation, I think Player 5 is a little suspicious.

2 Good morning everyone. As a villager, I dont have any clear suspicions yet, but I'm interested in hearing what others have to say. Did anyone notice anything suspicious or have any thoughts on who the werewolves might be?

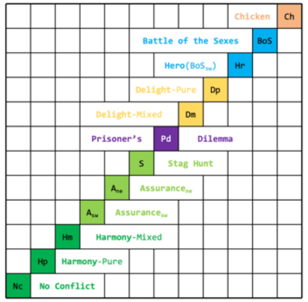
6 I observed that Player 5 was killed last night. As a witch, I used my antidote to save him. Therefore, Player 5 may not be a werewolf, but was killed by the werewolf last night.

7 I agree with Player1,2 and 6. I think we have to actively gather information and seriously speculate about who the werewolves are. We should protect ourselves as much as possible and avoid being killed by the werewolves. I will use my special ability wisely at night.

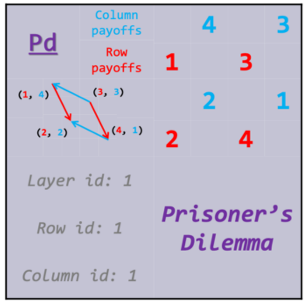
5 I have to defend myself bacause I am innocent. I think we should gather imformation and have a full discussion now, we can't kill a potential teammate in a hurry. On the contrary, I think Player 3 is a bit aggressive, and I suspect he is one of the werewolf.

Choice-Focusing: **TMGBench**

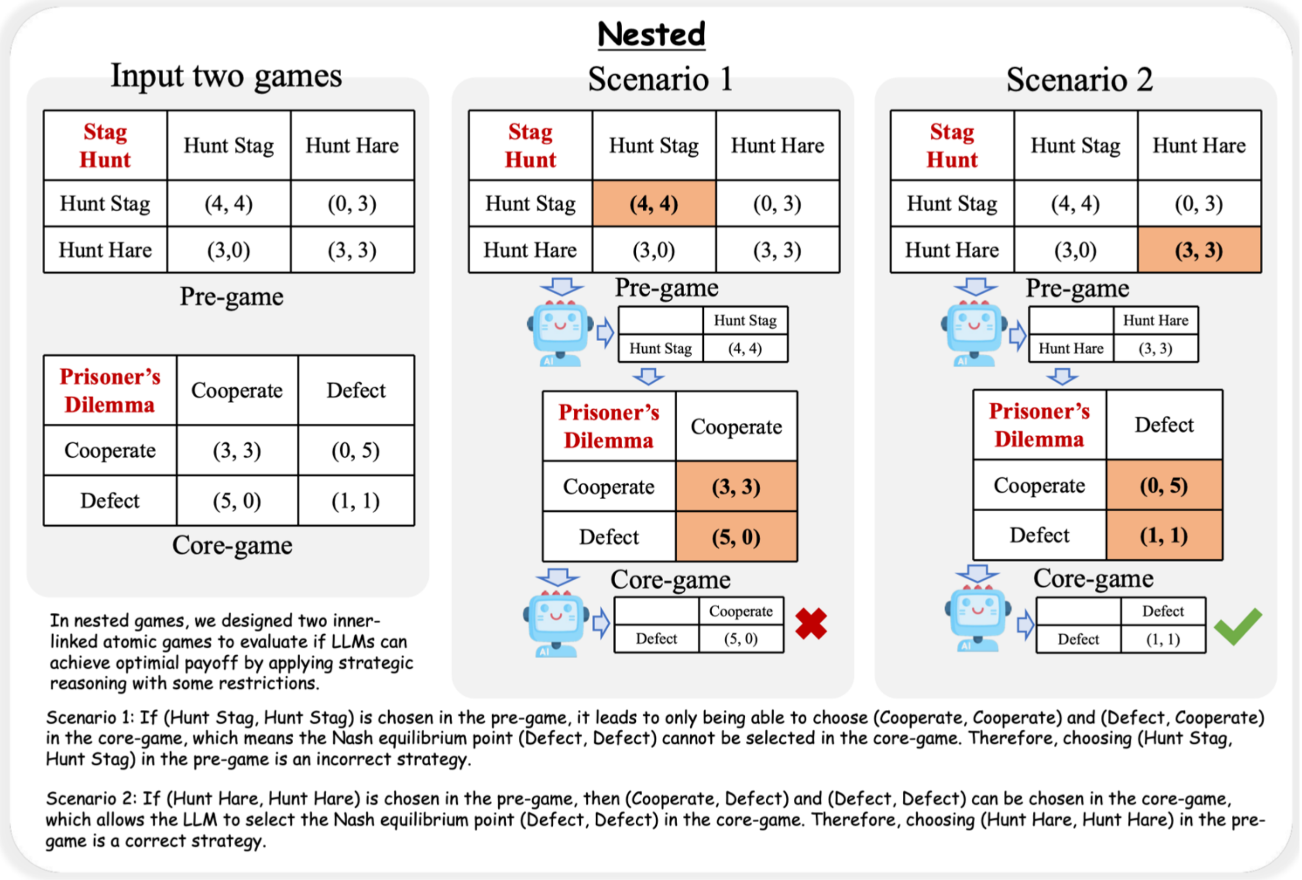
- Advanced LLMs like GPT-4o and Claude 3.5 Sonnet struggle to generalize across diverse contexts and scenarios.
- Complex-form games derived from atomic units in TMGBench pose significant challenges for LLMs — including DeepSeek-R1 and O1-mini — which often falter as the number of games increases.



(a) Most Famous Games



(b) Details in a Grid



Nested

Input two games

Stag Hunt	Hunt Stag	Hunt Hare
Hunt Stag	(4, 4)	(0, 3)
Hunt Hare	(3,0)	(3, 3)

Pre-game

Prisoner's Dilemma	Cooperate	Defect
Cooperate	(3, 3)	(0, 5)
Defect	(5, 0)	(1, 1)

Core-game

Scenario 1

Stag Hunt	Hunt Stag	Hunt Hare
Hunt Stag	**(4, 4)**	(0, 3)
Hunt Hare	(3,0)	(3, 3)

Pre-game

	Hunt Stag
Hunt Stag	(4, 4)

Prisoner's Dilemma	Cooperate
Cooperate	**(3, 3)**
Defect	**(5, 0)**

Core-game

	Cooperate
Defect	(5, 0)

✗

Scenario 2

Stag Hunt	Hunt Stag	Hunt Hare
Hunt Stag	(4, 4)	(0, 3)
Hunt Hare	(3,0)	**(3, 3)**

Pre-game

	Hunt Hare
Hunt Hare	(3, 3)

Prisoner's Dilemma	Defect
Cooperate	**(0, 5)**
Defect	**(1, 1)**

Core-game

	Defect
Defect	(1, 1)

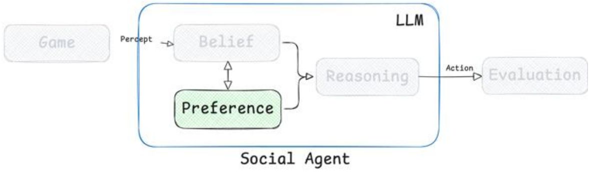
✓

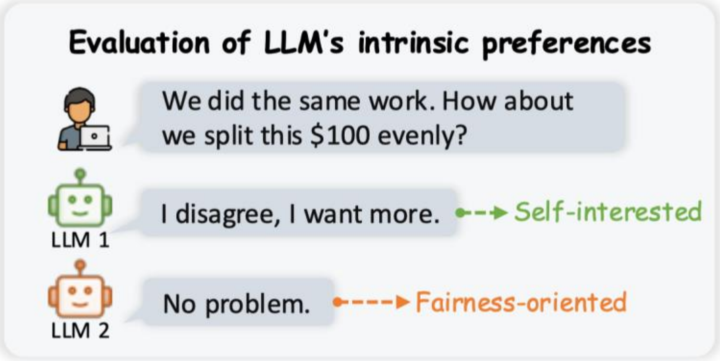
In nested games, we designed two inner-linked atomic games to evaluate if LLMs can achieve optimal payoff by applying strategic reasoning with some restrictions.

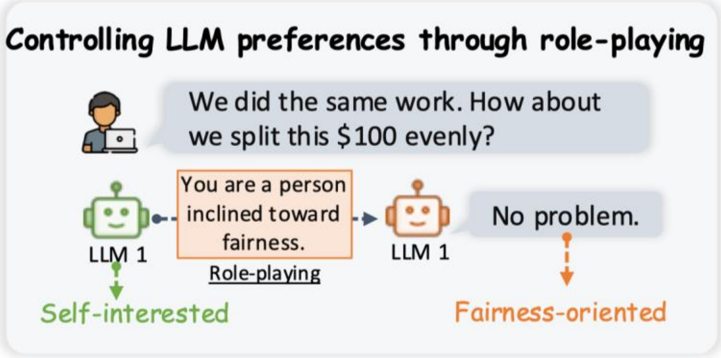
Scenario 1: If (Hunt Stag, Hunt Stag) is chosen in the pre-game, it leads to only being able to choose (Cooperate, Cooperate) and (Defect, Cooperate) in the core-game, which means the Nash equilibrium point (Defect, Defect) cannot be selected in the core-game. Therefore, choosing (Hunt Stag, Hunt Stag) in the pre-game is an incorrect strategy.

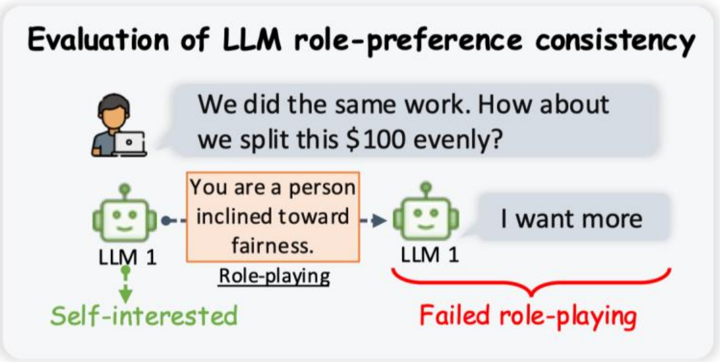
Scenario 2: If (Hunt Hare, Hunt Hare) is chosen in the pre-game, then (Cooperate, Defect) and (Defect, Defect) can be chosen in the core-game, which allows the LLM to select the Nash equilibrium point (Defect, Defect) in the core-game. Therefore, choosing (Hunt Hare, Hunt Hare) in the pre-game is a correct strategy.

Preference Module




Evaluation of LLM's intrinsic preferences
We did the same work. How about we split this $100 evenly?
I disagree, I want more.
Self-interested
LLM 1
No problem.
Fairness-oriented
LLM 2


Controlling LLM preferences through role-playing
We did the same work. How about we split this $100 evenly?
You are a person inclined toward fairness.
Role-playing
LLM 1
LLM 1
No problem.
Self-interested
Fairness-oriented


Evaluation of LLM role-preference consistency
We did the same work. How about we split this $100 evenly?
You are a person inclined toward fairness.
Role-playing
LLM 1
LLM 1
I want more
Self-interested
Failed role-playing

GPT-4 include reciprocity preferences, responsiveness to group identity cues, engagement in indirect reciprocity, and social learning capabilities. However, differences emerged as GPT-4 displayed a stronger inclination toward fairness than humans and responded decisively to negative stimuli, often retaliating against perceived uncooperative or harmful behaviours with heightened consistency.[1]

LLMs possess a basic ability to form clear preferences based on textual prompts. LLMs with high openness, conscientiousness, and neuroticism exhibited fair tendencies, while those with low agreeableness and low openness displayed rational tendencies, and low conscientiousness were associated with high toxicity. [2]

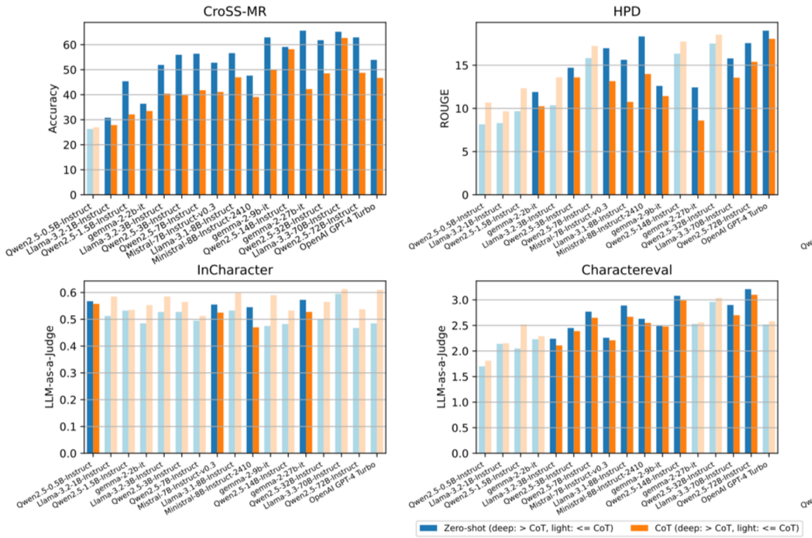
LLMs struggle with desires rooted in less common preferences.
Merely including persona details in the system prompt may not sufficiently capture the depth of certain personality preferences or the expertise of professional players, leading to lower consistency between strategic decision-making behaviour and preferences. [3]

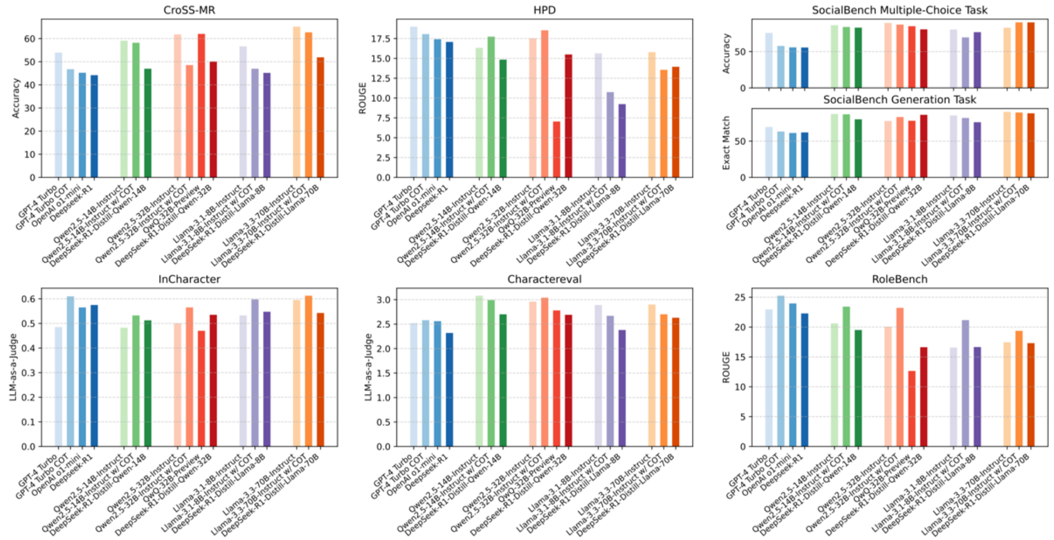
[1] Do llm agents exhibit social behavior?
[2] Llms with personalities in multi-issue negotiation games.
[3] Alympics: Language agents meet game theory.

Role-playing

- CoT may reduce the role-playing capabilities of LLMs.
- Reasoning-optimized LLMs are less suitable for role-playing tasks.

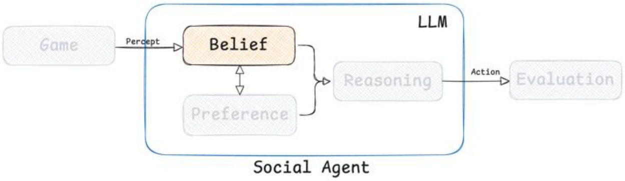
- (1) “Attention Diversion”: The model must simultaneously engage in reasoning and role-playing modes, which dilutes its focus on the role-playing task.
- (2) “Linguistic Style Drift”: Reasoning responses tend to be structured, logical, and formal, whereas effective role-playing requires a vivid, expressive, and character-consistent linguistic style.





Reasoning Does Not Necessarily Improve Role-Playing Ability

Belief Module



- Three key research questions:
  - Do agents possess internal beliefs?
  - How can the belief modelling capabilities of agents be enhanced?
  - Can agents revise their beliefs?

Example

Noor is working as a barista at a busy coffee shop. Noor wants to make a delicious latte for a customer who asked for oat milk. Noor grabs a milk pitcher and fills it with oat milk.

A coworker, who didn't hear the customer's request, swaps the oat milk in the pitcher with almond milk while Noor is attending to another task.

Scenario 1

Noor **does not see** her coworker swapping the milk.

What does Noor believe is in the milk pitcher?

Noor believes that the milk pitcher contains oat milk.

LLM

False Belief

Scenario 2

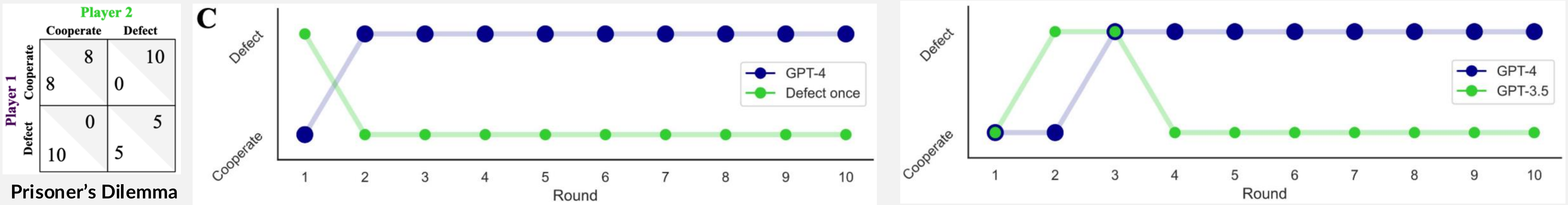
Noor **sees** her coworker swapping the milk.

What does Noor believe is in the milk pitcher?

Noor believes that the milk pitcher contains almond milk.

LLM

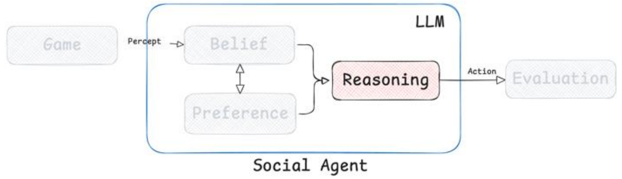
True Belief



Prisoner's Dilemma

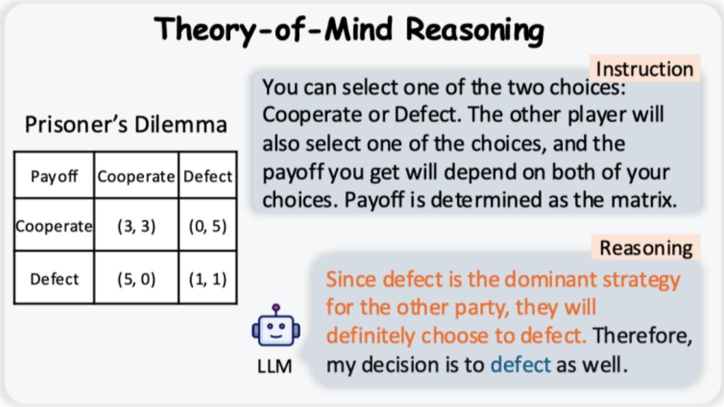
Playing repeated games with Large Language Models

Reasoning Module



- The involvement of multiple participants requires reasoning about the opponents' mental states.
  - Theory-of-Mind Reasoning

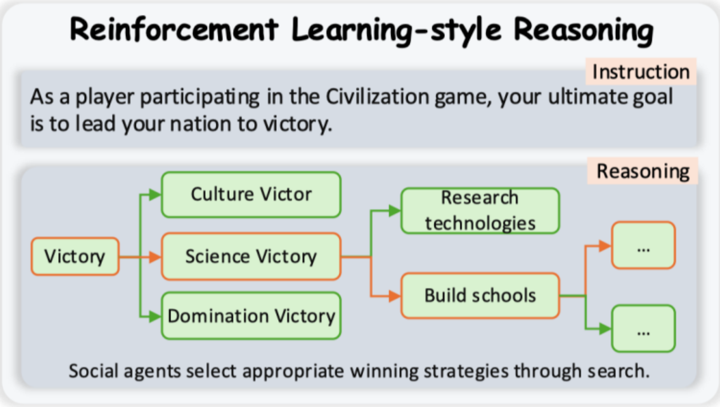
- The dynamic nature of the environment necessitates proactive exploration and evaluation of current and future possible states.
  - Reinforcement Learning-style Reasoning



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

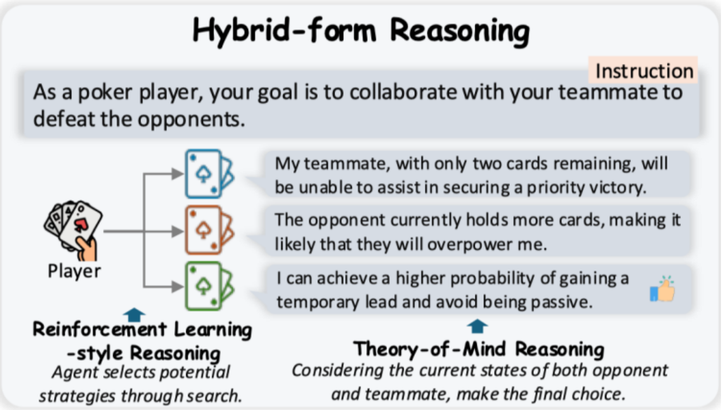
You can select one of the two choices: Cooperate or Defect. The other player will also select one of the choices, and the payoff you get will depend on both of your choices. Payoff is determined as the matrix.

Since defect is the dominant strategy for the other party, they will definitely choose to defect. Therefore, my decision is to defect as well.



As a player participating in the Civilization game, your ultimate goal is to lead your nation to victory.

Social agents select appropriate winning strategies through search.



As a poker player, your goal is to collaborate with your teammate to defeat the opponents.

My teammate, with only two cards remaining, will be unable to assist in securing a priority victory.

The opponent currently holds more cards, making it likely that they will overpower me.

I can achieve a higher probability of gaining a temporary lead and avoid being passive.

Agent selects potential strategies through search.

Considering the current states of both opponent and teammate, make the final choice.

Social Impact

Stage	Description	Potential Risks	Mitigation Strategies
Designing Social Agents	Focuses on creating the underlying algorithms that shape the agent's behavioral preferences.	Poorly designed algorithms may lead to negative behaviors (e.g., deception, manipulation, bias amplification).	✓ Enhance alignment algorithms (safety and moral alignment). ✓ Develop behavioral plugins as dynamic controllers.
Evaluating Social Agents	Involves rigorous testing of agents before real-world deployment to assess their behavior.	Agents with undetected negative behaviors (e.g., aggression, exploitation) may proceed to deployment.	✓ Evaluate agents in diverse game scenarios. ✓ Establish a benchmarking framework for behavioral assessment.
Deploying Social Agents	Covers the rollout of agents into real-world applications, starting with controlled environments.	Unforeseen negative consequences (e.g., misinformation, trust erosion) may emerge at scale.	✓ Start with low-risk, small-scale deployments. ✓ Gradually expand while monitoring anomalies in real time.
Supervising Social Agents	Ensures ongoing oversight and management of deployed agents to prevent harm.	Scalability of harm, impersonation, or subtle decision manipulation may go unchecked.	✓ Design automated monitoring systems for real-time surveillance. ✓ Use behavioral analysis for early warnings.

Conclusion

- Preference, belief, and reasoning are the three core modules within a social agent.
- Future work can continue to explore areas such as standardized benchmark generation, reinforcement learning agents, behavior pattern mining, and pluralistic game-theoretic scenarios.
- There is an urgent need for interdisciplinary research with the social sciences to clarify key scientific questions.
- Social agents are an essential pathway toward AGI, and more precise control as well as more effective simulation require further in-depth investigation.

SCHOOL OF
COMPUTING &
DATA SCIENCE
The University of Hong Kong

香港大學自然語言處理實驗室
Natural Language Processing Group, The University of Hong Kong

Thanks!