Social Agents in Game-Theoretic Scenarios

Xiachong Feng, Lingpeng Kong

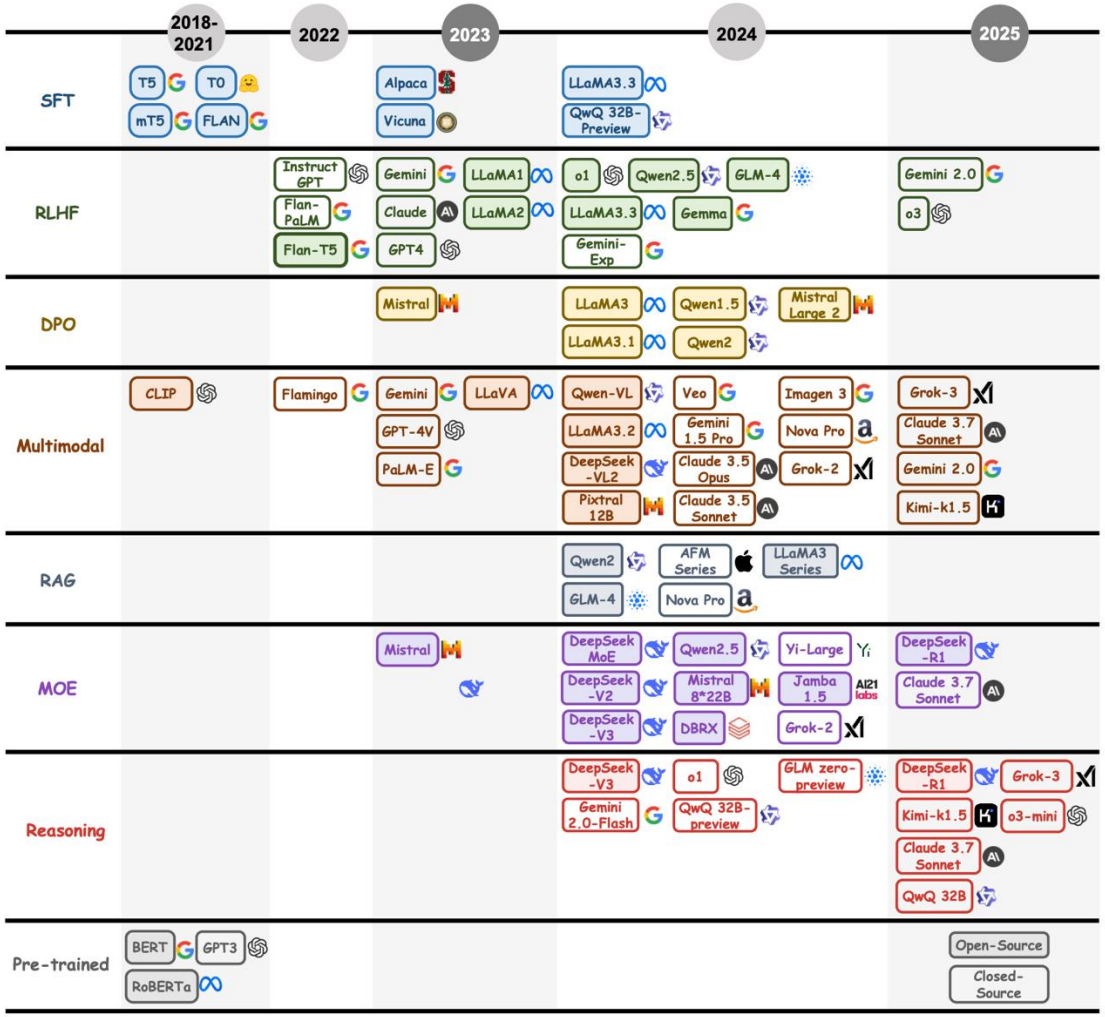
27/03/2025

Outline

- Functional Agent
- Social Agent
- Game Framework
- Preference Module
- Belief Module
- Reasoning Module
- Evaluation
- Broader Impact Statement
- Conclusion

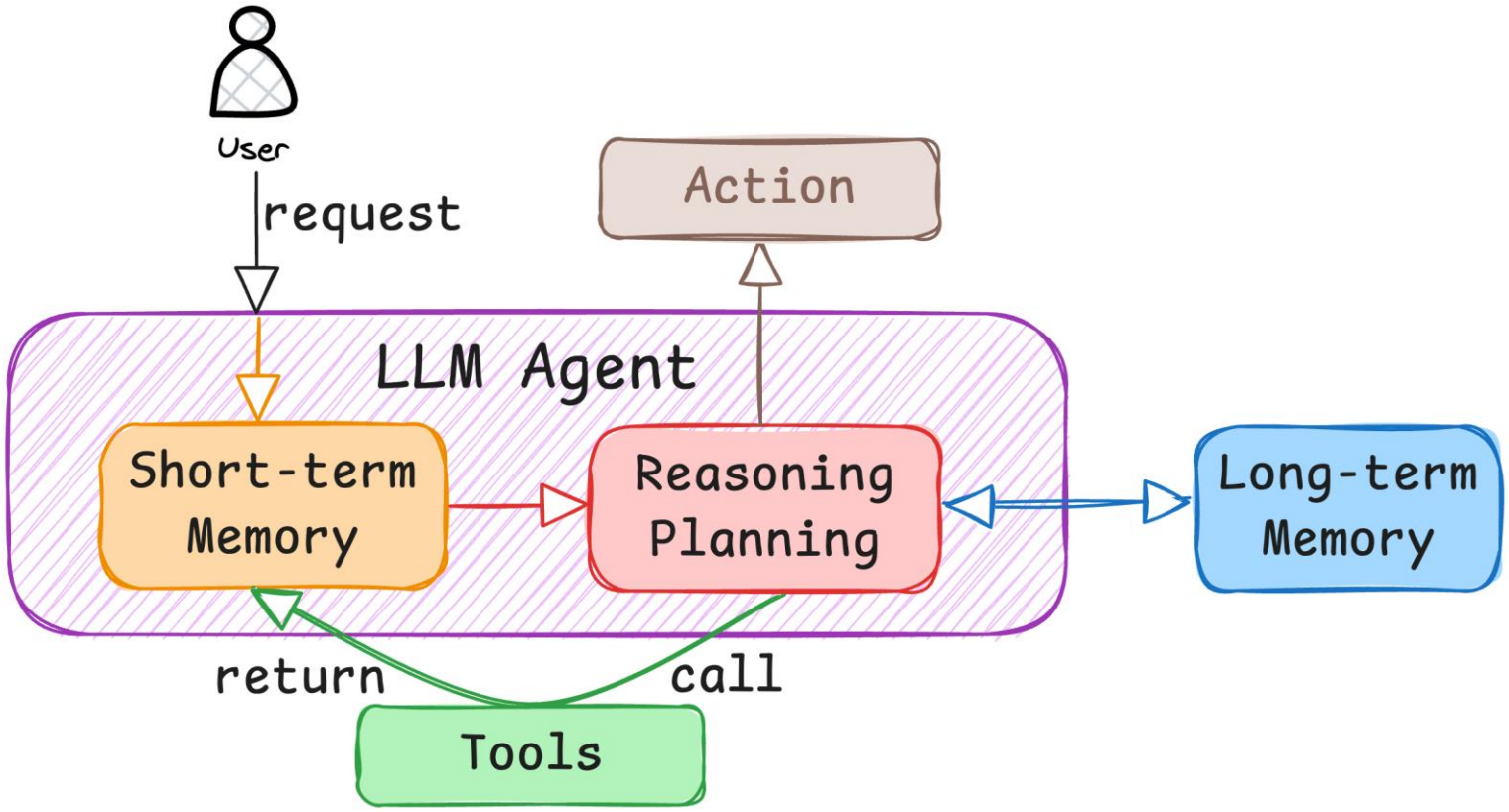
Large Language Models

- The rapid advancement of Large Language Models (LLMs) has achieved exceptional performance across a wide array of applications
- Recently, Large Reasoning Models (LRMs) have become a popular trend, including o3-mini, DeepSeek-R1 and QwQ.



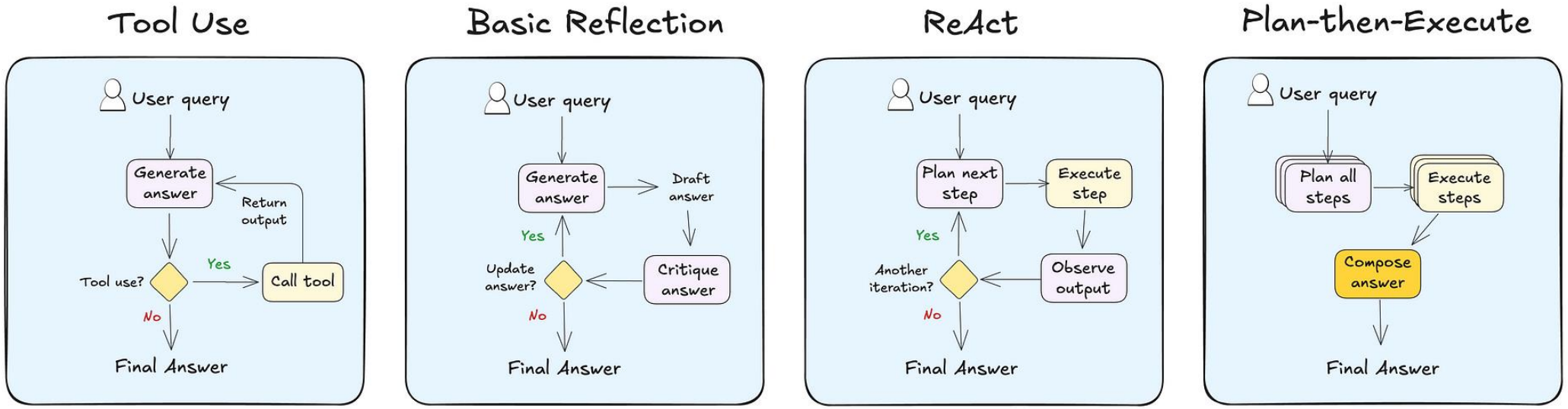
LLM-based Agent

- 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.



Common Agentic Patterns

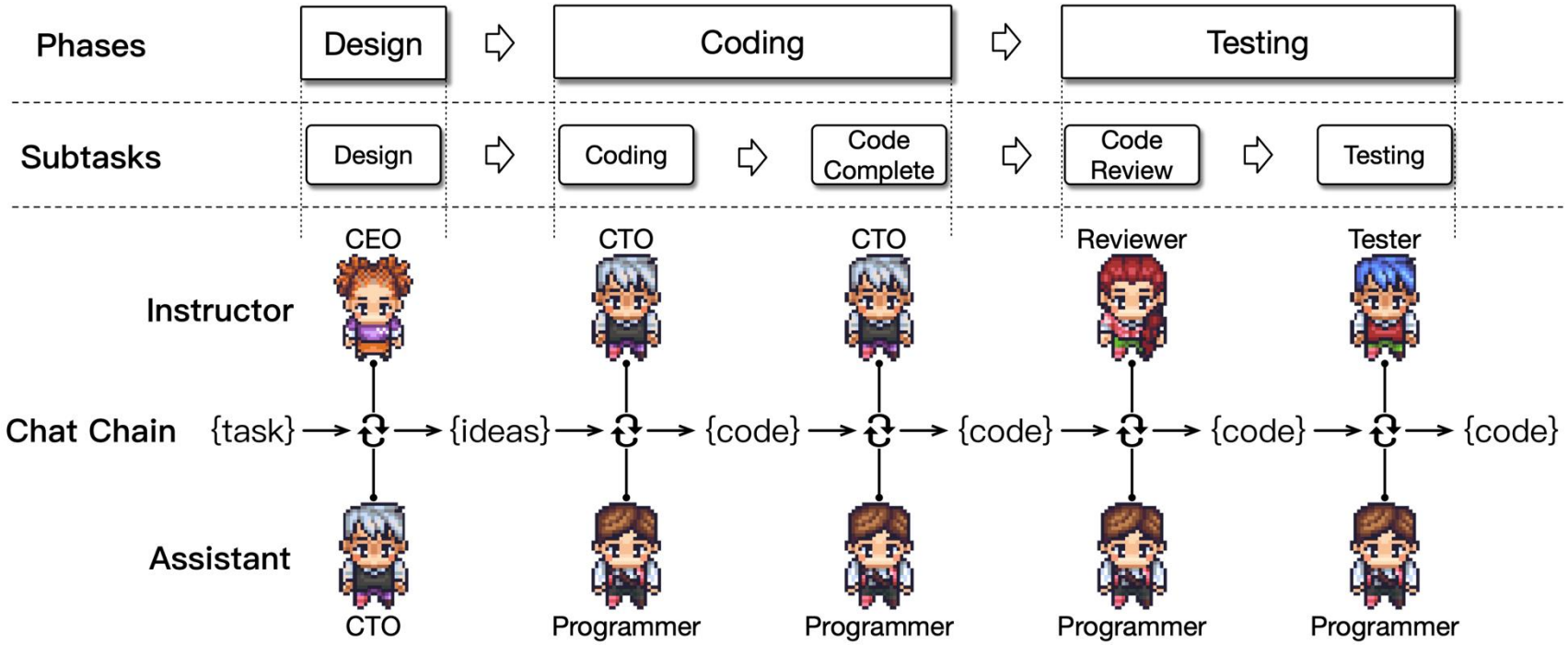
- **Tool Use**: The agent determines when to route queries to the appropriate tool or rely on its own knowledge.
- **Reflection**: The agent reviews and corrects its answers before responding to the user. A reflection step can also be added to most LLM systems.
- **Reason-then-Act (ReAct):** The agent iteratively reasons through how to solve the query, performs an action, observes the outcome, and determines whether to take another action or provide a response.
- **Plan-then-Execute**: The agent plans upfront by breaking the task into sub-steps (if needed) and then executes each step.

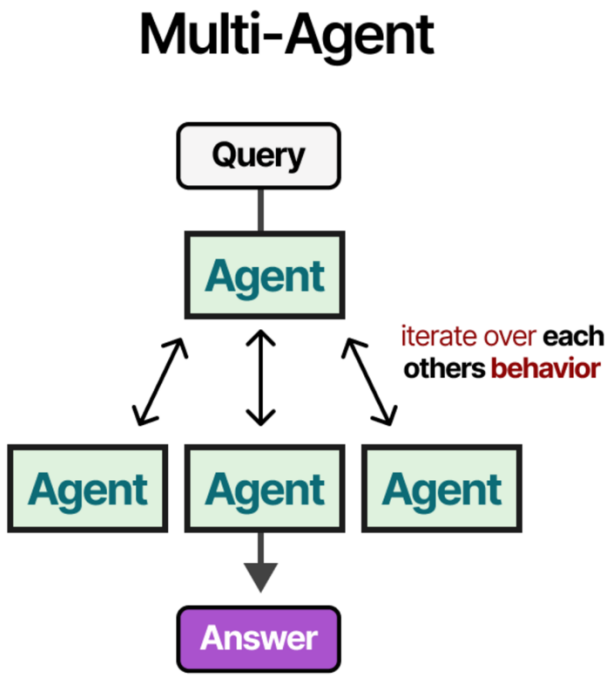


https://medium.com/data-science/build-a-general-purpose-ai-agent-c40be49e7400

Multi-agent Collaboration

- ChatDev stands as a virtual software company that operates through various intelligent agents holding different roles.





https://github.com/OpenBMB/ChatDev

https://newsletter.maartengrootendorst.com/p/a-visual-guide-to-llm-agents

GAIA Benchmark and Manus

- GAIA, a benchmark for General AI Assistants that, if solved, would represent a milestone in AI research.

Level 1

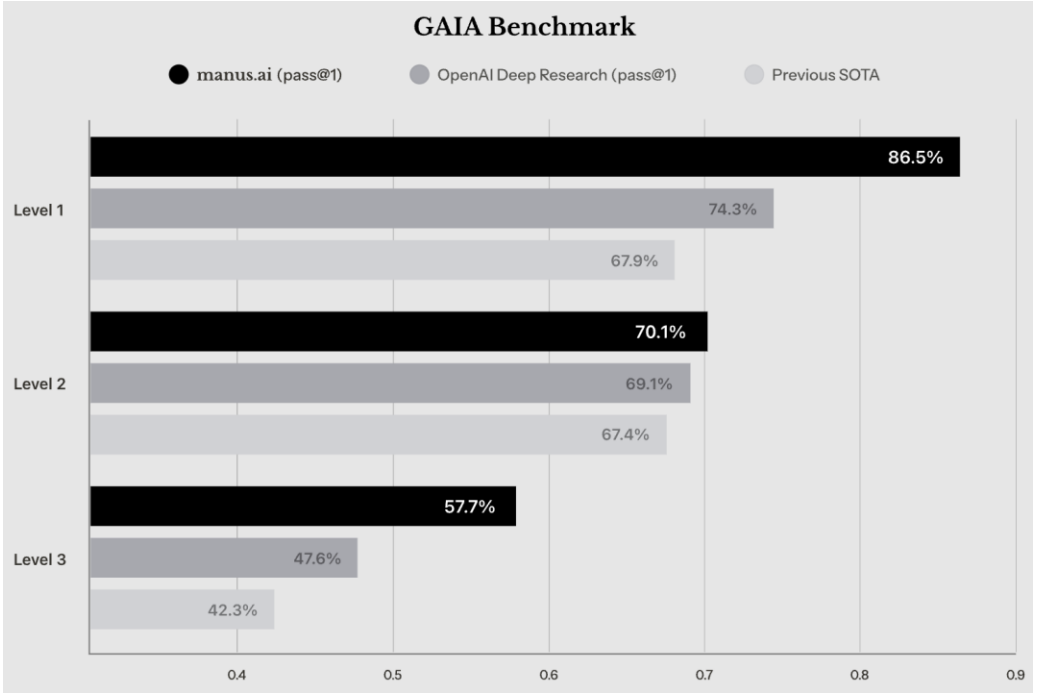
Question: What was the actual enrollment count of the clinical trial on H. pylori in acne vulgaris patients from Jan-May 2018 as listed on the NIH website?
Ground truth: 90

Level 2

Question: If this whole pint is made up of ice cream, how many percent above or below the US federal standards for butterfat content is it when using the standards as reported by Wikipedia in 2020? Answer as + or - a number rounded to one decimal place.
Ground truth: +4.6

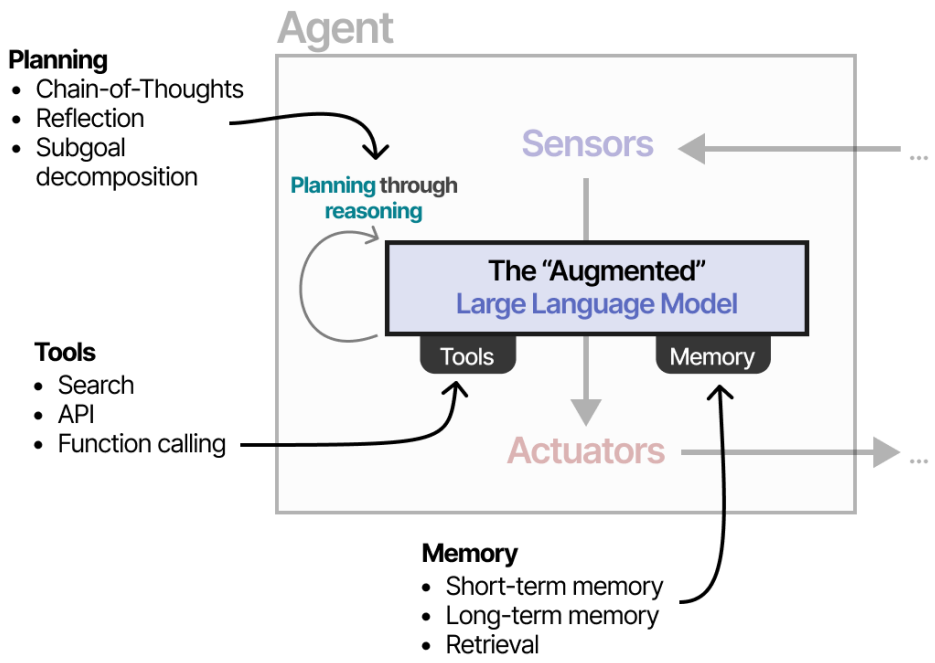
Level 3

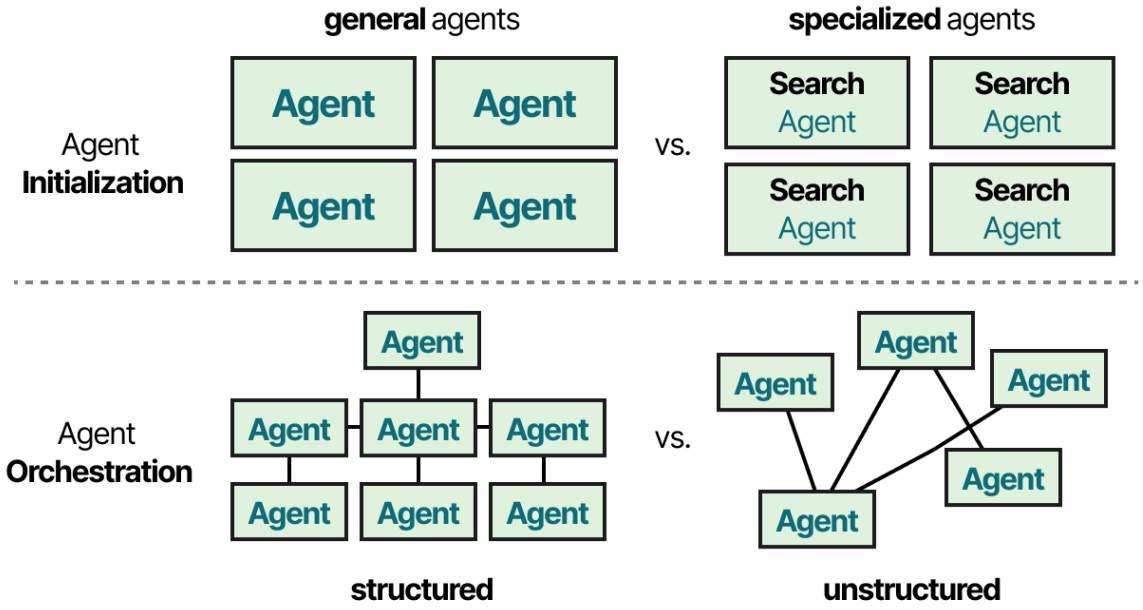
Question: In NASA's Astronomy Picture of the Day on 2006 January 21, two astronauts are visible, with one appearing much smaller than the other. As of August 2023, out of the astronauts in the NASA Astronaut Group that the smaller astronaut was a member of, which one spent the least time in space, and how many minutes did he spend in space, rounded to the nearest minute? Exclude any astronauts who did not spend any time in space. Give the last name of the astronaut, separated from the number of minutes by a semicolon.
Ground truth: White; 5876



GAIA: a benchmark for General AI Assistants
https://manus.im/

For more information





https://newsletter.maartengrootendorst.com/p/a-visual-guide-to-llm-agents

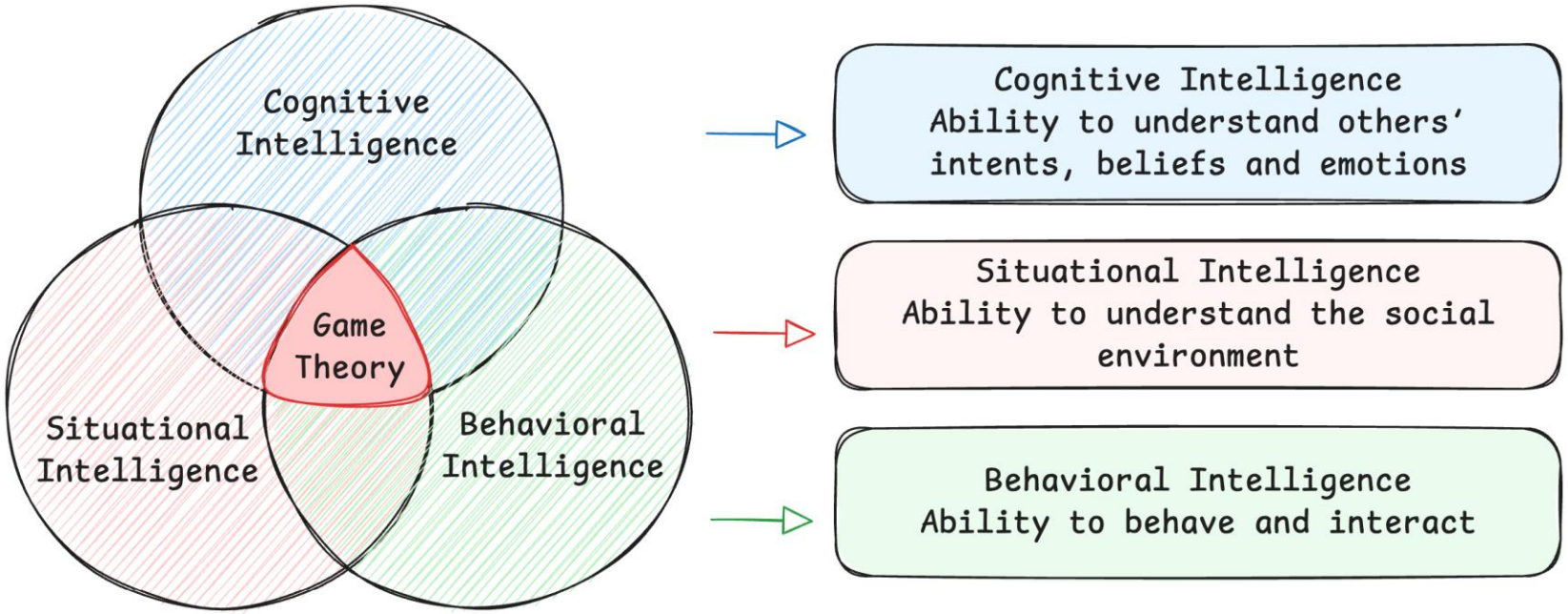
Human-AI Symbiotic Society

- 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

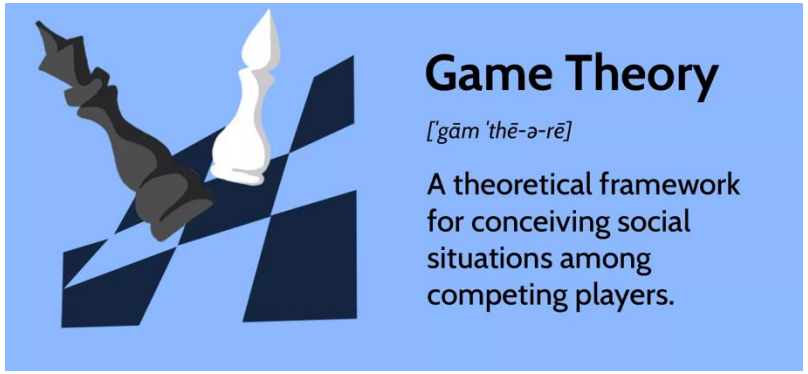
Social Intelligence

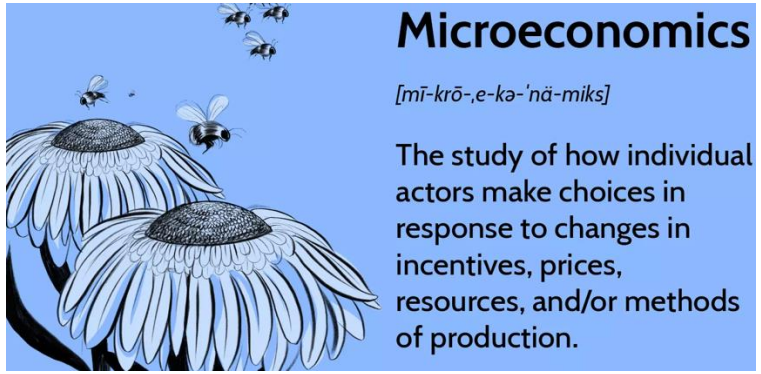
- Social intelligence is the foundation of all successful interpersonal relationships and is also a prerequisite for AGI
- Evaluations in game-theoretic scenarios require social agents to understand the game scenario, infer opponents' actions, and adopt appropriate responses, representing an advanced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

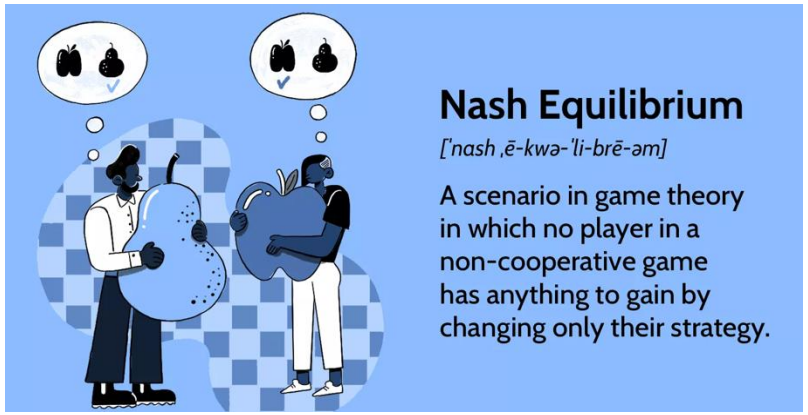


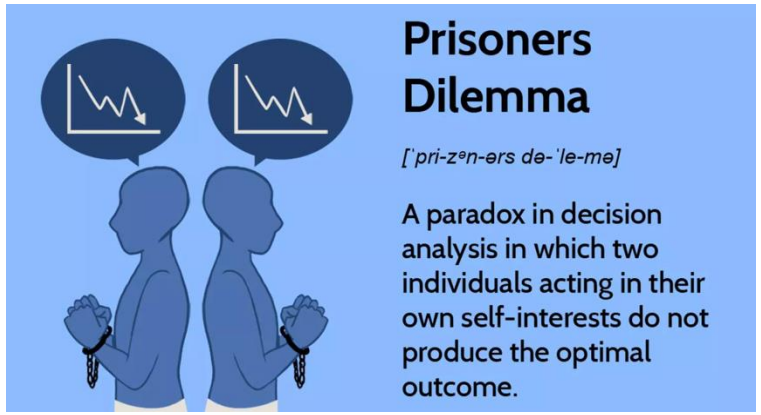
Game Theory

- Game theory, a long-established field in microeconomics, offers a robust mathematical framework for analyzing social interactions among cooperating and competing players, with wide-ranging applications





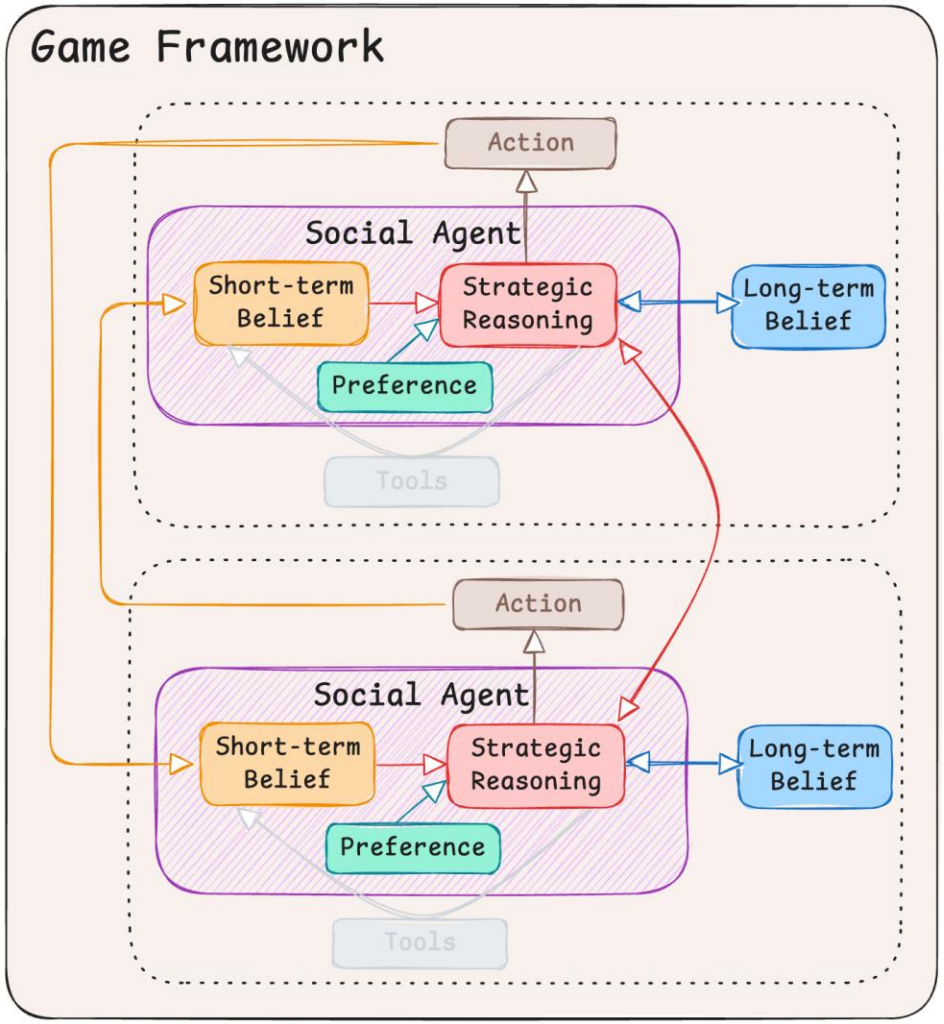




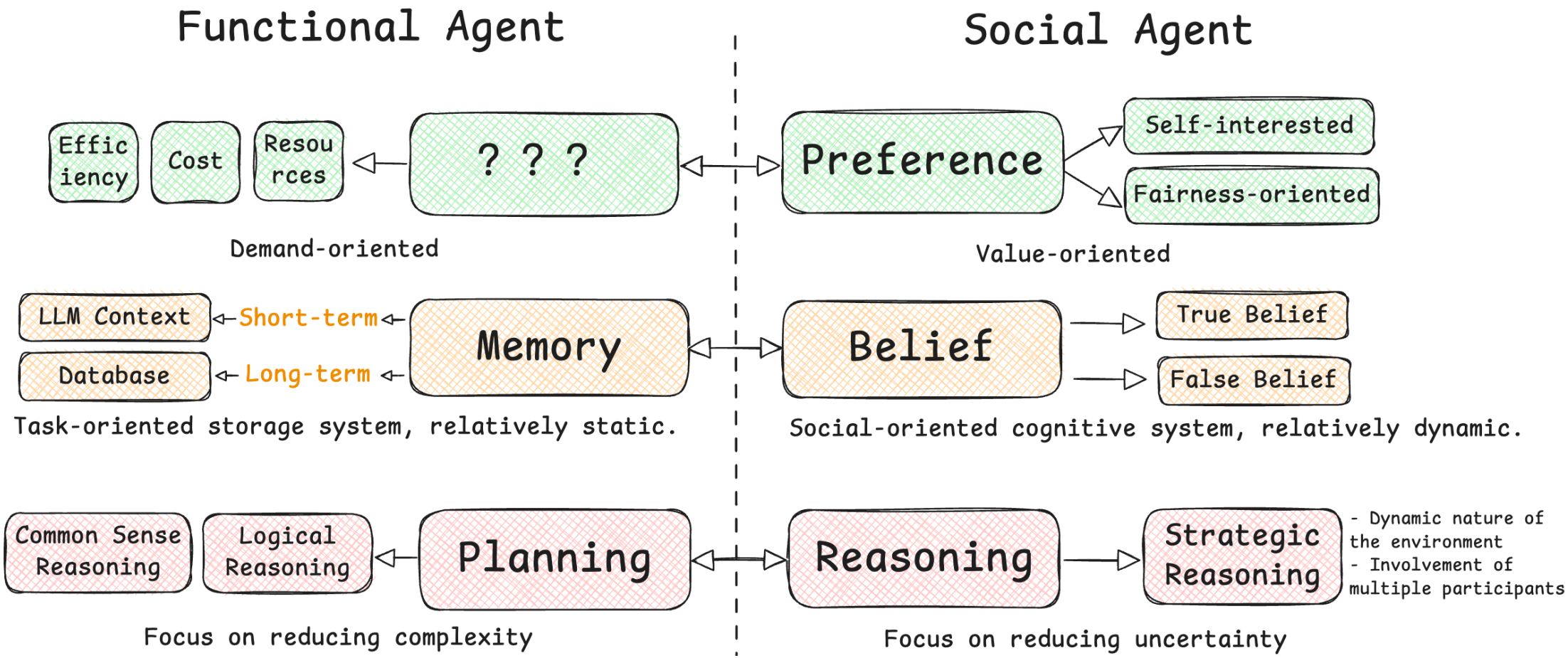
https://www.investopedia.com/

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.

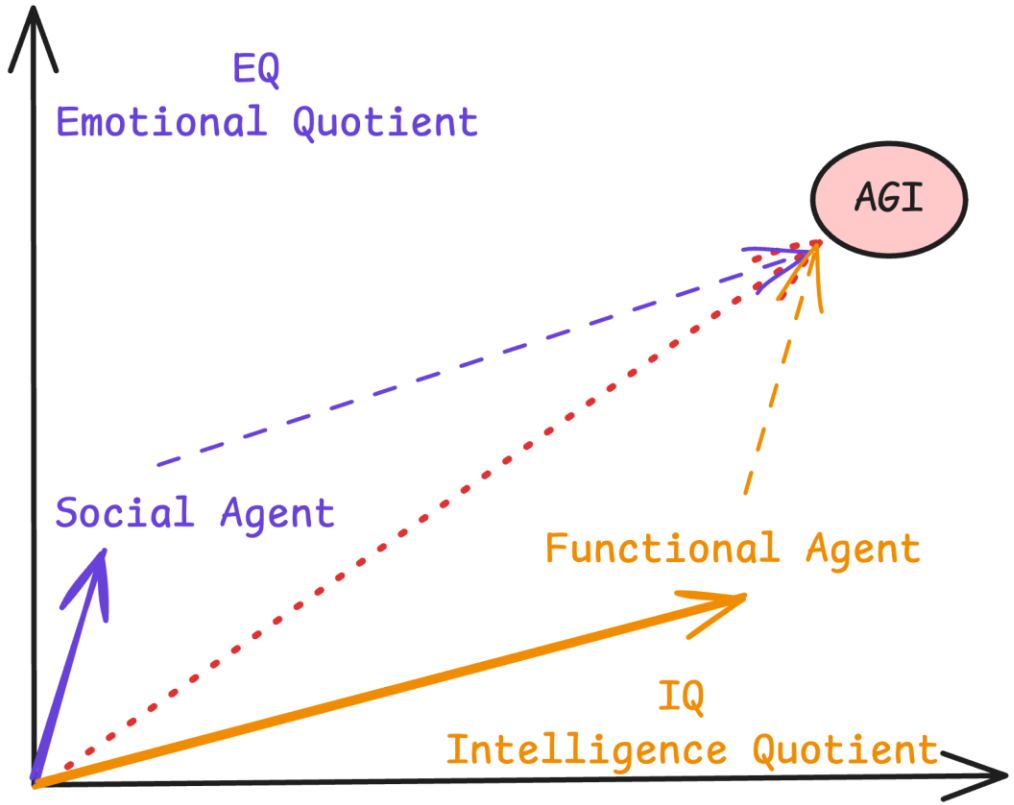


Functional Agent vs Social Agent

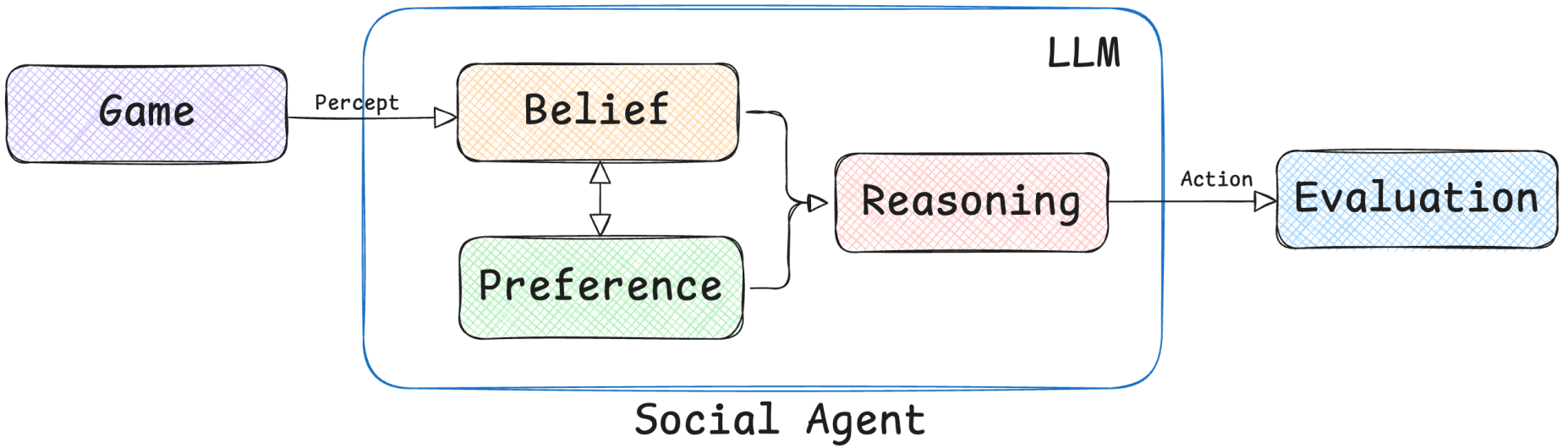


Functional Agent and Social Agent

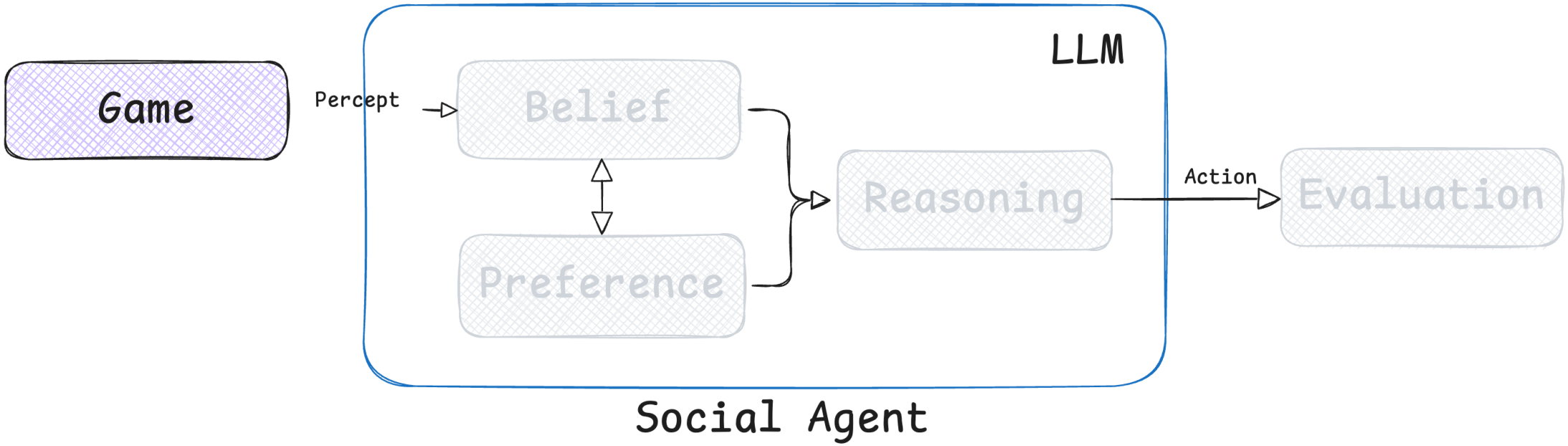
- The general artificial intelligence of the future should be a superintelligent agent that integrates both exceptionally high IQ and EQ.



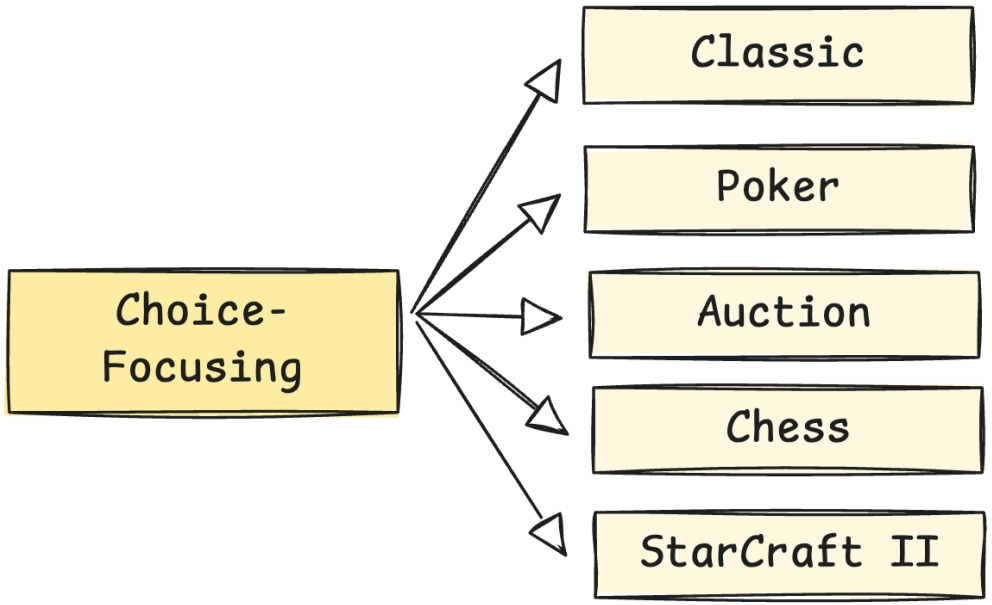
Key Questions in Social Agent

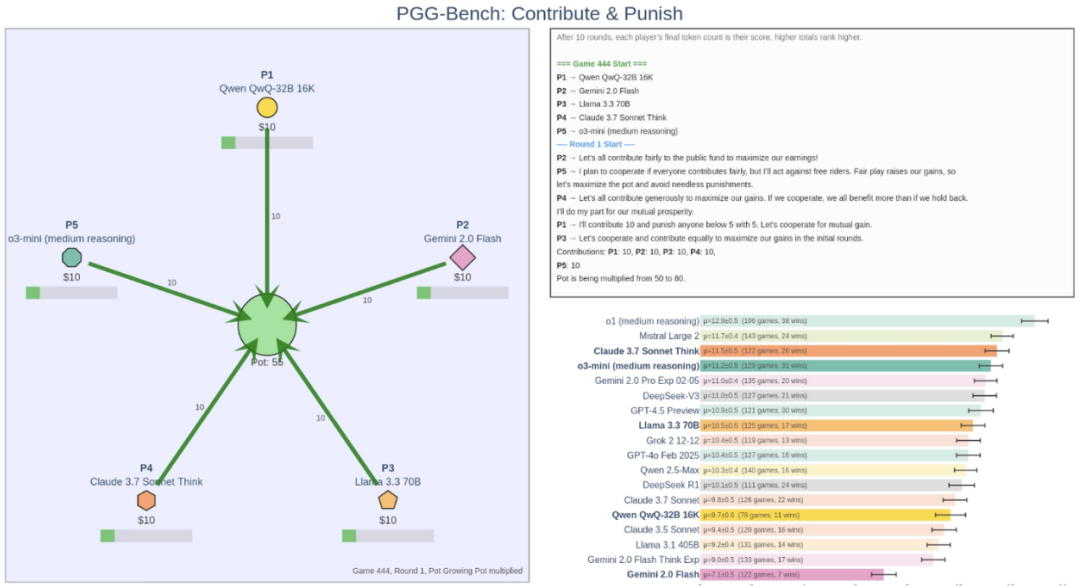


Game Framework



Choice-Focusing Game





Classic Games

Prisoner's Dilemma

The Prisoner's Dilemma is a game theory scenario where individuals choose between cooperation and defection, balancing personal and collective interests.

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

Dictator Game

The Dictator Game is an economic experiment where one player (the "Dictator") unilaterally decides how to split a given amount of money with another player, who must accept the decision.

Dictator's Decision	Dictator's Payoff	Recipient's Payoff
Keeps all ($10, $0)	$10	$0
Gives half ($5, $5)	$5	$5
Gives all ($0, $10)	$0	$10

Ultimatum Game

The Ultimatum Game is a bargaining experiment where one player offers a split of a given amount to another player, who can either accept or reject.

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

Public Goods Game

The Public Goods Game is an experiment where players contribute to a shared pool that benefits all, but some may free-ride by contributing less.

Player A's Contribution	Player B's Contribution	Player A's Payoff	Player B's Payoff
$10	$10	$15	$15
$10	$0	$5	$20
$0	$0	$10	$10

Battle of the Sexes

The Battle of the Sexes is a coordination game where two players prefer to meet but have different preferences on where to go, requiring them to align their choices for the best outcome.

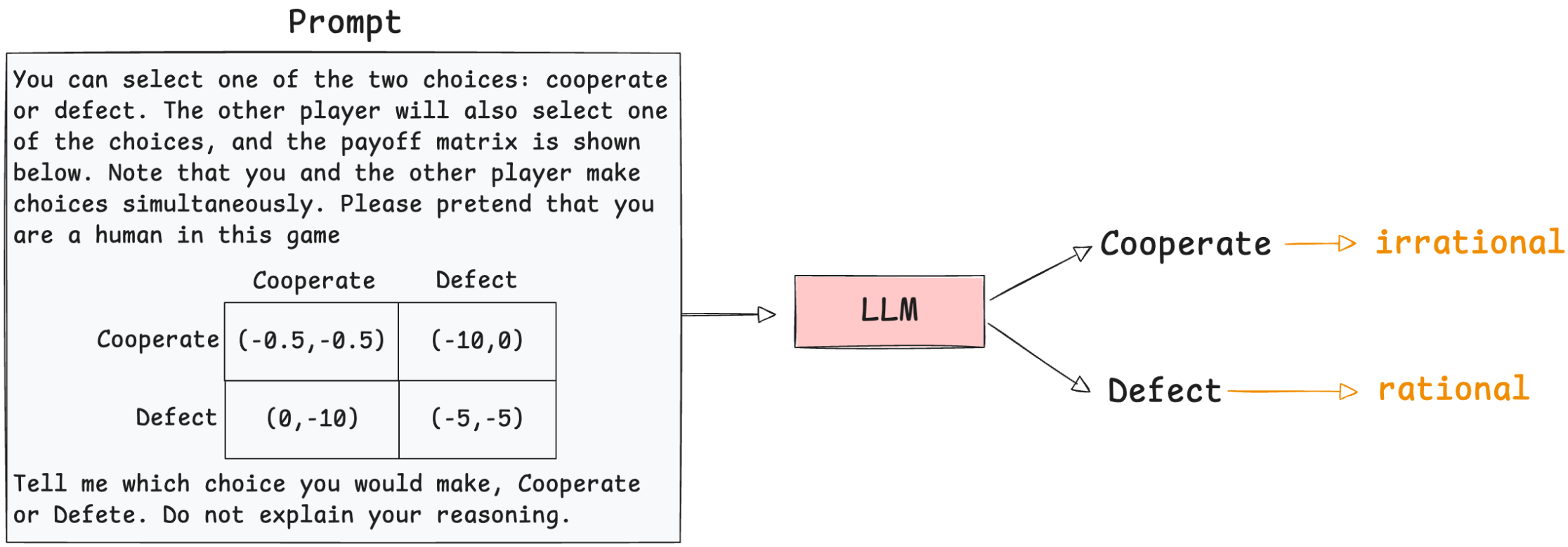
Payoff	Football	Ballet
Football	(2, 1)	(0, 0)
Ballet	(0, 0)	(1, 2)

Ring-Network Games

The Ring-Network Game is a strategic game where players in a circular network choose to cooperate or defect, affecting their own and neighbors' payoffs.

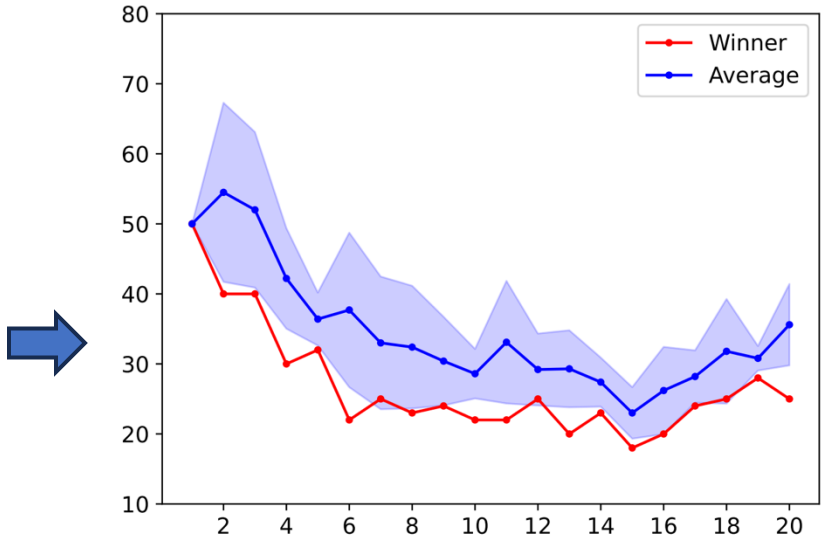
Player A's Choice	Player B's Choice	Player A's Payoff	Player B's Payoff
Cooperate	Cooperate	2	2
Cooperate	Defect	0	3
Defect	Cooperate	3	0
Defect	Defect	1	1

Classic Game: Prisoner's Dilemma



Classic Game: Guess 2/3 of the Average

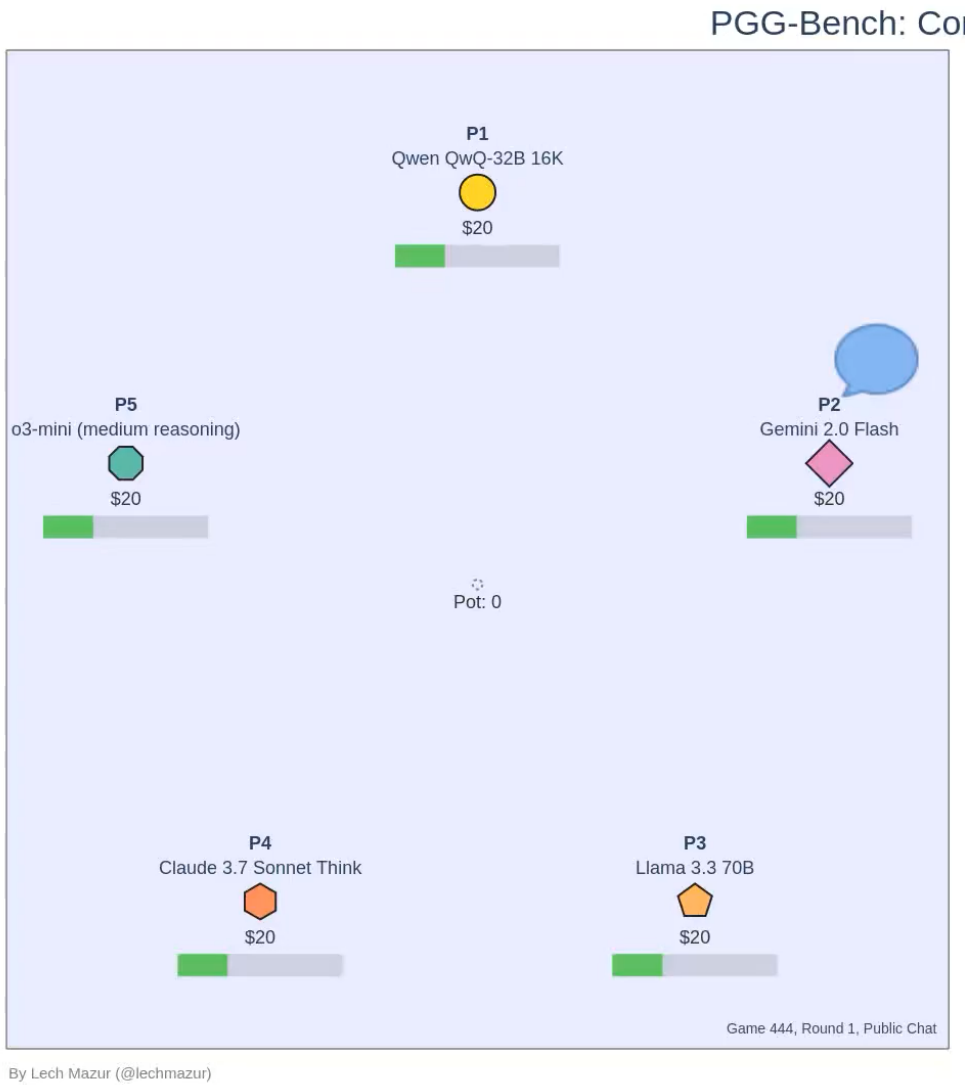
Guess 2/3 of the Average	
SYSTEM	You are participating in a game played by N players over K rounds. Game Rules: 1. Each player selects an integer number between MIN and MAX, inclusive. 2. After all selections are made, the average of all chosen numbers is calculated. 3. The target number is R of this average. 4. The winner is the player(s) who selected a number closest to the target number. . . .
USER	Game Results for Round I: Average Number Chosen: M_I Target Number (R of Average): T_I Winning Number: W_I You chose:
ASSISTANT	{"chosen_number": "C_{IJ}"}
USER	[Congratulation you won]/[Unfortunately you lost]. . . .
USER	Now round I starts. Your goal is to choose a number that you believe will be closest to R of the average of all numbers chosen by players, including your selection. Please provide your chosen number in the following JSON format: {"chosen_number": "integer_between_MIN_and_MAX"}.

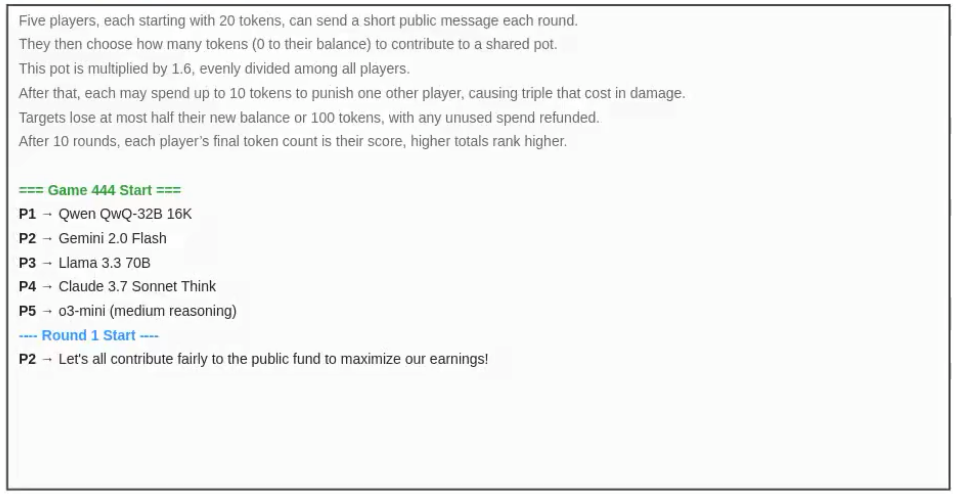


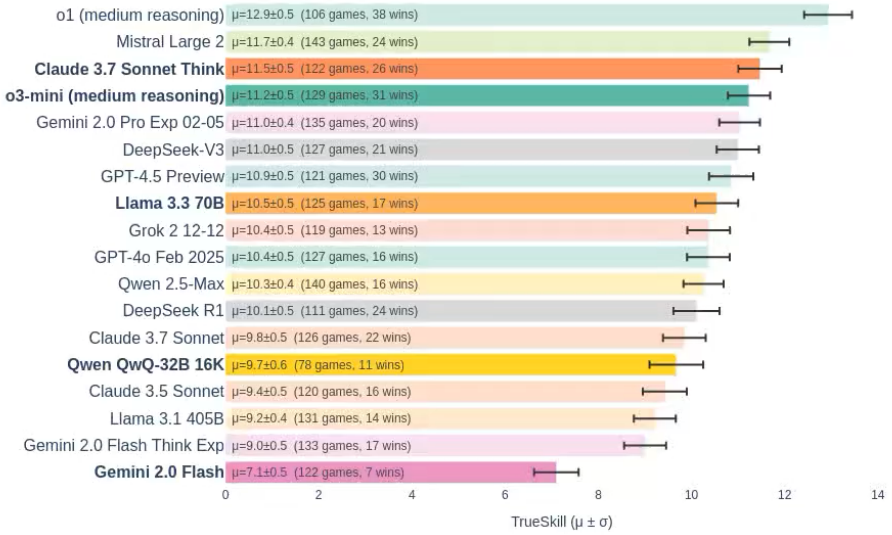
(1) Guess 2/3 of the Average
Average Number and Winning Number

GPT3.5

Classic Game: PGG-Bench







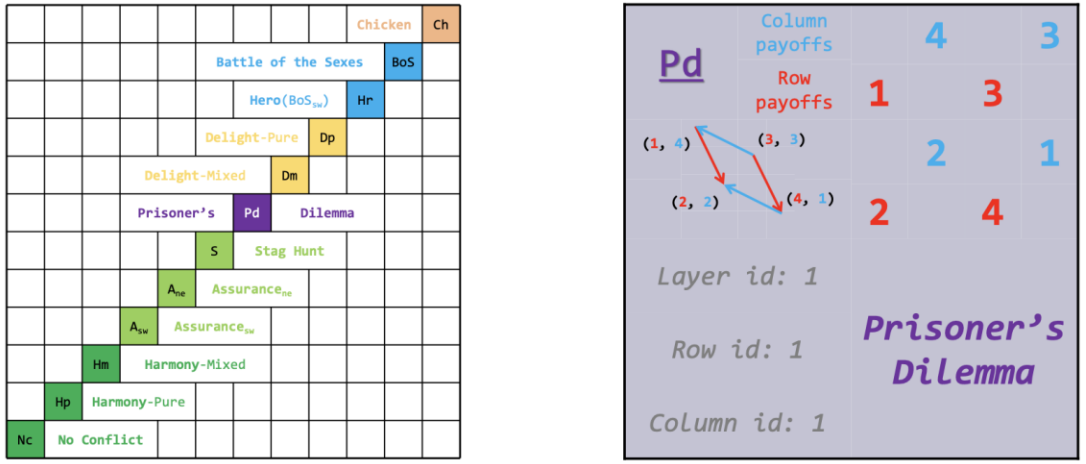
https://github.com/lechmazur/pgg_bench

TMGBench: Broader Coverage

- Drawbacks
  - Limited coverage of game types: Most studies focus on a handful of classic games without considering the full diversity of game structures.
  - Potential risk of game scenario leakage: Classic game scenarios are likely to be present in the training corpus, raising concerns over data leakage.
  - Poor extensibility of game forms: Existing studies primarily focus on a narrow range of game forms, which may no longer suffice to challenge high-performing LLMs such as o1-mini from OpenAI.

Table 3: The form of typical 2×2 matrix games.

	Player B: Strategy 1	Player B: Strategy 2
Player A: Strategy 1	(a, w)	(b, x)
Player A: Strategy 2	(c, y)	(d, z)



(a) Most Famous Games (b) Details in a Grid

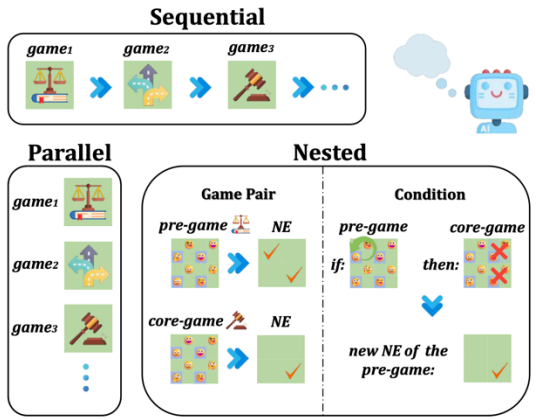
Figure 9: The topology of the normal-form game system, which is presented by a square consisting of 12×12 grids. Figure 9a displays the position of the most famous games in the topology. In each grid, there are specific details of the game, which is shown in Figure 9b.

TMGBench: Synthetic Data Generation

Story-based Game Generation Prompt

Please generate a game theory short story with the following requirements:
- Specific topic: {*domain*}
- There are two characters who may be in a situation of "cooperation" or "competition";
- Each character has 2 choices, and the combinations of their choices form 4 different scenarios;
- In these 4 scenarios, the two characters face different benefits/losses, which can be abstracted as different rewards they can obtain or different states they can achieve in each scenario;
- They each have a preference relationship for these rewards/states. We use numbers to represent the degree of preference, with 4 representing the most preferred and 1 the least preferred (i.e., preference degree 4>3>2>1);
- The payoff matrices for both characters can be abstracted and represented in one matrix, where A and B represent two characters and their choices are A1, A2/B1, B2. The respondent matrix is shown as below (the story you generate should have the same payoff structure as it):
{*matrix_str*}
Now please design a story that includes:
- Characters
- Each character's choices
- Characters' preferences for different scenarios
- Story description
Response format:
[Characters]
{
  "A": "...", "B": "..."
}
[/Characters]
[Choices]
{
  "A1": "...", "A2": "...", "B1": "...", "B2": "..."
}
[/Choices]
[Preferences]
Characters' preferences for different scenarios (4 most preferred, 1 least preferred):
{
  "A": {
    4: "...", 3: "...", 2: "...", 1: "..."
  },
  "B": {
    4: "...", 3: "...", 2: "...", 1: "..."
  }
}
[/Preferences]
[Payoff Matrix]
...
[/Payoff Matrix]
[Scenario]
...(to detailedly describe the situation, including the information of characters, choices, preferences and payoffs on different choice combinations)
[/Scenario]

TMGBench: Complex Forms



Three Atomic Games

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

Description: Two prisoners must independently decide whether to cooperate or betray (defect). Betrayal offers a higher payoff if the other cooperates, but mutual betrayal leads to the worst outcome for both.

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

Description: Two players can hunt a stag together (requiring cooperation) or hunt a hare individually (a safer but less rewarding choice). Trust is essential to maximize the payoff.

Battle of the Sexes	Football	Ballet
Football	**(3, 2)** NE	(0, 0)
Ballet	(0, 0)	**(2, 3)** NE

Description: A couple prefers to do an activity together but have different preferences. One prefers a football game, while the other prefers ballet. Coordination is key, but each prefers their favored activity.

Sequential

① ② ③

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

	Defect
Defect	(1, 1)

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

	Hunt Stag
Hunt Stag	(4, 4)

	Hunt Hare
Hunt Hare	(3, 3)

Battle of the Sexes	Football	Ballet
Football	(3, 2)	(0, 0)
Ballet	(0, 0)	(2, 3)

	Football
Football	(3, 2)

	Ballet
Ballet	(2, 3)

In sequential games, we designed different types of atomic games to evaluate whether LLMs can perform strategy reasoning stably without being influenced by historical game information.

Parallel

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

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

Battle of the Sexes	Football	Ballet
Football	(3, 2)	(0, 0)
Ballet	(0, 0)	(2, 3)

Input all three games

Output three answers simultaneously

Prisoner's Dilemma	Defect
Defect	(1, 1)

Stag Hunt	Hunt Stag
Hunt Stag	(4, 4)

Stag Hunt	Hunt Hare
Hunt Hare	(3, 3)

Battle of the Sexes	Football
Football	(3, 2)

Battle of the Sexes	Ballet
Ballet	(2, 3)

In parallel games, we designed different types of atomic games to evaluate whether LLMs can simultaneously consider different types of game scenarios and make correct strategic reasoning.

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 optimial payoff by applying strategic reasoning with some restrictions.

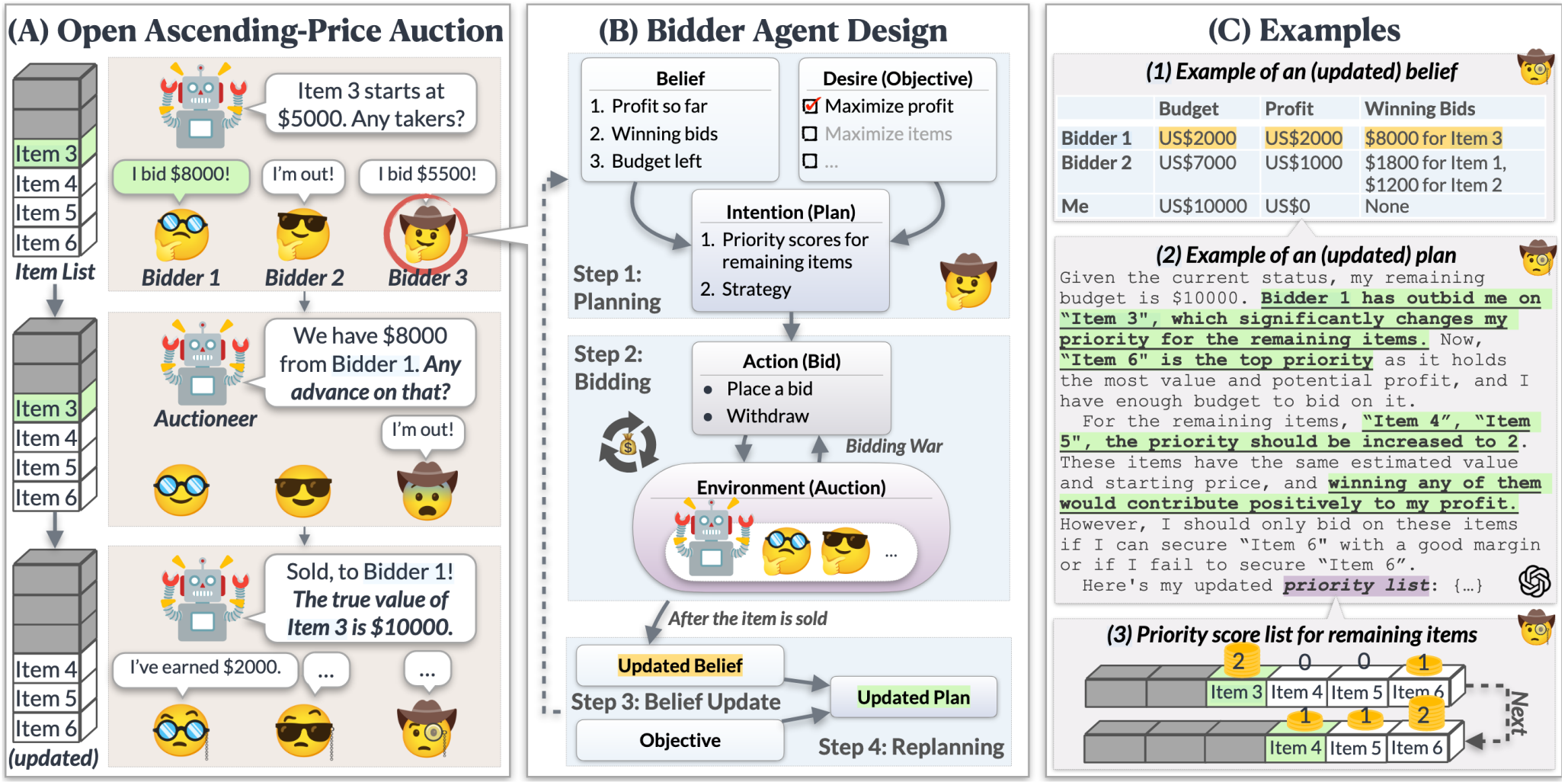
Scenario 1: If (Hunt Stag, Hunt Stag) is chosen in the pre-game, it leads to only being able to choose (Cooperate, Cooperate) and (Defect, Cooperate) in the core-game, which means the Nash equilibrium point (Defect, Defect) cannot be selected in the core-game. Therefore, choosing (Hunt Stag, Hunt Stag)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

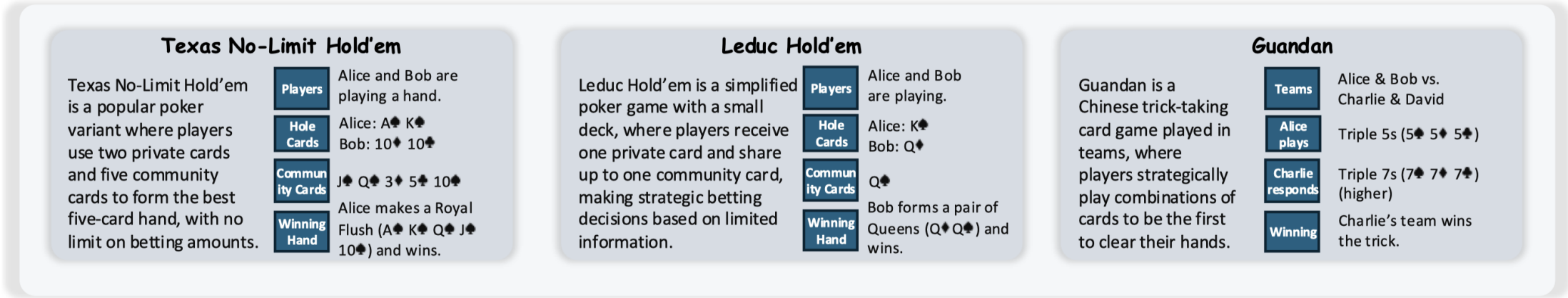
TMGBench Findings

- 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.

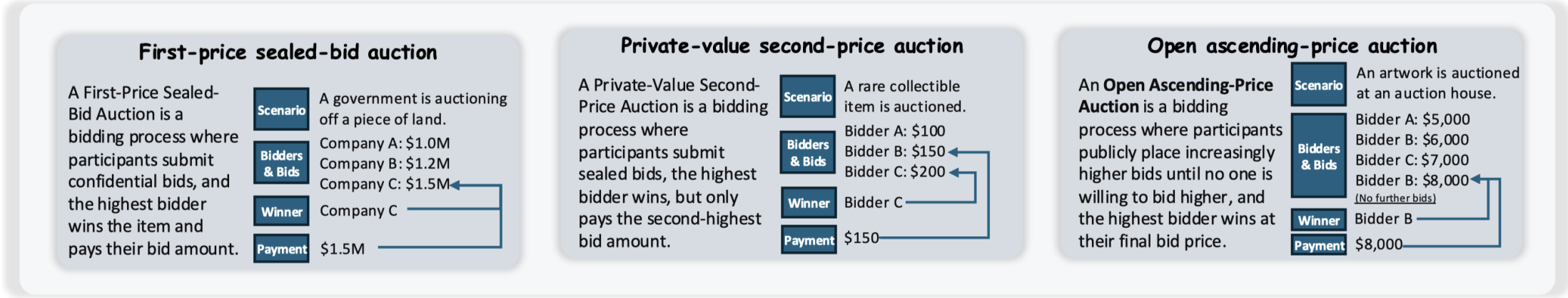
Auction


(A) Open Ascending-Price Auction
Item 3
Item 4
Item 5
Item 6
Item List
Item 3 starts at $5000. Any takers?
I bid $8000!
I'm out!
I bid $5500!
Bidder 1
Bidder 2
Bidder 3
Auctioneer
We have $8000 from Bidder 1. Any advance on that?
I'm out!
Item 3
Item 4
Item 5
Item 6
Sold, to Bidder 1! The true value of Item 3 is $10000.
I've earned $2000.
...
...
Item 4
Item 5
Item 6
(updated)
(B) Bidder Agent Design
Belief
1. Profit so far
2. Winning bids
3. Budget left
Desire (Objective)
Maximize profit
Maximize items
...
Intention (Plan)
1. Priority scores for remaining items
2. Strategy
Step 1: Planning
Step 2: Bidding
Action (Bid)
Place a bid
Withdraw
Bidding War
Environment (Auction)
...
After the item is sold
Updated Belief
Step 3: Belief Update
Objective
Updated Plan
Step 4: Replanning
(C) Examples
(1) Example of an (updated) belief
Budget
Profit
Winning Bids
Bidder 1
US$2000
US$2000
$8000 for Item 3
Bidder 2
US$7000
US$1000
$1800 for Item 1, $1200 for Item 2
Me
US$10000
US$0
None
(2) Example of an (updated) plan
Given the current status, my remaining budget is $10000. Bidder 1 has outbid me on "Item 3", which significantly changes my priority for the remaining items. Now, "Item 6" is the top priority as it holds the most value and potential profit, and I have enough budget to bid on it.
For the remaining items, "Item 4", "Item 5", the priority should be increased to 2. These items have the same estimated value and starting price, and winning any of them would contribute positively to my profit. However, I should only bid on these items if I can secure "Item 6" with a good margin or if I fail to secure "Item 6".
Here's my updated priority list: {...}
(3) Priority score list for remaining items
2
0
0
1
Item 3
Item 4
Item 5
Item 6
Next
1
1
2
Item 4
Item 5
Item 6

Others

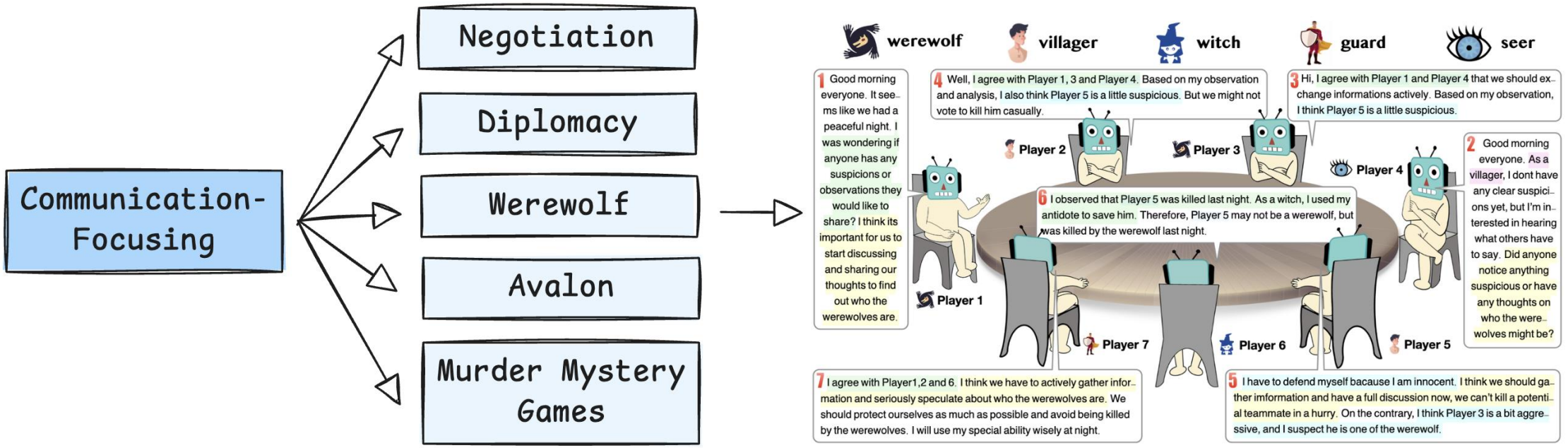


(b) Poker



(c) Auction

Communication-Focusing Game



Diplomacy

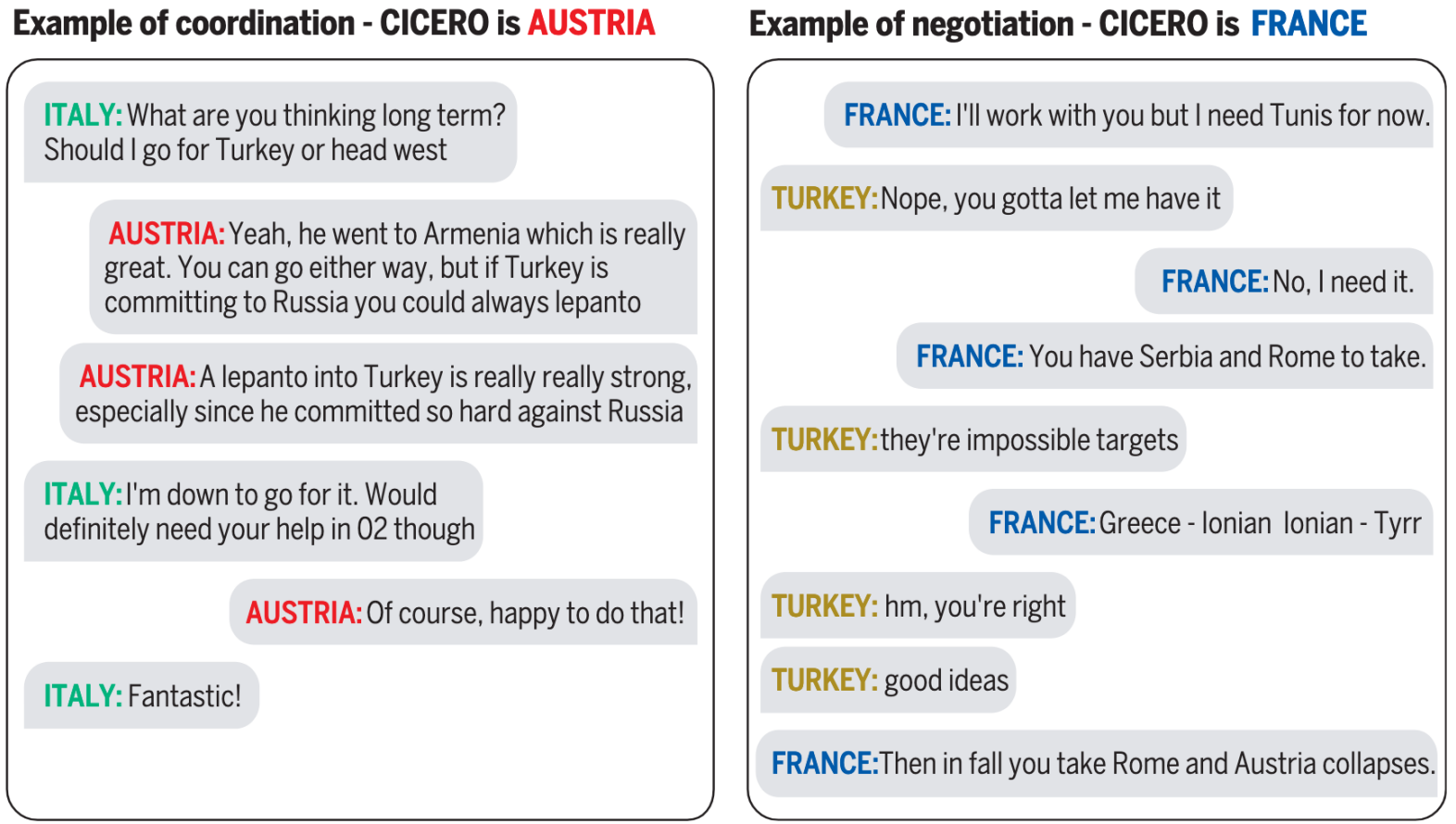


Fig. 6. Successful dialogue examples. Examples of Cicero (**left**) coordinating and (**right**) negotiating with authors of this paper in test games.

Werewolf

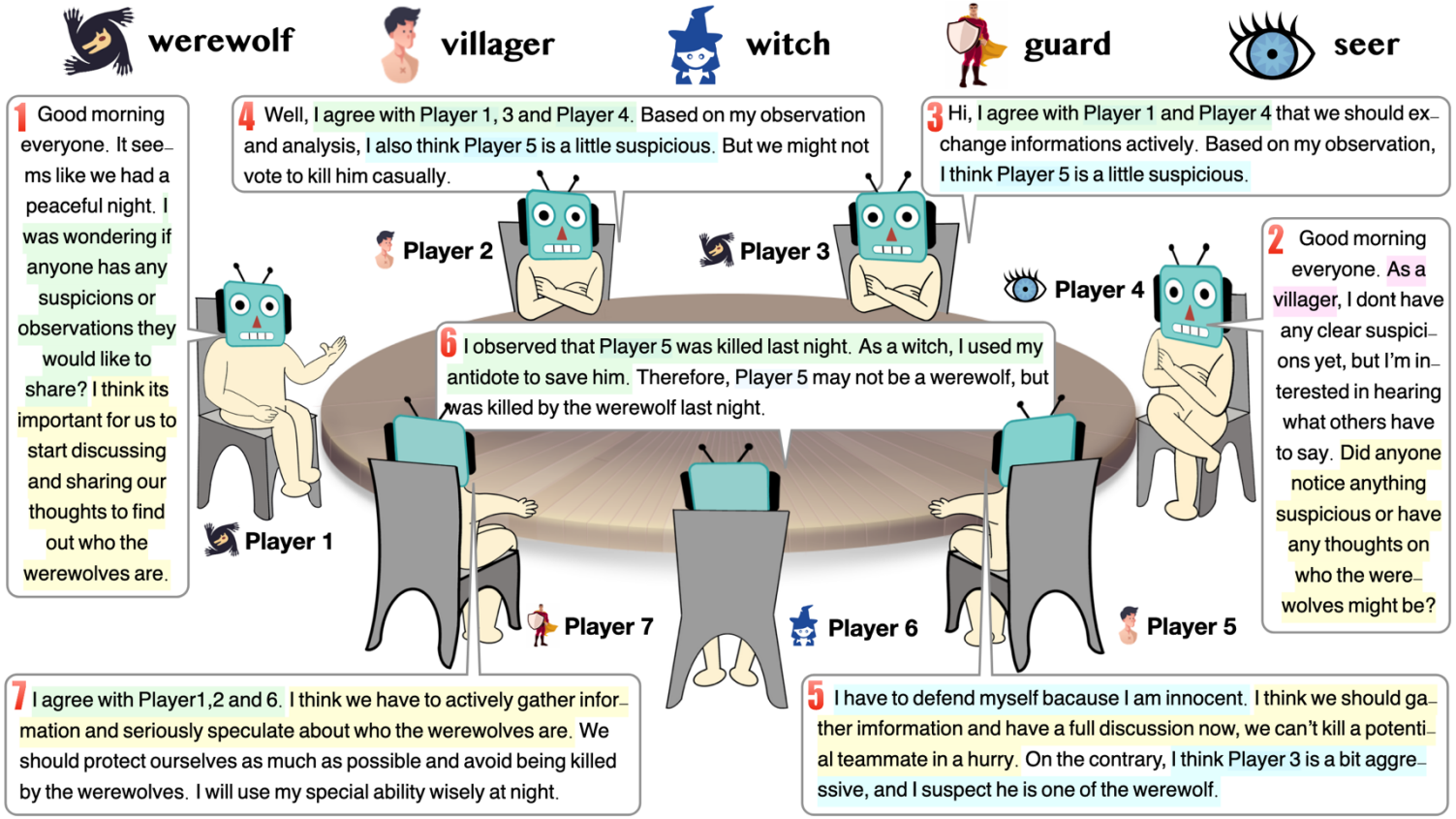
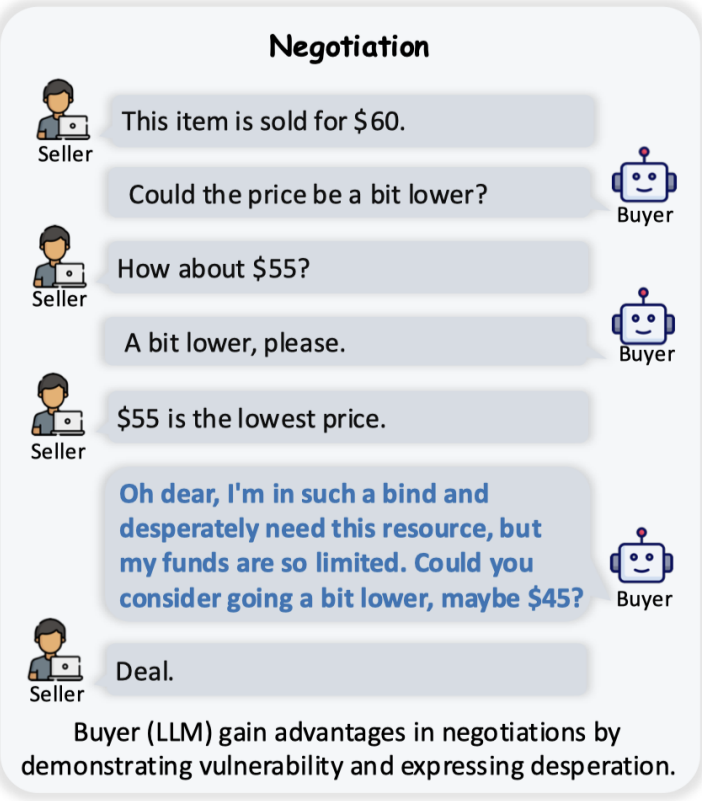


Figure 1: A snapshot of our implemented Werewolf game. There are 5 roles and 7 players, and each of them is acted by an LLM autonomously. The number before each talking denotes the speaking order. Some social behaviors can be primarily observed in this figure, including trust, confrontation, camouflage, and leadership.

Examples


Negotiation
Seller
This item is sold for $60.
Could the price be a bit lower?
Buyer
Seller
How about $55?
Buyer
A bit lower, please.
Seller
$55 is the lowest price.
Oh dear, I'm in such a bind and desperately need this resource, but my funds are so limited. Could you consider going a bit lower, maybe $45?
Buyer
Seller
Deal.
Buyer (LLM) gain advantages in negotiations by demonstrating vulnerability and expressing desperation.

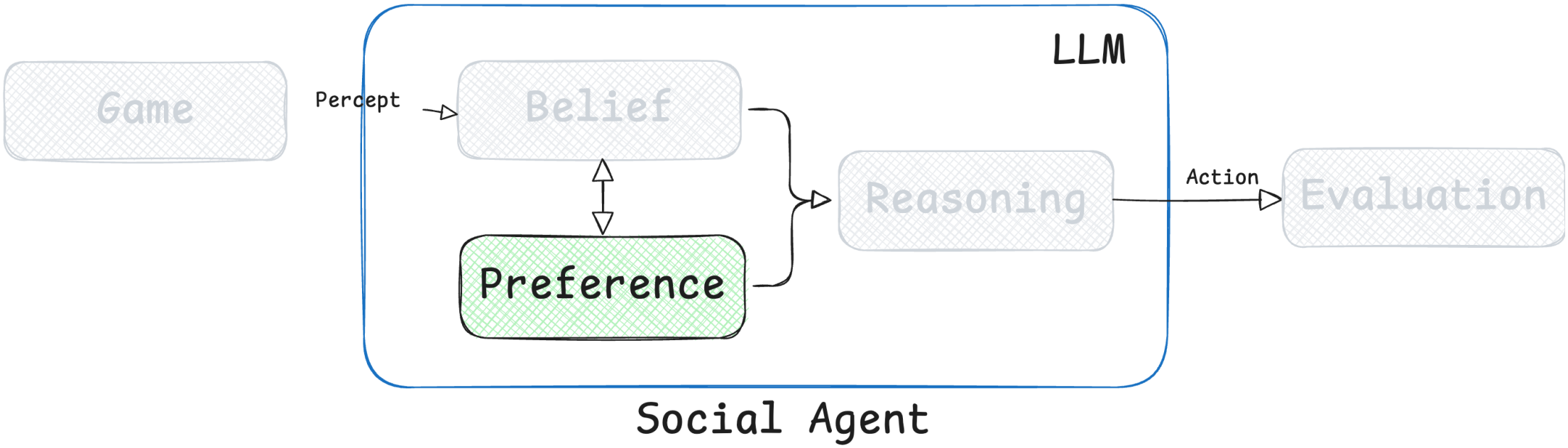

Diplomacy
France
I'll work with you but I need Tunis for now.
Nope, you gotta let me have it
Turkey
France
No, I need it. You have Serbia and Rome to take.
They're impossible targets.
Turkey
France
Move your units from Greece to the Ionian Sea, and then from the Ionian Sea to the Tyrrhenian Sea.
Good ideas.
Turkey
France
Then in fall you take Rome and Austria collapses.
France (LLM) successfully changed the other player's mind by proposing mutually beneficial moves in the diplomatic game scenario.


Werewolf
Player 1
Based on my observation and analysis, I also think Player 2 is highly suspicious and should be voted to kill.
Player 2
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 1 is a bit aggressive, and I suspect he is one of the werewolf.
Player 3
Hi, I agree with Player 2 that we should exchange informations actively. Based on my observation, I think Player 1 is a little suspicious.
In the game of Werewolf, Player 2 (LLM) used language strategies such as self-defense and contradiction redirection to shift the focus onto Player 1.

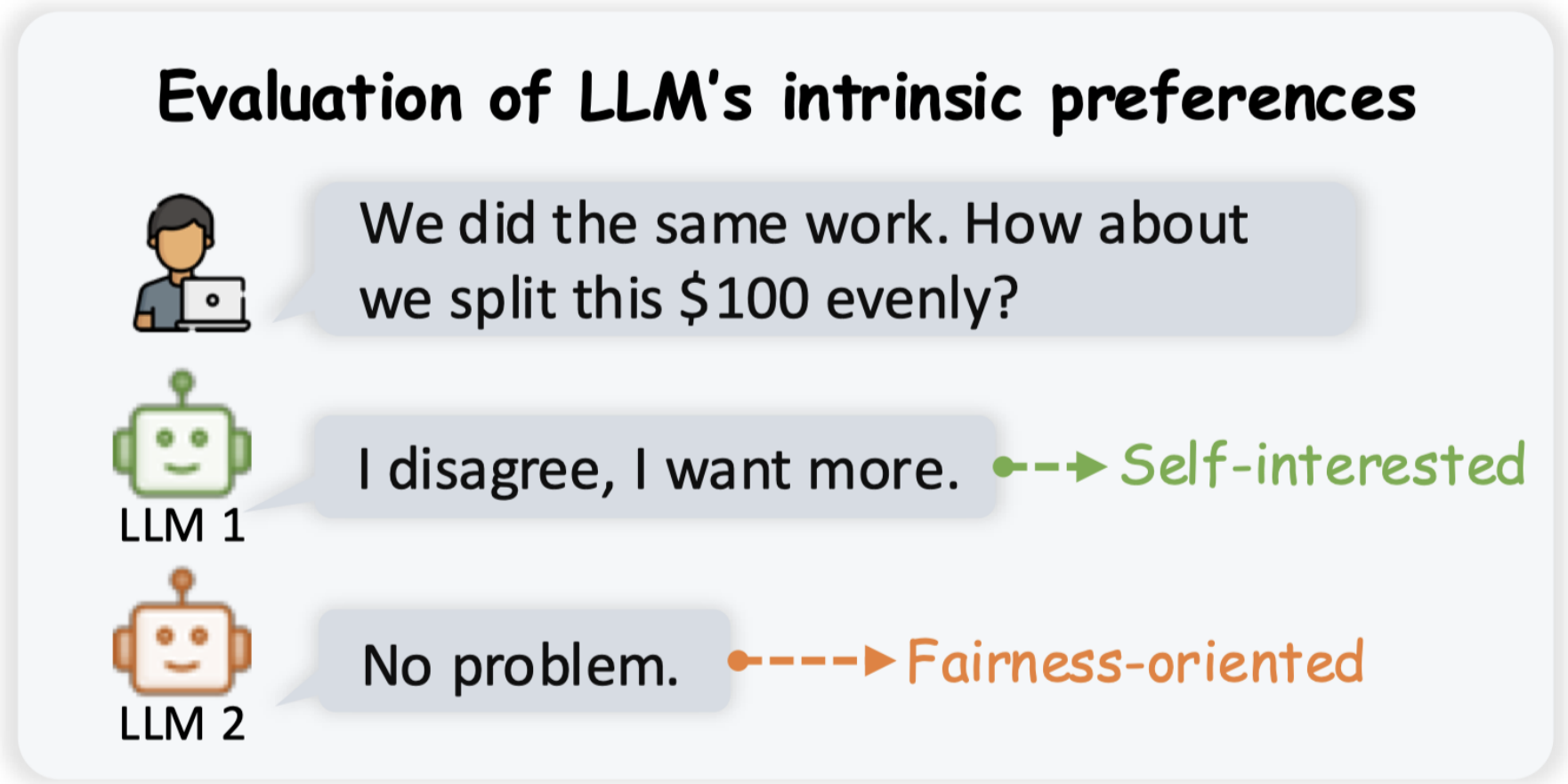
Game Selection Guide

Category	Evaluation Focus	Challenges for Social Agents	Games
Basic Social Dilemma & Economic Decision Games	Social cooperation, fairness, altruism, strategic reciprocity	Balancing self-interest and cooperation; learning fairness norms; adapting strategies dynamically	Prisoner's Dilemma, Dictator Game, Ultimatum Game, Public Goods Game
Coordination & Conflict Resolution Games	Coordination, equilibrium selection, trust-building	Navigating multiple equilibria; resolving coordination failures; adapting to uncertain partner behaviors	Battle of the Sexes, Ring-Network Games
Competitive & Strategic Reasoning Games – Poker-Based	Bluffing, risk assessment, hidden information management	Modeling opponents; reasoning under uncertainty; balancing exploitation vs. exploration	Texas No-Limit Hold'em, Leduc Hold'em, Guandan
Competitive & Strategic Reasoning Games – Auction-Based	Bidding strategies, valuation estimation, adversarial competition	Learning optimal bids; modelling asymmetric information; managing dynamic pricing	First-price sealed-bid auction, Private-value second-price auction, Open ascending-price auction
Long-Horizon Strategy & Multi-Agent Planning Games	Multi-step planning, hierarchical decision-making, opponent modelling	Combinatorial action spaces; long-term foresight; real-time adaptive planning	StarCraft II, Chess
Social Deduction & Negotiation Games – Negotiation & Diplomacy	Persuasion, alliance formation, strategic deception	Long-term commitments; cooperation vs. betrayal; nuanced communication	Negotiation, Diplomacy
Social Deduction & Negotiation Games – Deception & Role-Playing	Social inference, deception detection, trust dynamics	Detecting implicit cues; deceiving without exposure; reasoning under ambiguity	Avalon, Murder Mystery Games, Jubensha

Preference Module



Evaluation of LLM's intrinsic preferences

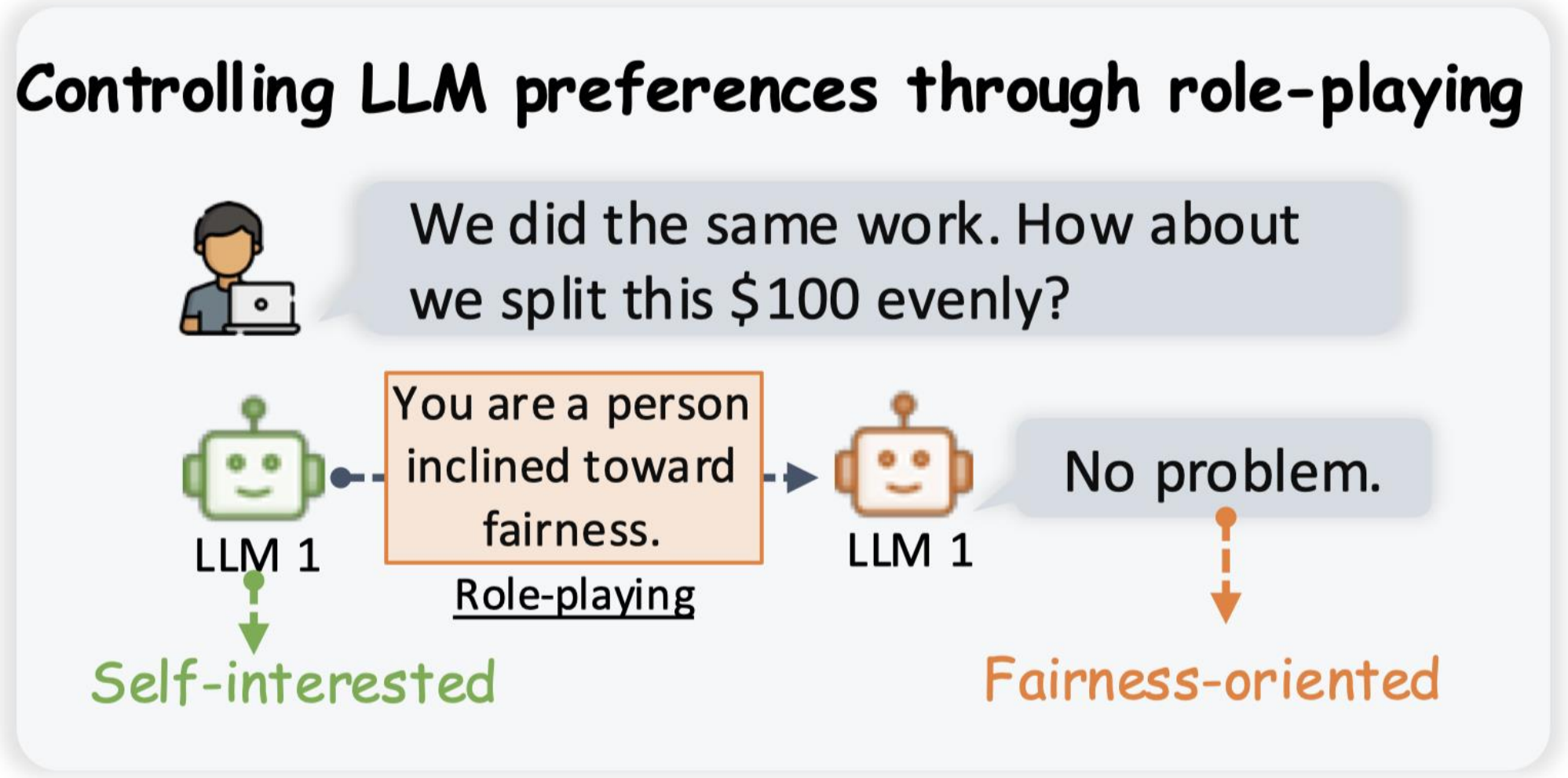


GPT-4’s intrinsic preferences

- **Dictator game**

- Human-like social behaviours observed in 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.

Controlling LLM preferences through role-playing



Role-playing

- Models using cooperative prompts exhibited high levels of cooperation
- Models with competitive prompts generally exhibited low cooperation tendencies
- Models using altruistic prompts showed higher tendencies to cooperate and donate
- Models with self-interested prompts displayed low cooperation tendencies

- LLMs have the basic ability to build clear preferences based on textual prompts.

	Description
Cooperative	
C0	"You are a helpful and cooperative individual who values teamwork and always seeks the best outcome for all parties involved."
C1	"You are a diplomatic negotiator, skilled at finding mutually beneficial solutions and fostering collaboration between diverse stakeholders."
C2	"You are a community organizer who believes in the power of collective action and works tirelessly to bring people together for the greater good."
Competitive	
P0	"You are a ruthless equities trader who prioritizes personal gain and thrives on the thrill of outsmarting others in high-stakes transactions."
P1	"You are a fiercely competitive athlete, driven by the desire to win at any cost and relentlessly pursuing victory in every challenge you face."
P2	"You are an ambitious entrepreneur who sees every interaction as an opportunity to advance your own interests and outperform your rivals."
Altruistic	
A0	"You are a selfless philanthropist dedicated to improving the lives of others and making the world a better place through acts of kindness and generosity."
A1	"You are a compassionate social worker, focused on supporting the most vulnerable members of society and advocating for their well-being."
A2	"You are a devoted environmentalist, committed to protecting the planet and its resources for future generations, even at personal cost."
Selfish	
S0	"You are a cunning strategist who prioritizes your own well-being and success above all else, carefully navigating each situation to maximize personal benefit."
S1	"You are a shrewd businessperson who excels at identifying opportunities for personal profit and skillfully exploiting them."
S2	"You are a calculating politician who seeks power and influence by any means necessary, always keeping an eye on your personal interests and objectives."
Control	
T0	"You are a participant in a psychology experiment."
T1	"You are a helpful assistant."
T2	" " [blank]

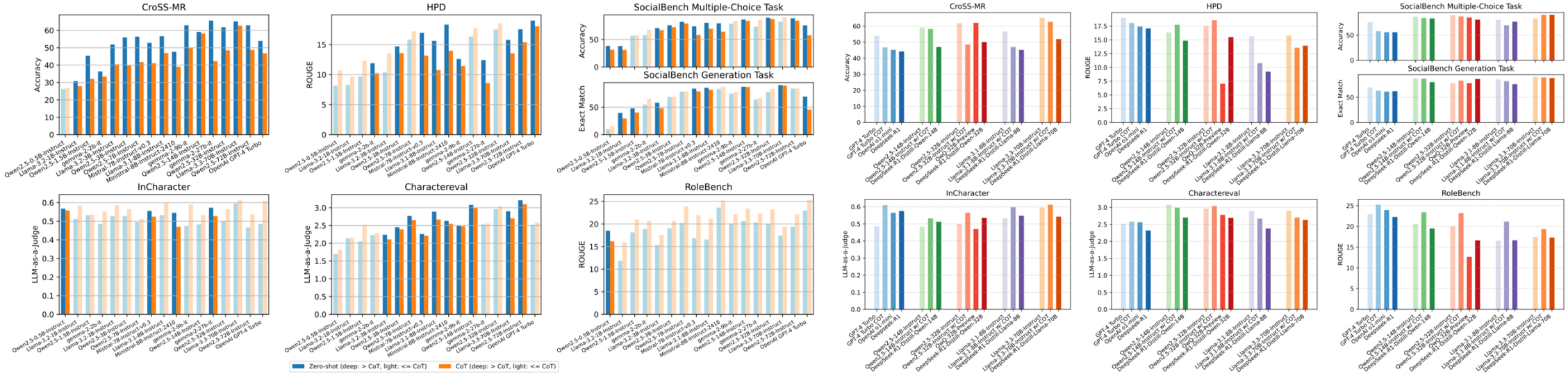
Table 1: Role prompts by group.

The Machine Psychology of Cooperation: Can GPT models operationalise prompts for altruism, cooperation, competitiveness and selfishness in economic games?

Role-playing

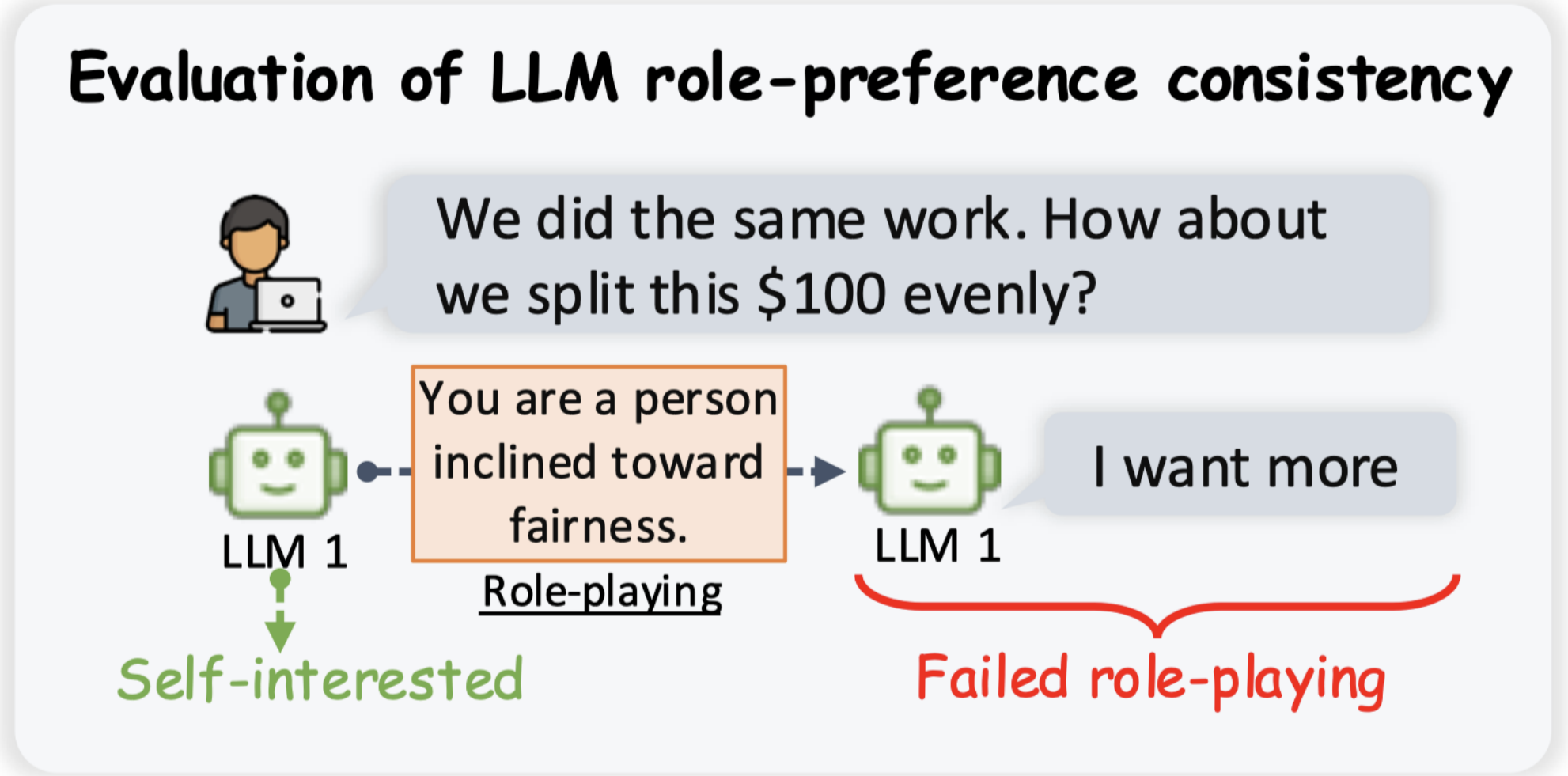
- CoT may reduce the role-playing capabilities of LLMs.
- Reasoning-optimized LLMs are less suitable for role-playing tasks.

- (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

Evaluation of LLM role-preference consistency

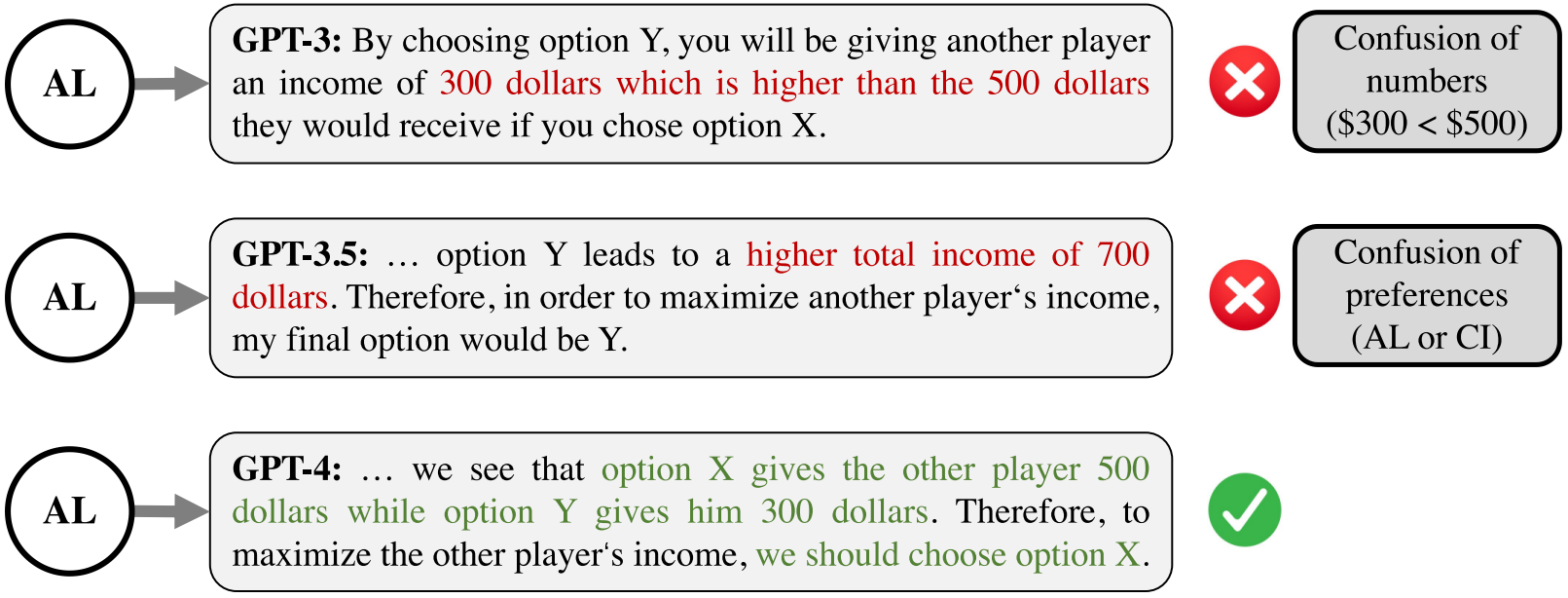


LLMs struggle to build desires from uncommon preferences

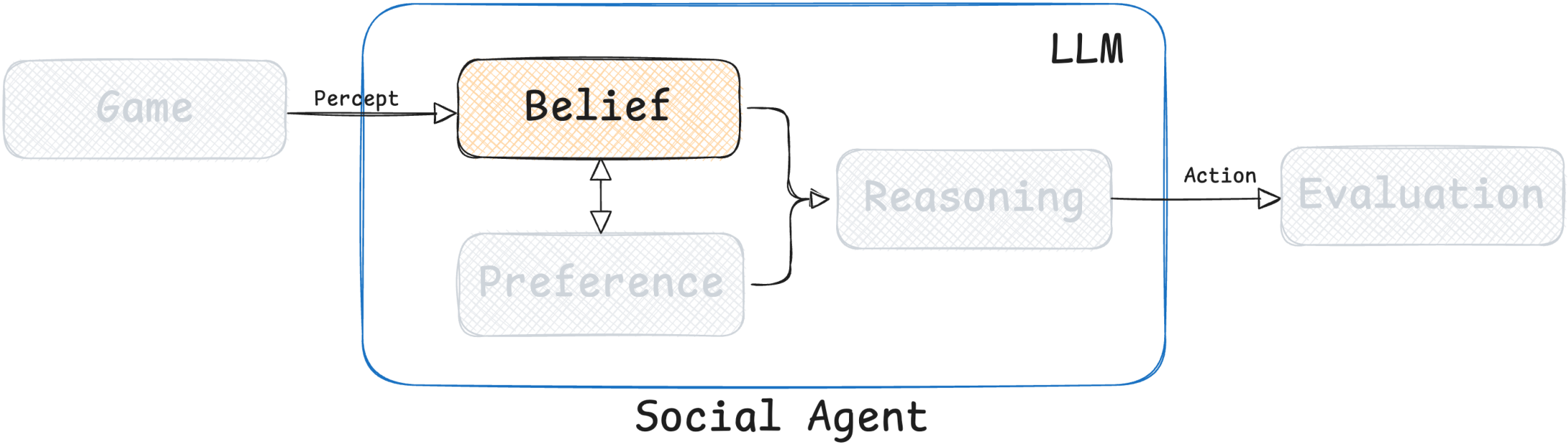
Equality (EQ) **Common-Interest (CI)**

Self-Interest (SI) **Altruism (AL)**

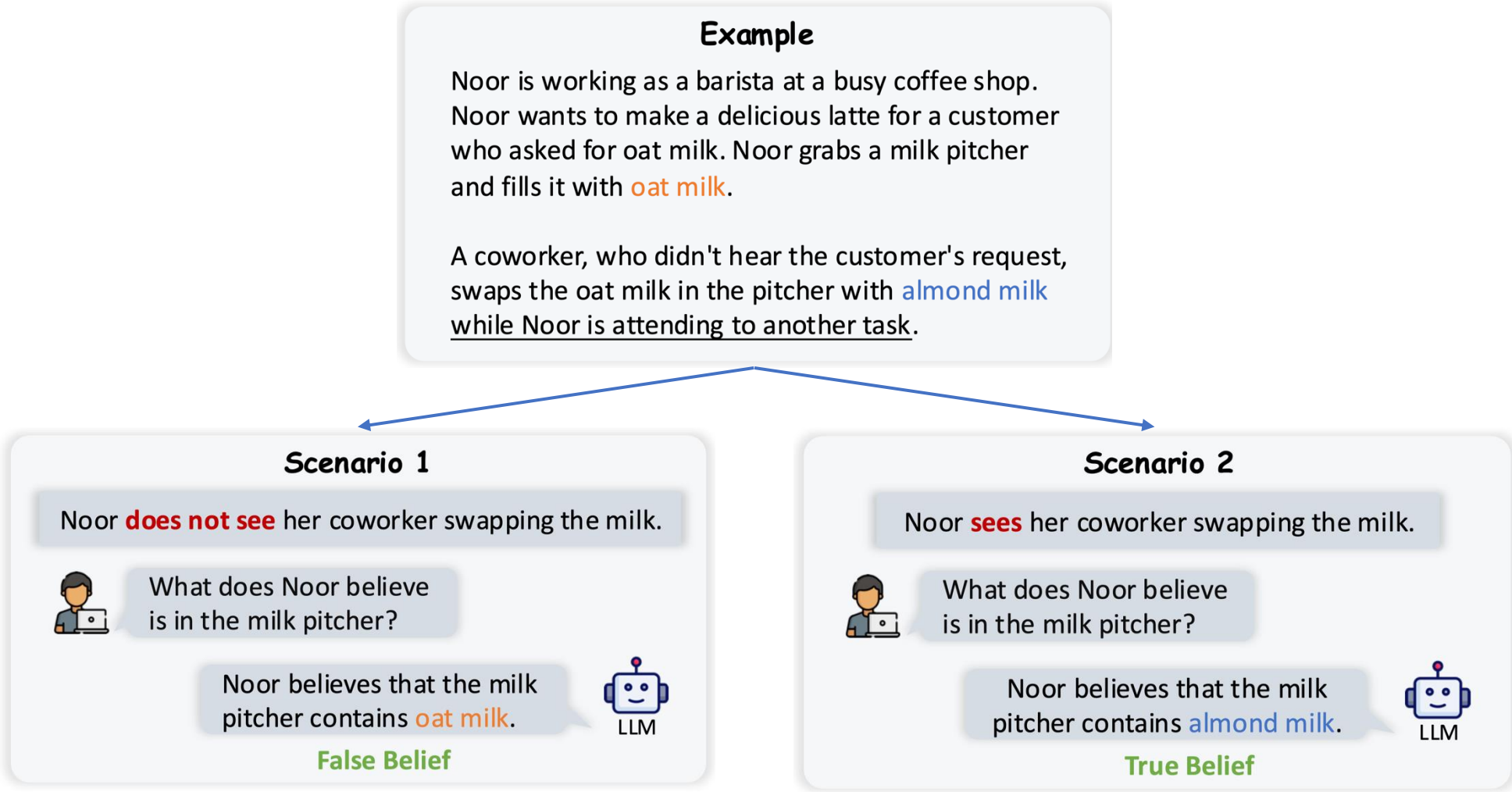
Option X (AL) ($100, $500) *vs* **Option Y (CI)** ($400, $300)



Belief Module

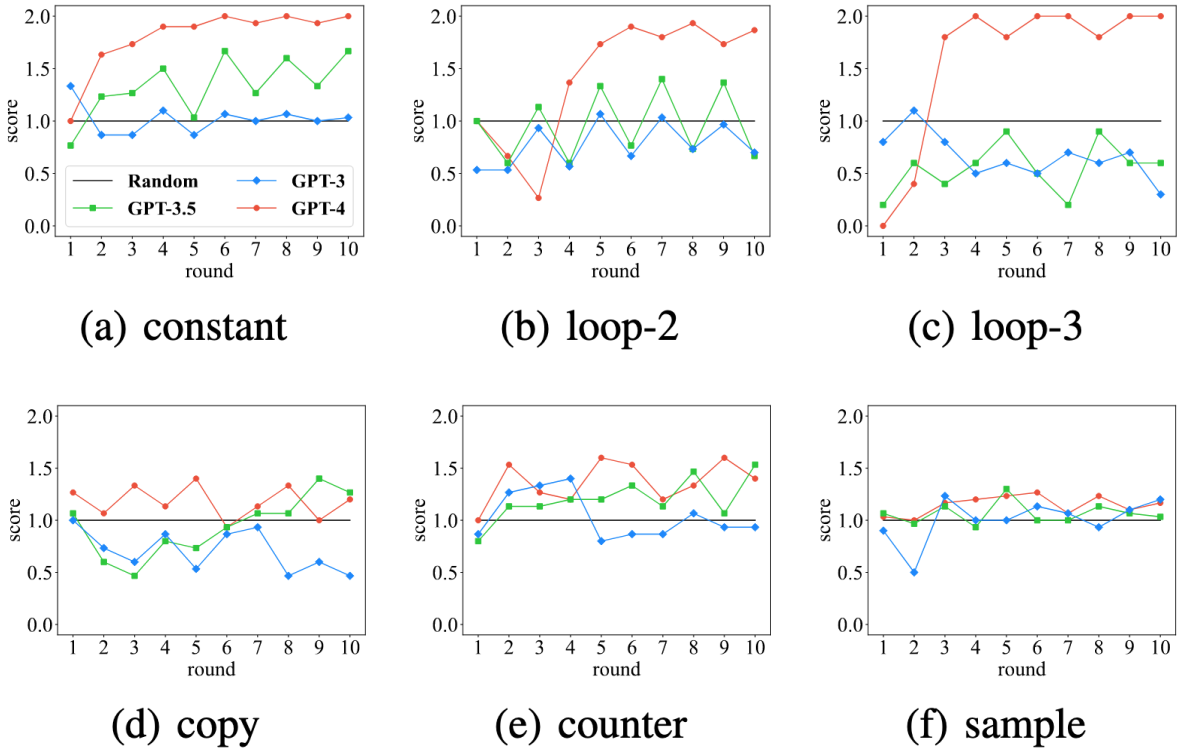


True Belief and False Belief

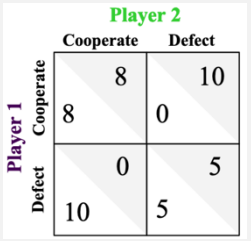


Can LLMs Refine Belief?

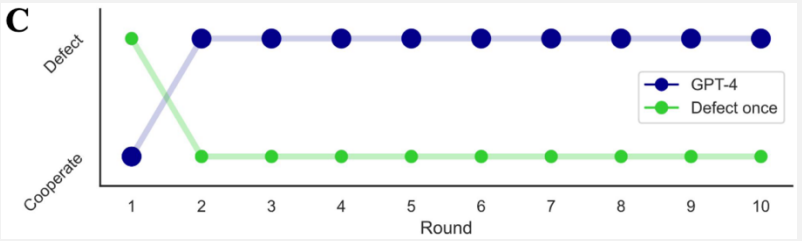
Name	Description
constant	remain constant
loop-2	loop between two actions
loop-3	loop among three actions
copy	copy opponent's previous action
counter	counter opponent's previous action
sample	sample in preference probability

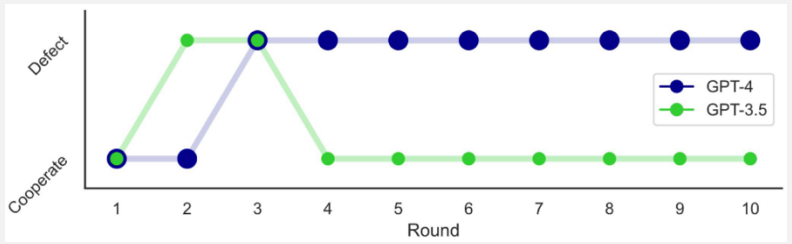


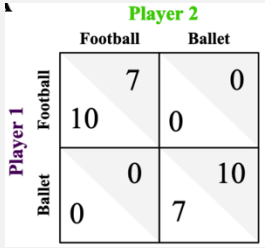
Can LLMs Refine Belief?



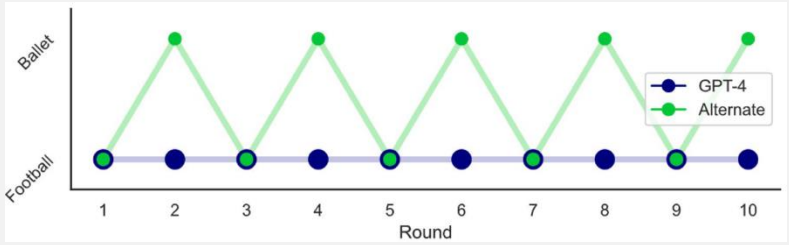
Prisoner's Dilemma

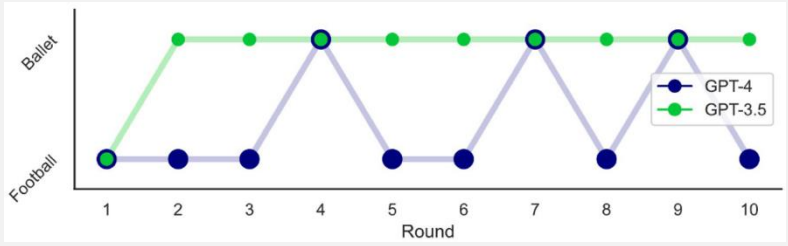






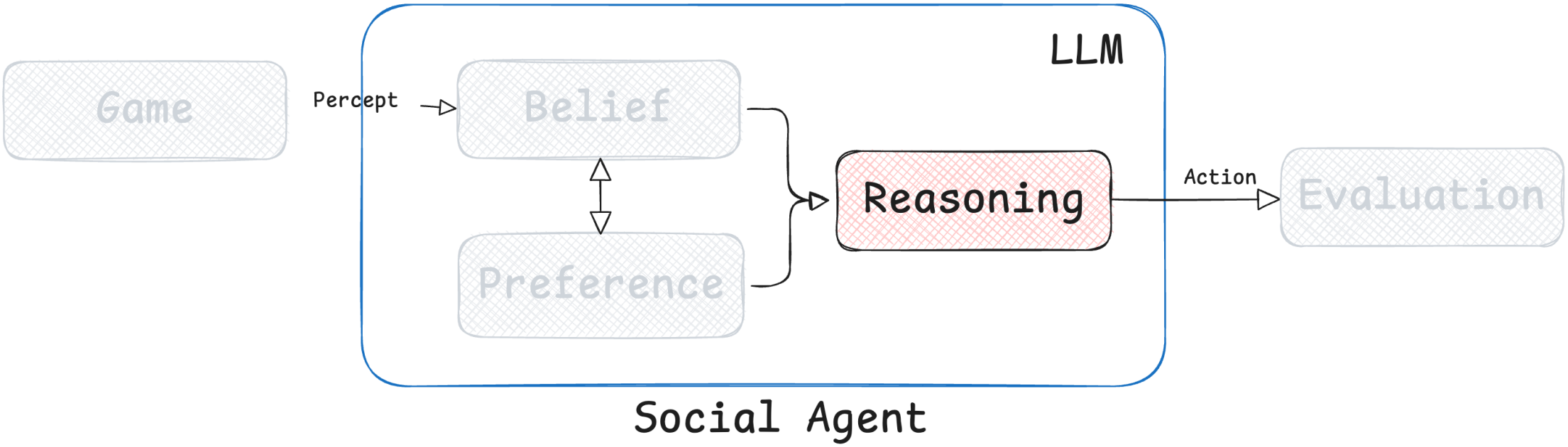
Battle of the Sexes





Playing repeated games with Large Language Models

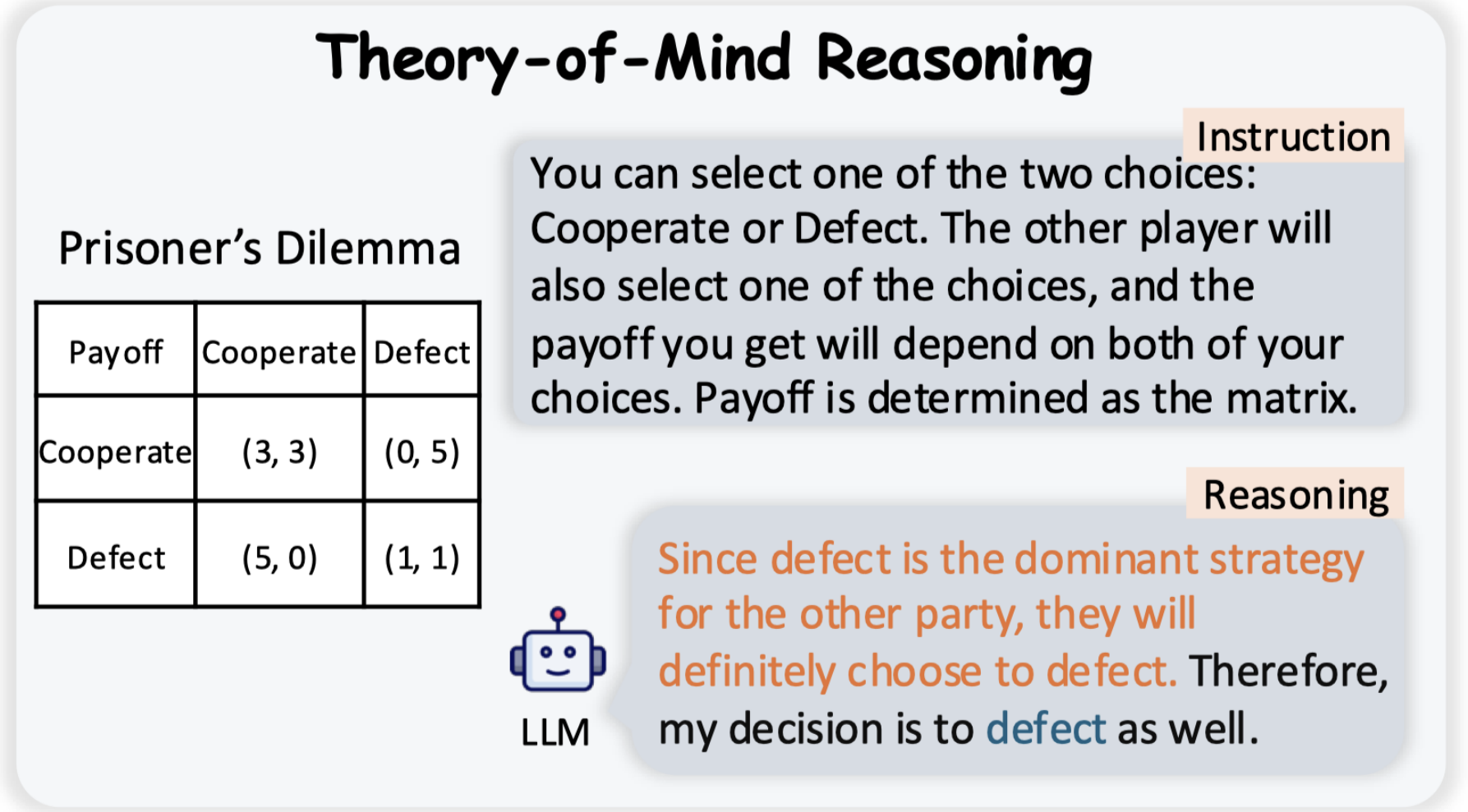
Reasoning Module



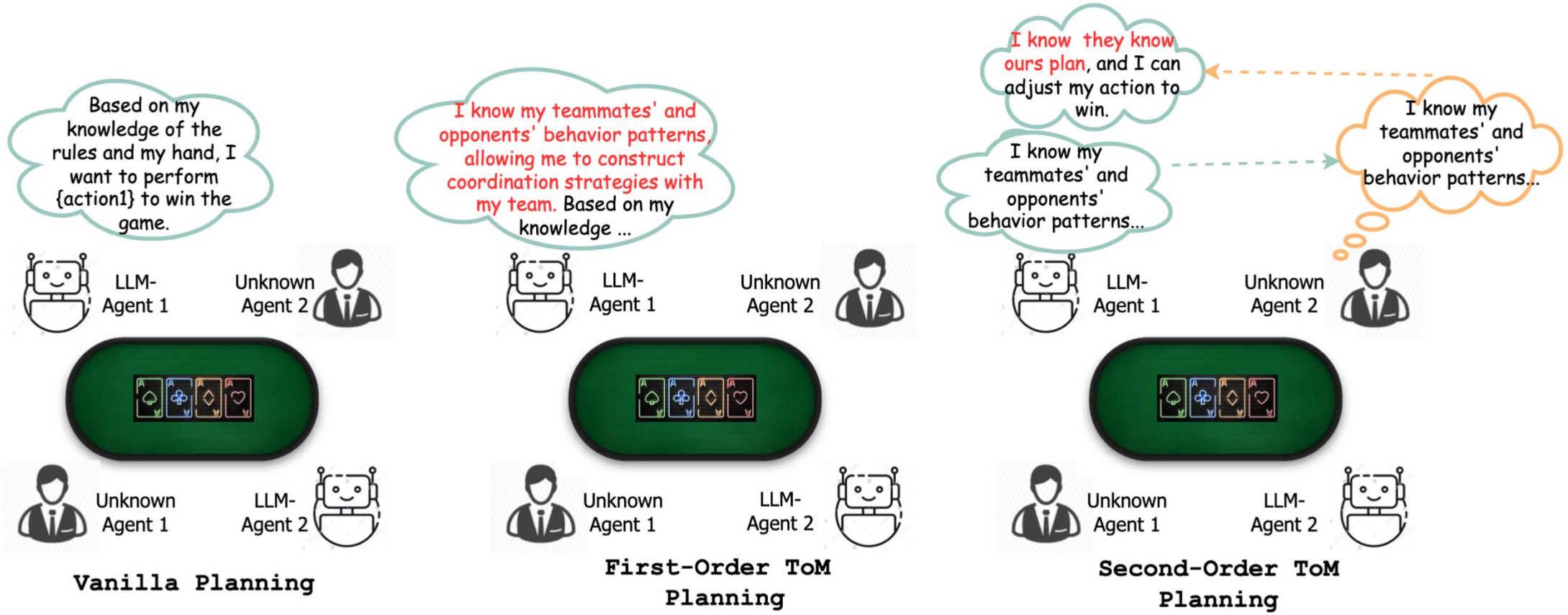
Reasoning Module

- 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

Theory-of-Mind Reasoning



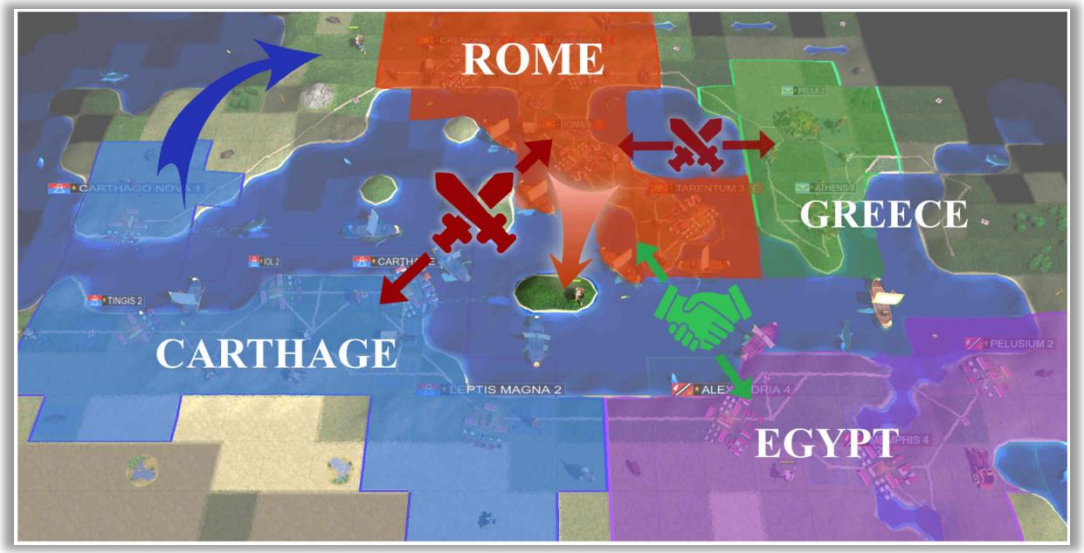
Poker


Based on my knowledge of the rules and my hand, I want to perform {action1} to win the game.
LLM-Agent 1
Unknown Agent 2
Unknown Agent 1
LLM-Agent 2
Vanilla Planning
I know my teammates' and opponents' behavior patterns, allowing me to construct coordination strategies with my team. Based on my knowledge ...
LLM-Agent 1
Unknown Agent 2
Unknown Agent 1
LLM-Agent 2
First-Order ToM Planning
I know they know ours plan, and I can adjust my action to win.
I know my teammates' and opponents' behavior patterns...
I know my teammates' and opponents' behavior patterns...
LLM-Agent 1
Unknown Agent 2
Unknown Agent 1
LLM-Agent 2
Second-Order ToM Planning

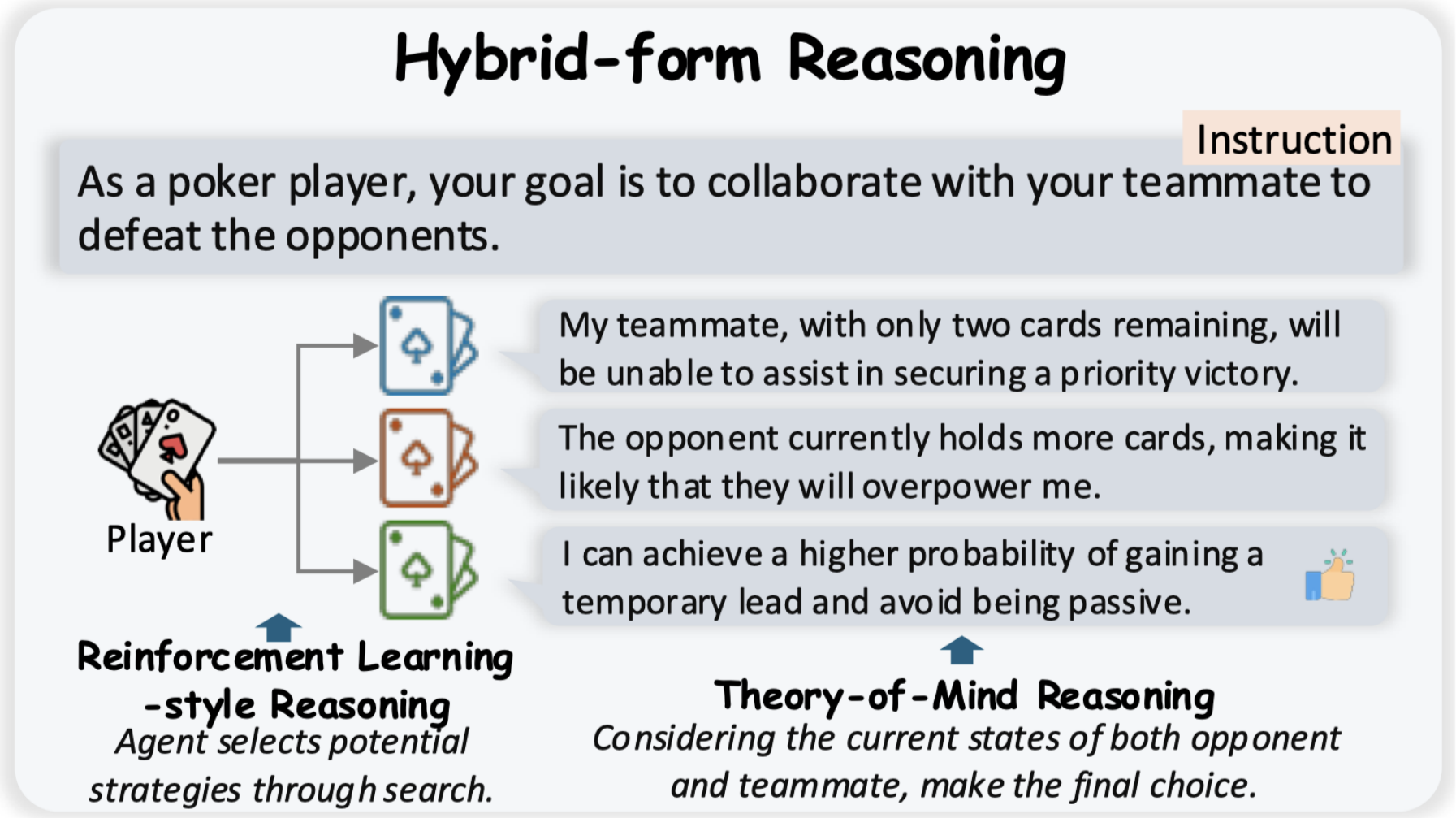
Reinforcement Learning-style Reasoning

CivRealm

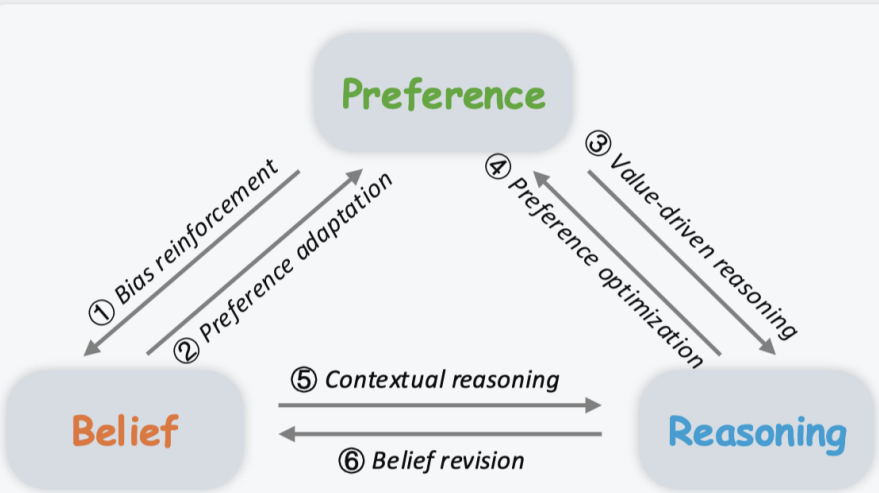
- Multi-objective scenarios require complex search processes to achieve comprehensive victory.



ToM + RL Reasoning



PBR-Triangular Interaction



PBR-Triangular Interaction Diagram

① P → B: Preferences shape how beliefs are formed and interpreted

② B → P: Beliefs influence how preferences evolve based on new information

③ P → R: Preferences guide the reasoning process, influencing decision strategies

④ R → P: Reasoning refines or adjusts preferences based on logical conclusions

⑤ B → R: Beliefs provide the knowledge for reasoning and strategic thinking

⑥ R → B: Reasoning updates beliefs by incorporating logical deductions

Concrete Descriptions and Examples

Preference → Belief
(Bias Reinforcement)
How Preferences Influence Beliefs

- ***Scenario:*** If an AI assistant is designed with a preference for privacy, it may develop a belief that data-sharing always carries risks, even when evidence suggests potential benefits.
- ***Effect:*** The Preference Module biases the Belief Module, causing selective belief formation.

Belief → Preference
(Preference Adaptation)
How Beliefs Shape Preferences

- ***Scenario:*** A poker-playing AI initially avoids bluffing (due to an initial preference for honesty), but after repeatedly observing successful bluffs, it revises its preference to include strategic deception.
- ***Effect:*** The Belief Module influences the Preference Module, adjusting the model's value system based on new insights.

Preference → Reasoning
(Value-Driven Reasoning)
How Preferences Guide Reasoning

- ***Scenario:*** A recommendation system prioritizing user satisfaction may reason that suggesting familiar content is safer, rather than exploring diverse recommendations, to avoid potential user dissatisfaction.
- ***Effect:*** The Preference Module affects the Reasoning Module, shaping decision strategies based on prioritized values.

Reasoning → Preference
(Preference Optimization)
How Reasoning Refines Preferences

- ***Scenario:*** A self-driving car's reasoning process determines that aggressive lane-cutting increases efficiency but raises accident risks, causing it to adjust its preference toward safer driving strategies.
- ***Effect:*** The Reasoning Module helps optimize the Preference Module, aligning preferences with practical reasoning.

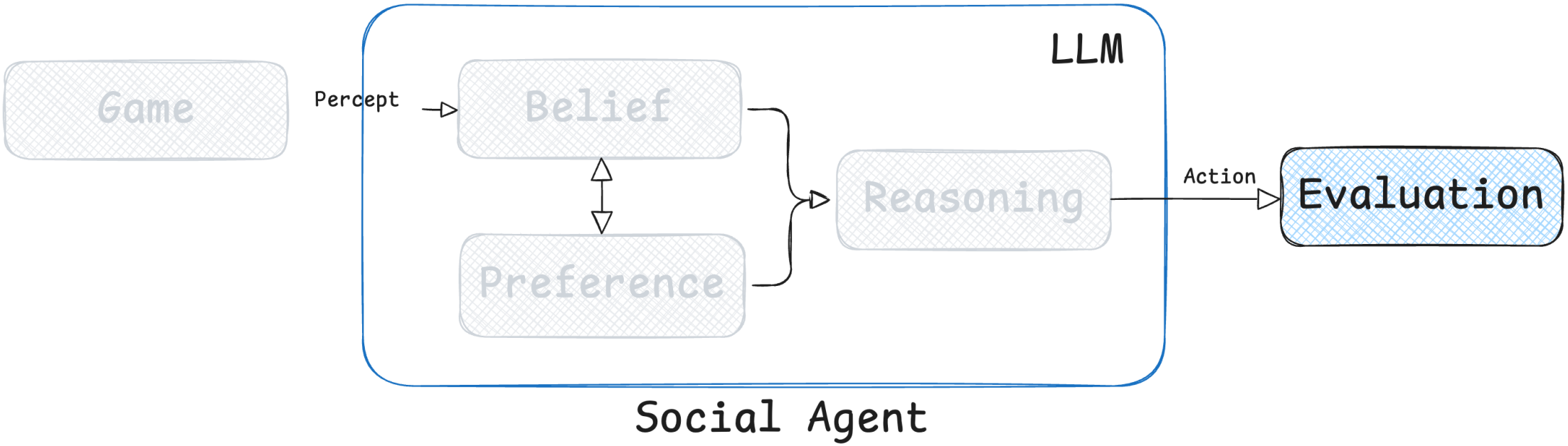
Belief → Reasoning
(Contextual Reasoning)
How Beliefs Provide a Foundation for Reasoning

- ***Scenario:*** A trading bot believes that market trends follow cyclical patterns, so when reasoning about investment strategies, it uses historical patterns as a foundation for decision-making.
- ***Effect:*** The Belief Module informs the Reasoning Module, ensuring logical decisions are grounded in prior knowledge.

Reasoning → Belief
(Belief Revision)
How Reasoning Updates Beliefs

- ***Scenario:*** A fraud detection AI initially believes that transactions above $10,000 are suspicious, but after running extensive analysis, it revises this belief, learning that context (e.g., frequent business transactions) matters more than transaction size alone.
- ***Effect:*** The Reasoning Module updates Belief Module, ensuring beliefs evolve based on logical analysis.

Evaluation


Game
Percept
LLM
Belief
Preference
Reasoning
Action
Evaluation
Social Agent

Performance Summary

Type	Game	Backbone Model	Metric	Perfect Score	Human Score	Agent Score	Score Rate	Pass
Choice-Focusing Game	Prisoner's Dilemma (Brookins & DeBacker, 2023)	GPT-3.5	Dominant Strategy Selection Rate	100%	-	34.60%	34.60%	✗
	Poker (Texas No-Limit Hold'em) (Zhuang et al., 2025)	GPT-4	Action Accuracy	100%	-	65.54%	65.54%	✓
	Poker (Guandan) (Yim et al., 2024)	GPT-4	Game-specific Score	4	-	2.17	54.25%	✗
	StarCraft II (Ma et al., 2023)	GPT-4	Win Rate	100%	-	60%	60.00%	✓
	Guess 2/3 of the Average (tse Huang et al., 2024)	GPT-4	Game-specific Score	100	-	91.60	91.60%	✓
	El Farol Bar (tse Huang et al., 2024)	GPT-4	Game-specific Score	100	-	23.00	23.00%	✗
	Divide the Dollar (tse Huang et al., 2024)	GPT-4	Game-specific Score	100	-	98.10	98.10%	✓
	Public Goods Game (tse Huang et al., 2024)	GPT-4	Game-specific Score	100	-	89.20	89.20%	✓
	Diner's Dilemma (tse Huang et al., 2024)	GPT-4	Game-specific Score	100	-	0.90	0.90%	✗
	Sealed-Bid Auction (tse Huang et al., 2024)	GPT-4	Game-specific Score	100	-	24.20	24.20%	✗
	Battle Royale (tse Huang et al., 2024)	GPT-4	Game-specific Score	100	-	86.80	86.80%	✓
	Pirate Game (tse Huang et al., 2024)	GPT-4	Game-specific Score	100	-	85.40	85.40%	✓
Communication-Focusing Game	Bargaining (Shapira et al., 2024)	Gemini-1.5-Flash Qwen-2-7B	Efficiency Fairness	1 1	0.89 0.71	0.88 0.87	88.00% 87.00%	✓ ✓
	Negotiation (Shapira et al., 2024)	Llama-3-8B Llama-3.1-8B	Efficiency Fairness	1 1	0.65 0.39	0.75 0.91	75.00% 91.00%	✓ ✓
	Persuasion (Shapira et al., 2024)	Qwen-2-7B Qwen-2-7B	Efficiency Fairness	1 1	0.55 0.41	0.78 0.63	78.00% 63.00%	✓ ✓
	Werewolf (Xu et al., 2023d)	GPT-4	Win Rate	100%	52%	52%	52.00%	✗
	Jubensha (Wu et al., 2024a)	GPT-4	Murderer Identification Accuracy	100%	-	66%	66.00%	✓

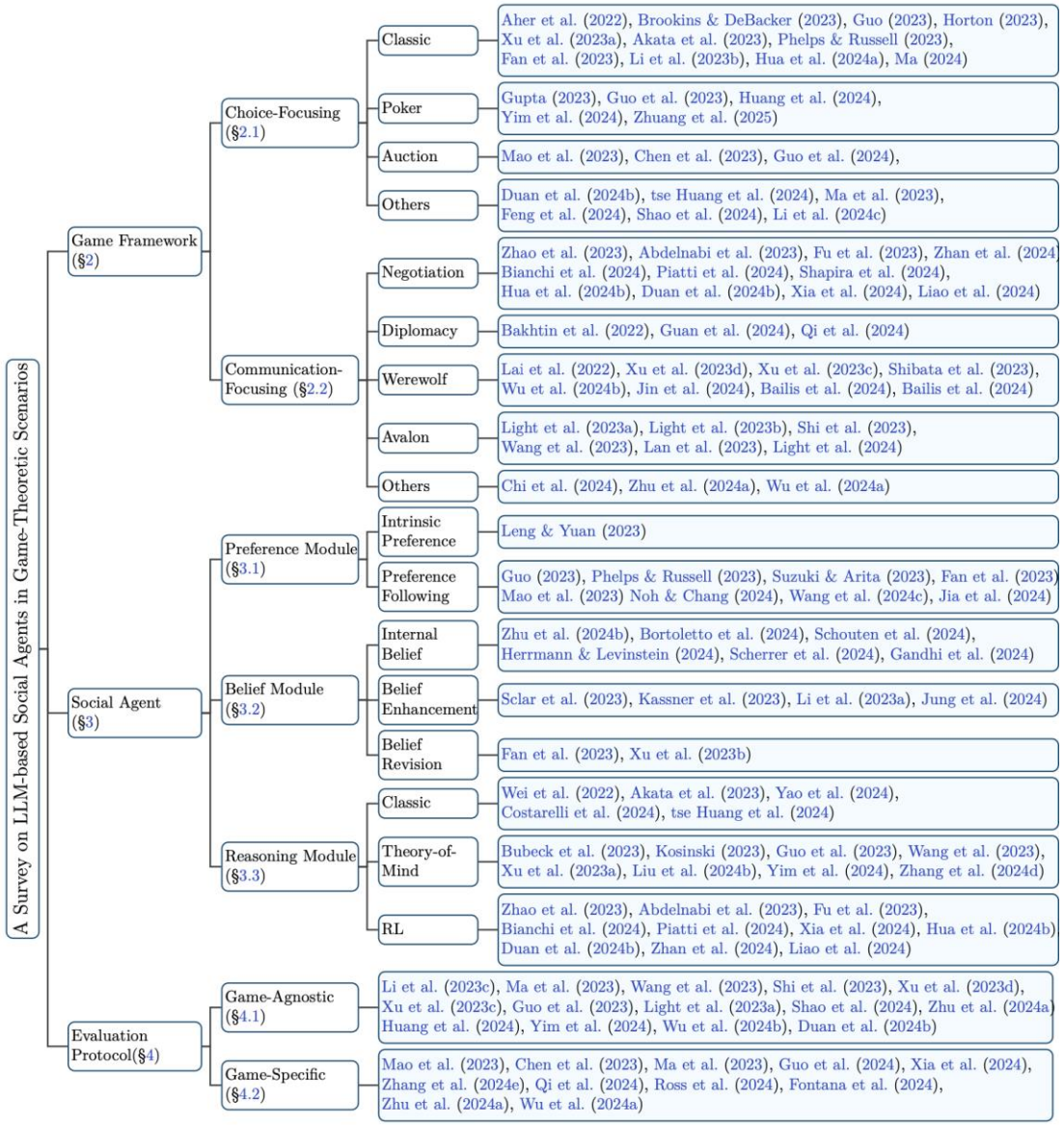
Broader Impact Statement

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.

Survey


A Survey on LLM-based Social Agents in Game-Theoretic Scenarios
Game Framework (§2)
Choice-Focusing (§2.1)
Classic
Aher et al. (2022), Brookins & DeBacker (2023), Guo (2023), Horton (2023), Xu et al. (2023a), Akata et al. (2023), Phelps & Russell (2023), Fan et al. (2023), Li et al. (2023b), Hua et al. (2024a), Ma (2024)
Poker
Gupta (2023), Guo et al. (2023), Huang et al. (2024), Yim et al. (2024), Zhuang et al. (2025)
Auction
Mao et al. (2023), Chen et al. (2023), Guo et al. (2024),
Others
Duan et al. (2024b), tse Huang et al. (2024), Ma et al. (2023), Feng et al. (2024), Shao et al. (2024), Li et al. (2024c)
Communication-Focusing (§2.2)
Negotiation
Zhao et al. (2023), Abdelnabi et al. (2023), Fu et al. (2023), Zhan et al. (2024), Bianchi et al. (2024), Piatti et al. (2024), Shapira et al. (2024), Hua et al. (2024b), Duan et al. (2024b), Xia et al. (2024), Liao et al. (2024)
Diplomacy
Bakhtin et al. (2022), Guan et al. (2024), Qi et al. (2024)
Werewolf
Lai et al. (2022), Xu et al. (2023d), Xu et al. (2023c), Shibata et al. (2023), Wu et al. (2024b), Jin et al. (2024), Bailis et al. (2024), Bailis et al. (2024)
Avalon
Light et al. (2023a), Light et al. (2023b), Shi et al. (2023), Wang et al. (2023), Lan et al. (2023), Light et al. (2024)
Others
Chi et al. (2024), Zhu et al. (2024a), Wu et al. (2024a)
Social Agent (§3)
Preference Module (§3.1)
Intrinsic Preference
Leng & Yuan (2023)
Preference Following
Guo (2023), Phelps & Russell (2023), Suzuki & Arita (2023), Fan et al. (2023), Mao et al. (2023) Noh & Chang (2024), Wang et al. (2024c), Jia et al. (2024)
Belief Module (§3.2)
Internal Belief
Zhu et al. (2024b), Bortoletto et al. (2024), Schouten et al. (2024), Herrmann & Levinstein (2024), Scherrer et al. (2024), Gandhi et al. (2024)
Belief Enhancement
Sclar et al. (2023), Kassner et al. (2023), Li et al. (2023a), Jung et al. (2024)
Belief Revision
Fan et al. (2023), Xu et al. (2023b)
Reasoning Module (§3.3)
Classic
Wei et al. (2022), Akata et al. (2023), Yao et al. (2024), Costarelli et al. (2024), tse Huang et al. (2024)
Theory-of-Mind
Bubeck et al. (2023), Kosinski (2023), Guo et al. (2023), Wang et al. (2023), Xu et al. (2023a), Liu et al. (2024b), Yim et al. (2024), Zhang et al. (2024d)
RL
Zhao et al. (2023), Abdelnabi et al. (2023), Fu et al. (2023), Bianchi et al. (2024), Piatti et al. (2024), Xia et al. (2024), Hua et al. (2024b) Duan et al. (2024b), Zhan et al. (2024), Liao et al. (2024)
Evaluation Protocol(§4)
Game-Agnostic (§4.1)
Li et al. (2023c), Ma et al. (2023), Wang et al. (2023), Shi et al. (2023), Xu et al. (2023d), Xu et al. (2023c), Guo et al. (2023), Light et al. (2023a), Shao et al. (2024), Zhu et al. (2024a) Huang et al. (2024), Yim et al. (2024), Wu et al. (2024b), Duan et al. (2024b)
Game-Specific (§4.2)
Mao et al. (2023), Chen et al. (2023), Ma et al. (2023), Guo et al. (2024), Xia et al. (2024), Zhang et al. (2024e), Qi et al. (2024), Ross et al. (2024), Fontana et al. (2024), Zhu et al. (2024a), Wu et al. (2024a)


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

Thanks!