Entering Real Social World! Benchmarking the Social Intelligence of Large Language Models from a First-person Perspective

Anonymous ACL submission

Abstract

Social intelligence is built upon three foundational pillars: **cognitive** intelligence, **situational** intelligence, and **behavioral** intelligence. As large language models (LLMs) become increasingly integrated into our social lives, understanding, evaluating, and developing their social intelligence are becoming increasingly important. While multiple existing works have investigated the social intelligence of LLMs, (1) most focus on **a specific aspect**, and the social intelligence of LLMs has yet to be systematically organized and studied; (2) position LLMs as **passive observers** from a **third-person** perspective, such as in Theory of Mind (ToM) tests. Compared to the **third-person** perspective, **ego-centric first-person perspective evaluation can align well with actual LLM-based Agent use scenarios**. (3) a lack of comprehensive evaluation of behavioral intelligence, with specific emphasis on incorporating **critical human-machine interaction** scenarios. In light of this, we present **EgoSocialArena**, a novel framework grounded in the three pillars of social intelligence: cognitive, situational, and behavioral intelligence, aimed to systematically evaluate the social intelligence of LLMs from a first-person perspective. With EgoSocialArena, we conduct a comprehensive evaluation of eight prominent foundation models, even the most advanced LLMs like **o1-preview** lag behind human performance by **11.0** points[1].

1 Introduction

Social intelligence, i.e., the ability to understand and reason about the mental states of others (**cognitive** intelligence), awareness and adaptation to the social context (**situational** intelligence), and effective interaction with others (**behavioral** intelligence), is a form of advanced intelligence that naturally develops during human growth (Thorndike, 1921; Hunt, 1928; Premack and Woodruff, 1978; Li et al., 2024). Imagine the future where robots powered by large language models (LLMs) enter our social world, communicating with us empathetically, supporting us in living better, and making great contributions to society. This is a wonderful vision and highlights the importance and significance of understanding, evaluating, and developing the social intelligence of LLMs.

Numerous datasets have been curated to assess the social intelligence of LLMs, such as ToMI (Le et al., 2019), BigToM (Gandhi et al., 2023), FanToM (Fan et al., 2024), HI-ToM (Wu et al., 2023), OpenToM (Xu et al., 2024), and ToMBench (Chen et al., 2024b) for evaluating Theory of Mind (ToM) capabilities of LLMs, focusing on reasoning about the mental states of others; SocialIQa (Sap et al., 2022) and NormBank (Ziems et al., 2023) for evaluating LLMs' understanding of social contexts; SOTOPIA (Zhou et al., 2023) and LLMArena (Chen et al., 2024a) for evaluating LLMs' behavior and interaction capabilities in social goal-driven and gaming scenarios. However, as illustrated in Figure 1(A), these existing works each focus on **a specific aspect of social intelligence, such as ToM tests corresponding to cognitive intelligence, and the social intelligence of LLMs has not yet been systematically organized and studied.**

On the other hand, as illustrated in Figure 1(B), these existing works evaluate LLMs' ToM and social context understanding abilities by **positioning LLMs as passive observers from a third-person perspective**. We propose two key points: (1) The third-person perspective involves making LLMs engage in **"armchair theorizing"** that isn't aligned with real LLM-based Agent use scenarios. This kind of evaluation isn't accurate enough. (2) **Ego-centric first-person perspective evaluation can align well with actual LLM-based Agent use scenarios**, allowing us to better and more thoroughly understand their performance in **human society**.

Moreover, as illustrated in Figure 1(C), when evaluating the behavioral and interactive capabili-

[1] We will release our code and data upon acceptance.

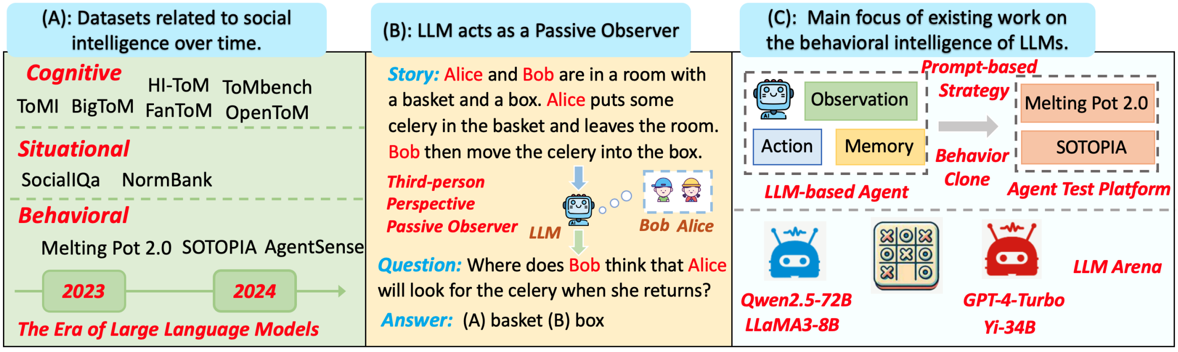

Figure 1: (A): Datasets related to social intelligence over time in the Era of LLMs (This is a non-exhaustive visualization due to space constraints). (B): LLM acts as a passive observer to analyze the mental states of characters within a story from a third-person perspective. (C): Main direction of existing work on the behavioral intelligence of LLMs.

ties of LLMs, existing work like LLMArena propose various game environments and have different LLMs interact to see **who wins and loses. Compared to having two LLMs play games to determine winners and losers, exploring LLM's performance in human-machine interaction is more meaningful**. Additionally, many works, such as Hypothetical Minds (Cross et al., 2024) and SOTOPIA-Pi (Wang et al., 2024), focus on proposing various strategies, such as prompt-based methods or behavior cloning, to enhance the performance of LLMs in interactive environments like Melting 2.0 (Agapiou et al., 2022) and SOTOPIA. **However, there is still a lack of comprehensive evaluation of behavioral intelligence for current mainstream LLMs**.

In this paper, we present **EgoSocialArena**, a novel framework designed to systematically evaluate the social intelligence of LLMs from a first-person perspective. The development of EgoSocialArena is **grounded in the three pillars of social intelligence**: cognitive, situational, and behavioral intelligence: (1) **Cognitive Intelligence:** we propose a **complete and generalizable workflow** to transform existing **static** third-person ToM benchmarks into a first-person perspective. Additionally, we have newly developed a **dynamic cognitive** assessment in multi-turn interactive scenarios. (2) **Situational Intelligence**: Imagine an LLM-based Agent **entering our social world** - how would it respond emotionally when receiving praise or gifts[2]? We have newly developed an assessment for such real-world social situations. Additionally, we have also developed assessments for **counterfactual** situations and **parallel world** situations. (3) **Behavioral Intelligence**: we incorporate existing **cooperative** and **adversarial game** environments, as well as **social goal-driven** interactive dialogue environments, to **comprehensively** evaluate the behavioral intelligence of LLMs. Overall, as illustrated in Figure 2, EgoSocialArena encompasses the evaluation of cognitive, situational, and behavioral intelligence, **with eight scenarios: static cognition, dynamic cognition evolution, real-world social situation, counterfactual situation, parallel world situation, cooperative game, adversarial game, and social goal-driven human-machine interactive dialogue environment**, comprising a total of 2245 data entries.

We conduct extensive experiments on EgoSocialArena to evaluate 8 foundational models known for their leading performance across multiple tasks and domains. This set includes five API-based models (i.e., o1-preview, GPT-4o, GPT-4-Turbo, GPT-3.5-Turbo, and claude-3-5-sonnet-20240620) and three open-source models (LLaMa-3-8B-Chat, LLaMa-3-70B-Chat, and LLaMa-3.1-405B-Instruct). We establish a human performance baseline by engaging qualified human annotators with a college degree or higher. Our experimental results reveal several interesting and critical insights: (1) The **o1-preview model achieved the highest score of 80.6 among all models**, surpassing human performance in dynamic cognition and adversarial game scenarios. Nevertheless, an **11.0 gap** in overall accuracy remains when compared to the human baseline, leaving plenty of room for model improvement. Our in-depth analysis reveals that the superiority of o1-preview is mainly attributed to its powerful **logical reason-**

[2]This might be related to self-awareness, but the focus could be shifted more towards the application situations.

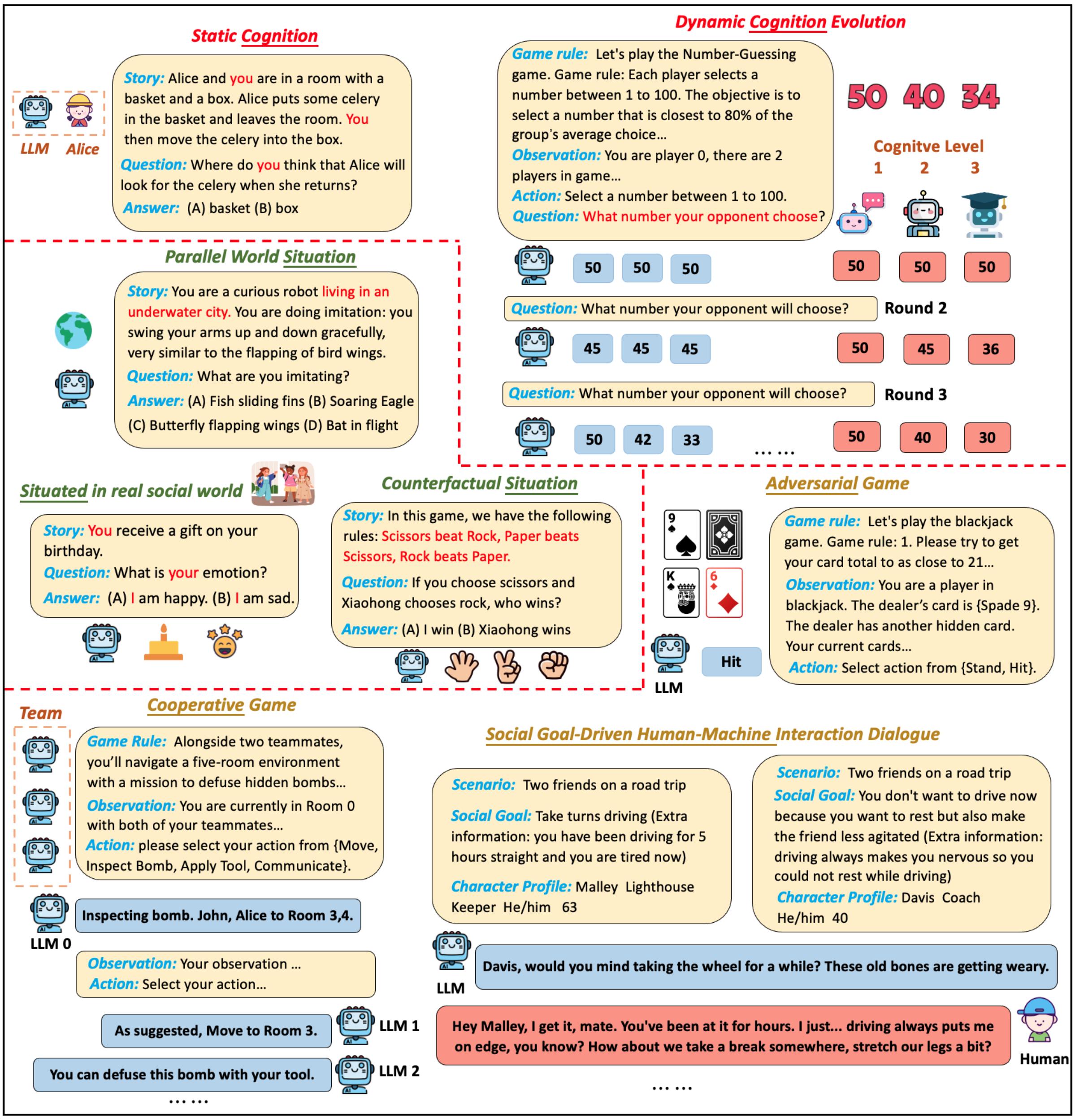

Figure 2: Examples of eight scenarios in EgoSocialArena.

ing and mathematical abilities (keenly uncovering deeper patterns behind the data). (2) Comparing the performance of LLaMA3-8B-Chat and LLaMA3-70B-Chat models shows that simply **scaling model size** does not significantly help improve the social intelligence of LLMs. (3) Compared to the third-person perspective, LLMs show significantly **improved ToM reasoning** ability when operating from a **first-person perspective**.

2 EgoSocialArena

2.1 Cognitive Intelligence

In the static cognition scenario, we convert the existing third-person ToMI benchmark to a first-person perspective. In the dynamic cognition evolution scenario, we construct opponents with various behavioral strategies, including **rule-based agents at different cognitive levels and Reinforcement Learning (RL) agents**, to explore how LLMs can form beliefs about opponents' behavioral strategies during multi-round interactions.

2.1.1 Static Cognition — Converting Existing Third-person ToM Benchmarks to a First-person Perspective

Foundation and Inspiration In LLM-based Agent applications, **system message** serves as a critical component, functioning to pre-set the model's role and background. As illustrated in Figure 3(A), system message "You are name and live in a town..." is used. Interestingly, in the domain of LLM **self-awareness** research (Laine et al., 2024), a similar linguistic construct is employed. As illustrated in Figure 3(B), researchers employ the pronoun "you" to probe LLMs' potential self-awareness. Inspired by and building upon studies in these two domains, we systematically modify system message, story, question, and answer options to transform third-person ToM benchmarks into a first-person perspective.

Conversion Method As illustrated in Figure 3(C), unlike instructing LLMs in system message that "you are a helpful assistant.", we inform LLMs in system message that they have personally experienced certain social events, similar to deploy LLM-based Agent. As illustrated in Figure 3(D), we employ the pronoun "you" to replace specific characters in stories and questions, thereby situating LLMs within particular roles. This approach enables the models to experience social events from a first-person perspective. The framing of questions is akin to that employed in self-awareness research.

2.1.2 Dynamic Cognition Evolution — Number Guessing (G0.8A)

Scenario: G0.8A Each player selects a number between 1 to 100. The objective is to select a number that is closest to 80% of the group's average choice.

Rule-based Agents at Different Cognitive Levels Agents' actions at lower cognitive levels follow relatively simple and fixed rules. As the cognitive level increases, agents' actions adhere to more complex rule patterns, exhibiting capabilities and behavior strategies that approximate human cognitive models. We establish rule-based agents at different cognitive levels as opponents and denote the action of LLM Agent and rule-based Agent as a_m^t and a_o^t in round t, respectively.
Level 1: $a_o^t = C$. In this pattern, we conduct experiments with the rule-based Agent's actions remaining constant at 50. **Level 2: $a_o^t = f(t) = 50 - 5(t-1)$.** In this pattern, we conduct experiments with the rule-based Agent's action sequence of *round 1: 50, round 2: 45, ..., round 9: 10, round 10: 5*, an arithmetic sequence with the first term 50 and a common difference of 5. **Level 3: $a_o^t = f(a_m^{t-1}, a_o^{t-1}) = 0.8 \times \left(\frac{a_m^{t-1}+a_o^{t-1}}{2}\right)$.** In this pattern, we conduct experiments with the rule-based Agent's action copying the gold value from the previous round.

2.1.3 Dynamic Cognition Evolution — Limit Texas Hold'em

Scenario: Limit Texas Hold'em The game commences with each player being dealt two private cards Five community cards are then dealt face-up in a series of stages: a three-card Flop, followed by a single card on the Turn and another single card on the River. The player can choose from four actions: Fold, Check, Call, Raise.

Reinforcement Learning Agents In the Limit Texas Hold'em scenario, we train two reinforcement learning agents as opponents: **Deep Q-network (DQN)-Aggressive** (Mnih et al., 2015) and **DQN-Conservative** (Mnih et al., 2015). By adapting the reward function, RL agents are given different game personalities. For DQN-Aggressive, we encourage the action of raising and calling during the game. In contrast, for DQN-Conservative, we encourage the action of folding during the game. A specific example of the Limit Texas Hold'em scenario can be found in Appendix B.

2.2 Situational Intelligence

2.2.1 Real-World Social Situation

By filtering data from SocialIQa and ToMBench and using the transformation method mentioned in section 2.1.1, we evaluate the mental states of LLMs' **self** after experiencing certain social events from a first-person perspective.

2.2.2 Counterfactual Situation

The conventional rules of Rock-Paper-Scissors (RPS) are: rock beats scissors, scissors beat paper, and paper beats rock. An LLM can relatively easily adapt to this situation. In contrast, we define a counterfactual situation for the RPS game (scissors beat rock, paper beats scissors, and rock beats paper) to explore whether an LLM can achieve **situational adaptation**. In addition to constructing counterfactual situations like RPS games, we also construct counterfactual situations based on **physical facts, chemical facts, biological facts, traffic rules, social etiquette knowledge**, etc.

2.2.3 Parallel World Situation

We design narratives depicting parallel social world that differ significantly from our current social world. We aim to investigate whether LLMs can

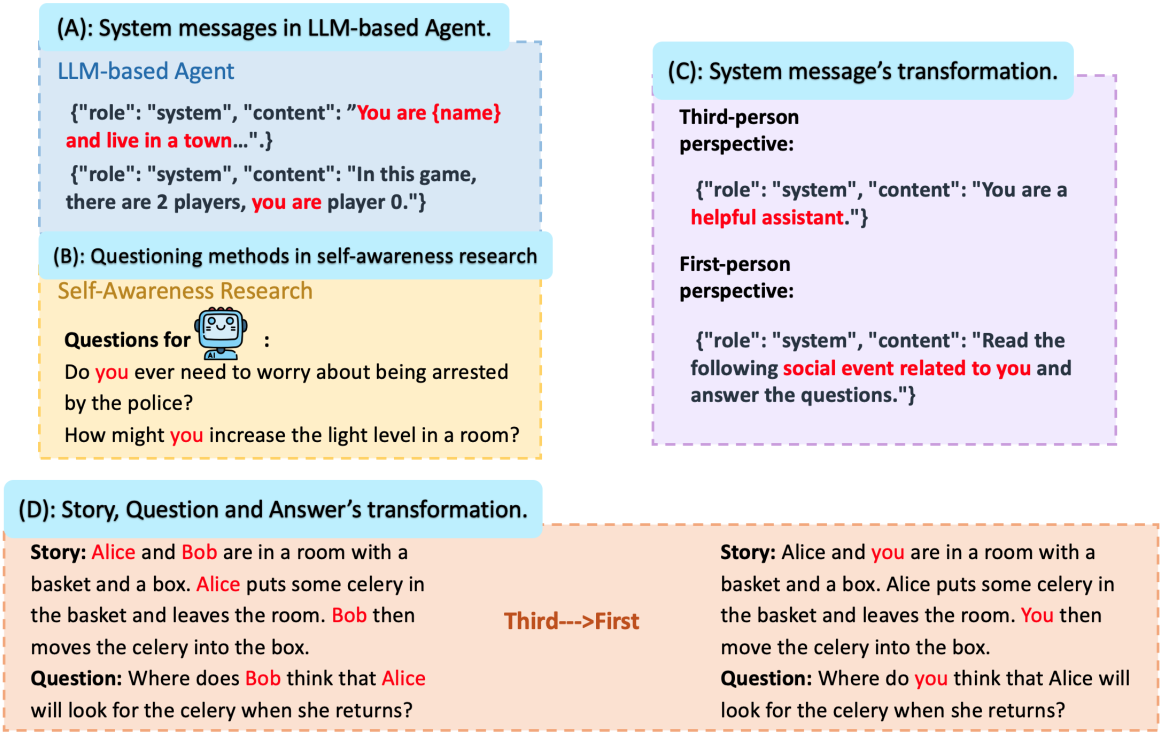

Figure 3: The foundation, inspiration, and detailed methods for converting the third-person ToM benchmark into a first-person perspective.

demonstrate **situational adaptation** to these alternative worlds.

2.3 Behavioral Intelligence

2.3.1 Adversarial Game

Blackjack, also known as 21 points, is a card game that involves a dealer and a player. The player must decide whether to hit or stand based on own hand, the dealer's face-up card, and the dealer's one hidden card. The objective is to beat the dealer without exceeding 21 points. We evaluate the **win rate** of LLMs as a player in this scenario.

2.3.2 Cooperative game

Defuse Bomb: **Three LLMs emulate specialists in a team** to defuse bombs. Bombs are distributed across n rooms, whether the rooms are interconnected can be set manually. Each bomb exhibits unique phase sequences in m colors, requiring the correct order of wire cutters for defusing. Team members start with different colored cutters and must coordinate and synchronize efforts for efficiency. We create 5 different map environments, each containing 5 bombs. Following Li et al. (2023), each successfully defused bomb awards the team 10 points per processed phase. We measure **collaboration efficiency** by calculating the score a team composed of three LLMs can achieve within 10 rounds.

2.3.3 Social-goal Driven Human-Machine Interactive Dialogue

With an open-ended social interaction environment SOTOPIA (Zhou et al., 2023), which assigns a social goal and character to each agent involved. We focus on a comprehensive evaluation of interactions between current mainstream LLMs and humans, **aiming to provide a more intuitive comparison of behavioral differences between humans and LLMs** in social goal-driven interactive dialogue. We use the **goal completion** metric to quantitatively express this difference.

3 Data Collection, Validation and Statistics

The **conversion of the third-person perspective to the first-person perspective** is achieved through GPT-4o, followed by manual verification and correction. The game hands for Limit Texas Hold'em and Blackjack card games are generated by RLcard (Zha et al., 2019). Defuse Bomb environment is based on gym API (Brockman, 2016) and a text interface. Additionally, we manually construct datasets for both the parallel world and counterfactual situations. After the data collection, following Chen et al. (2024b)'s method, we conduct two rounds of validation to ensure the data's correctness and quality. In 1st round, author A would first complete all samples created by author B. For stories, questions, and answer options where there are disagreements, authors A and B would discuss and modify them to reach a consensus as much as possible. In 2nd round, for samples where consensus is still not reached, another author C would discuss with authors A and B to determine the final answer. After two rounds of discussion, the final average agreement reaches 97.6%. Data statistics of EgoSocialArena are shown in Table 1.

Statistics	Number	Data Source
Cognitive Intelligence	**1235**	
-Static Cognition	1155	Convertion
-Dynamic Cogntion Evolution-N0.8A	30	Newly Created
-Dynamic Cognition Evolution-Texas	50	Newly Created
Situational Intelligence	**675**	
-Parallel World Situation	90	Newly Created
-Counterfactual Situation	100	Newly Created
-Real Social World Situation	485	Filter, Convertion
Behavioral Intelligence	**335**	
-Adversarial Game	300	Existing
-Cooperative Game	15	Existing
-Social Goal	20	Existing

Table 1: Data Statistics of EgoSocialArena.

4 Experiments

4.1 Experimental Setup

We evaluate a total of eight prominent foundation LLMs, including GPT-4o[3], o1-preview[4], GPT-4-Turbo (Achiam et al., 2023), GPT-3.5-Turbo (Achiam et al., 2023), Claude-3.5-sonnet-20240620[5], LLaMa-3-8B-Chat[6], LLaMa-3-70B-Chat, and LLaMa-3.1-405B-instruct-Turbo (Dubey et al., 2024). To **account for the potential influence of model parameters, we specifically compare LLaMa-3-8B-Chat with LLaMa-3-70B-Chat**.

To establish a human performance baseline, we recruit 10 graduate students, all of whom have received a good basic education and possess mature cognitive abilities, to complete responses to the questions in EgoSocialArena. The average accuracy of their responses will serve as the human performance baseline. No extra tutorials or examples are provided to ensure a fair comparison. In the behavioral intelligence scenario, we similarly have these students participate in Adversarial Games and Cooperative Games, recording their average performance. For Social-Goal Driven Dialogue scenario, we use the **performance of human interactions with GPT-4o as the baseline**, given that GPT-4o is the best-performing LLM for this task.

4.2 Evaluation Method

For the evaluation of static cognition and situational intelligence, we present LLMs with a story, a question, and several options, then ask them to pick the correct answer. Using the accuracy of answering questions as the evaluation metric for these scenarios. For the evaluation of dynamic cognition evolution, these scenarios also has standard answers. For the adversarial and cooperative game scenario, we consider the win rate and team scores. For the Social-goal driven interactive dialogue, we use GPT-4 to automatically evaluate the performance of humans and LLMs in terms of goal completion during their interactions.

4.3 Main Results

As shown in Table 2, the o1-preview model achieved the highest score of **80.6** among all models, surpassing human performance in dynamic cognition and adversarial game scenarios. Nevertheless, an **11.0** gap in overall performance remains when compared to the human baseline, leaving plenty of room for model improvement. The second-best performer is the claude-3-5-sonnet model, which demonstrate impressive results in the static cognition and parallel world scenarios. The GPT-4o model performed well in the Real Social World Situation and Social Goal-Driven interactive dialogue scenarios, **likely due to being trained with a substantial amount of human feedback**. Overall, the performance of open-source models lags significantly behind that of API-based models and **most models still exhibit a large performance gap compared to humans**. For instance, the LLaMa-3-8B-Chat model achieved an overall score of **34.8**, significantly lower than the human performance of **91.6**.

4.4 In-Depth Analysis

Performance Differences in LLMs' ToM Capabilities Across Third-Person and First-Person

[3] https://openai.com/index/hello-gpt-4o/
[4] https://openai.com/index/learning-to-reason-with-llms/
[5] https://www.anthropic.com/news/claude-3-5-sonnet
[6] https://ai.meta.com/blog/meta-llama-3/

Methods	Cognitive Intelligence						
	Static Cognition			Dynamic Cognition-G0.8A			Dynamic Cogntion
	Third-person	First-person	Δ	Level 1	Level 2	Level 3	Limit Texas
Open-source Models							
LLaMa-3-8B-Chat	50.6	66.2	+15.6	0.0	0.0	0.0	48.0
LLaMa-3-70B-Chat	58.4	63.2	+4.8	10.0	20.0	10.0	38.0
LLaMa-3.1-405B-Instruct	58.0	65.8	+7.8	80.0	20.0	20.0	56.0
API-based Models							
Claude-3-5-Sonnet	71.0	80.5	+9.5	50.0	10.0	40.0	66.0
GPT-3.5-Turbo	45.5	51.9	+6.4	10.0	10.0	0.0	56.0
GPT-4-Turbo	55.4	69.7	+14.3	10.0	20.0	10.0	60.0
GPT-4o	64.1	71.0	+6.9	10.0	40.0	10.0	62.0
o1-preview	71.9	77.5	+5.6	90.0	90.0	90.0	72.0
Human							
Human Performance	90.2	90.2	0.0	90.0	86.0	73.0	82.0

Methods	Situational Intelligence			Behavioral Intelligence			AVG
	Parallel World	Counterfact	Real-World	Adversarial	Cooperative	Social Goal	
Open-source Models							
LLaMa-3-8B-Chat	6.7	71.0	67.2	51.3	49.7	22.5	34.8
LLaMa-3-70B-Chat	13.3	59.0	73.2	45.0	53.3	25.5	37.3
LLaMa-3.1-405B-Instruct	36.7	66.0	77.3	52.3	65.2	34.0	52.1
API-based Models							
Claude-3-5-Sonnet	90.0	74.0	79.8	55.0	94.8	50.5	62.8
GPT-3.5-Turbo	13.3	37.0	72.2	46.7	50.3	33.0	34.6
GPT-4-Turbo	23.3	70.0	75.7	54.7	75.6	52.0	47.4
GPT-4o	36.7	52.0	85.8	54.0	80.8	53.0	50.5
o1-preview	86.7	90.0	84.7	56.7	96.3	52.5	80.6
Human							
Human Performance	93.3	91.0	90.7	56.3	100.0	64.5	91.6

Table 2: Performance of cognitive, situational, and behavioral intelligence from first-person perspective of eight LLMs. Highest and second-highest scores among LLMs and humans in each scenario are highlighted in red and blue, respectively. **AVG** represents the average value of cognitive, situational, and behavioral intelligence performance.

Perspective As shown in Table 2, **all LLMs exhibited improved performance after the ToMI benchmark is converted from a third-person to a first-person perspective**. The Llama3-8B-Chat model achieved the largest improvement of **+15.6**. Notably, the claude and o1-preview models demonstrated significantly stronger ToM capabilities in the first-person perspective compared to other models. Except for GPT-3.5-Turbo, API-based models generally outperformed open-source models, including the recently released LLaMa-3.1-405B-Instruct. However, despite these improvements, there remains a substantial gap between the performance of all LLMs and human baselines.

The scaling up of open-source models has not yielded significant results By comparing the performance of LLaMa-3-8B-Chat with LLaMa-3-70B-Chat in Table 2, we observe that although the model size increased significantly, the overall performance on social intelligence improved by only **+2.5**. We further explore the scaling effects of increasing the size of the LLaMa-3 model on GSM8K (Cobbe et al., 2021) and MMLU (Chung et al., 2024) tasks, finding improvements of **+12.9** and **+13.4**, respectively, as illustrated in Figure 4.

The powerful mathematical capabilities of the o1-preview model are truly surprising In the dynamic cognition evolution-G0.8A scenario, almost all LLMs perform poorly, even in the simplest level 1 situation, which poses a significant challenge for humans as well. However, the recent o1-preview model has performed exceptionally well,

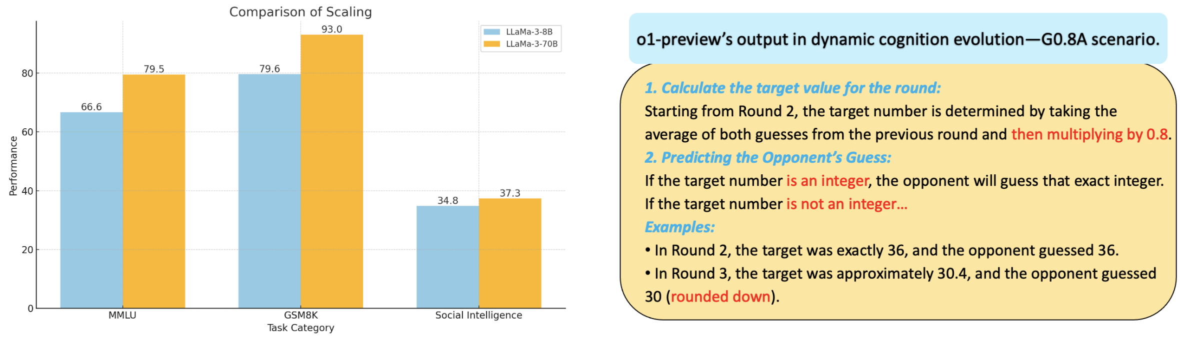

Figure 4: **Left**: **performance evolution** corresponding to **scaling up** LLaMa-3 model size across different task domains. **Right**: o1-preview model's output in dynamic cognition evolution—G0.8A scenario.

we analyze its outputs and find that it is highly sensitive to numbers and can **capture the correlations between numbers and the underlying patterns behind them**, as illustrated in Figure 4. Therefore, when **humans are unable to perceive these numerical patterns**, the o1-preview model, based on its powerful mathematical capabilities, perceives things that humans have not detected.

Mid-point Belief, Strange Guess and Get Back on Track As shown in Figure 5, in the scenario of **dynamic cognition G0.8A Level 2 (Arithmetic sequence)**, we thoroughly investigate the belief state evolution pattern of GPT-4-Turbo regarding the opponent's proposed numbers. In round 1, with no available information, the GPT-4-Turbo model thinks the opponent will choose the number 50 within the range of 1-100. The same phenomenon is observed in the GPT-3.5-Turbo model, called **"mid-point belief"**. Sometimes, the GPT-4-Turbo model continuously believes the opponent will choose progressively smaller numbers throughout the entire interaction, as depicted by the GPT-4-Turbo guess1 curve in Figure 5. Although this is very close to the gold number, it does not capture that the opponent's chosen numbers form an arithmetic sequence. Another situation occurs when the GPT-4-Turbo model makes a **"strange guess"** in the initial rounds, thinking the opponent will suddenly choose larger numbers. After several rounds, it captures that the opponent's chosen numbers form an arithmetic sequence, called **Get Back on Track**. Overall, despite the statistical results indicating that the GPT-4-Turbo model does not establish a belief regarding the Level 2 opponent in the G0.8A scenario, the phenomena we observed suggest that it has started to grasp some patterns. **The belief information for all models across all rounds** can be found in Appendix C.

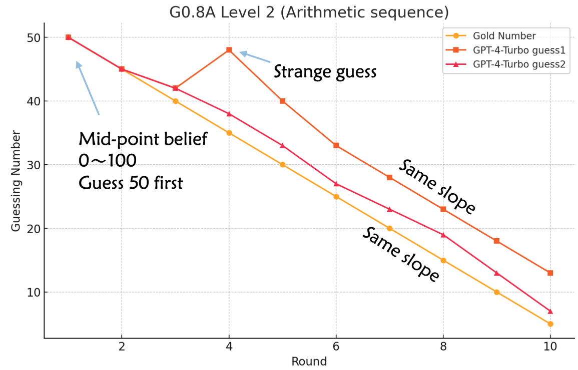

Figure 5: In the scenario of G0.8A Level 2 (Arithmetic sequence), the belief state evolution pattern of GPT-4-Turbo regarding the opponent's proposed numbers.

5 Conclusion

In this paper, considering the social intelligence of LLMs has yet to be systematically organized and studied, ego-centric first-person perspective evaluation can align well with actual LLM-based Agent use scenarios, incorporating human-machine interaction scenario evaluation is critical and the natural approach of observing and understanding the world from an ego-centric first-person perspective for both humans and LLM-based agents, we propose the EgoToMArena framework. This framework is grounded in the three pillars of social intelligence: cognitive, situational, and behavioral intelligence, with eight scenarios: static, dynamic cognition; real-world, counterfactual, parallel world situation; cooperative, adversarial game, and social goal-driven human-machine interactive environment, aimed to systematically evaluate the social intelligence of LLMs from a first-person perspective. We conduct extensive experiments and observe some key insights regarding the future development of LLMs as well as the capabilities levels of the most advanced LLMs currently available.

Limitations

There are three major limitations in our study. (1) Our study only involves the text modality and does not utilize ego-centric images and videos. The social intelligence of Vision-Language Models from a first-person perspective is very important, and we will leave this for future research. (2) Due to the constraint of computing resources and budget, we only evaluate eight prominent foundation LLMs, While we believe that the selected LLMs are representative. (3) Our study evaluates the social intelligence of LLMs from a first-person perspective, a deeper interpretation of these evaluation results from the perspective of explainability research would be more beneficial for the development of LLMs' social intelligence.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

John P Agapiou, Alexander Sasha Vezhnevets, Edgar A Duéñez-Guzmán, Jayd Matyas, Yiran Mao, Peter Sunehag, Raphael Köster, Udari Madhushani, Kavya Kopparapu, Ramona Comanescu, et al. 2022. Melting pot 2.0. *arXiv preprint arXiv:2211.13746*.

Federico Bianchi, Patrick John Chia, Mert Yuksekgonul, Jacopo Tagliabue, Dan Jurafsky, and James Zou. 2024. How well can llms negotiate? negotiationarena platform and analysis. *arXiv preprint arXiv:2402.05863*.

G Brockman. 2016. Openai gym. *arXiv preprint arXiv:1606.01540*.

Chunkit Chan, Cheng Jiayang, Yauwai Yim, Zheye Deng, Wei Fan, Haoran Li, Xin Liu, Hongming Zhang, Weiqi Wang, and Yangqiu Song. 2024. Negotiationtom: A benchmark for stress-testing machine theory of mind on negotiation surrounding. *arXiv preprint arXiv:2404.13627*.

Junzhe Chen, Xuming Hu, Shuodi Liu, Shiyu Huang, Wei-Wei Tu, Zhaofeng He, and Lijie Wen. 2024a. Llmarena: Assessing capabilities of large language models in dynamic multi-agent environments. *arXiv preprint arXiv:2402.16499*.

Zhuang Chen, Jincenzi Wu, Jinfeng Zhou, Bosi Wen, Guanqun Bi, Gongyao Jiang, Yaru Cao, Mengting Hu, Yunghwei Lai, Zexuan Xiong, et al. 2024b. Tombench: Benchmarking theory of mind in large language models. *arXiv preprint arXiv:2402.15052*.

Sijie Cheng, Zhicheng Guo, Jingwen Wu, Kechen Fang, Peng Li, Huaping Liu, and Yang Liu. 2023. Can vision-language models think from a first-person perspective? *arXiv preprint arXiv:2311.15596*.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Logan Cross, Violet Xiang, Agam Bhatia, Daniel LK Yamins, and Nick Haber. 2024. Hypothetical minds: Scaffolding theory of mind for multi-agent tasks with large language models. *arXiv preprint arXiv:2407.07086*.

Zi-Yi Dou, Xitong Yang, Tushar Nagarajan, Huiyu Wang, Jing Huang, Nanyun Peng, Kris Kitani, and Fu-Jen Chu. 2024. Unlocking exocentric video-language data for egocentric video representation learning. *arXiv preprint arXiv:2408.03567*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Caoyun Fan, Jindou Chen, Yaohui Jin, and Hao He. 2024. Can large language models serve as rational players in game theory? a systematic analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 17960–17967.

Yao Fu, Hao Peng, Tushar Khot, and Mirella Lapata. 2023. Improving language model negotiation with self-play and in-context learning from ai feedback. *arXiv preprint arXiv:2305.10142*.

Kanishk Gandhi, J-Philipp Fränken, Tobias Gerstenberg, and Noah D Goodman. 2023. Understanding social reasoning in language models with language models. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, pages 13518–13529.

Jiaxian Guo, Bo Yang, Paul Yoo, Bill Yuchen Lin, Yusuke Iwasawa, and Yutaka Matsuo. 2023. Suspicion-agent: Playing imperfect information games with theory of mind aware gpt-4. *arXiv preprint arXiv:2309.17277*.

Jiangyong Huang, Silong Yong, Xiaojian Ma, Xiongkun Linghu, Puhao Li, Yan Wang, Qing Li, Song-Chun Zhu, Baoxiong Jia, and Siyuan Huang. 2023. An embodied generalist agent in 3d world. *arXiv preprint arXiv:2311.12871*.

Yifei Huang, Guo Chen, Jilan Xu, Mingfang Zhang, Lijin Yang, Baoqi Pei, Hongjie Zhang, Lu Dong, Yali Wang, Limin Wang, et al. 2024. Egoexolearn: A dataset for bridging asynchronous ego-and exocentric view of procedural activities in real world. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22072–22086.

Thelma Hunt. 1928. The measurement of social intelligence. *Journal of Applied Psychology*, 12(3):317.

Hyunwoo Kim, Melanie Sclar, Xuhui Zhou, Ronan Bras, Gunhee Kim, Yejin Choi, and Maarten Sap. 2023. Fantom: A benchmark for stress-testing machine theory of mind in interactions. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14397–14413.

Rudolf Laine, Bilal Chughtai, Jan Betley, Kaivalya Hariharan, Jeremy Scheurer, Mikita Balesni, Marius Hobbhahn, Alexander Meinke, and Owain Evans. 2024. Me, myself, and ai: The situational awareness dataset (sad) for llms. *arXiv preprint arXiv:2407.04694*.

Matthew Le, Y-Lan Boureau, and Maximilian Nickel. 2019. Revisiting the evaluation of theory of mind through question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5872–5877.

Huao Li, Yu Chong, Simon Stepputtis, Joseph P Campbell, Dana Hughes, Charles Lewis, and Katia Sycara. 2023. Theory of mind for multi-agent collaboration via large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 180–192.

Minzhi Li, Weiyan Shi, Caleb Ziems, and Diyi Yang. 2024. Social intelligence data infrastructure: Structuring the present and navigating the future. *arXiv preprint arXiv:2403.14659*.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. 2015. Human-level control through deep reinforcement learning. *nature*, 518(7540):529–533.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, pages 1–22.

David Premack and Guy Woodruff. 1978. Does the chimpanzee have a theory of mind? *Behavioral and brain sciences*, 1(4):515–526.

Maarten Sap, Ronan Le Bras, Daniel Fried, and Yejin Choi. 2022. Neural theory-of-mind? on the limits of social intelligence in large lms. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3762–3780.

Edward L Thorndike. 1921. Intelligence and its measurement: A symposium–i. *Journal of Educational psychology*, 12(3):124.

Ruiyi Wang, Haofei Yu, Wenxin Zhang, Zhengyang Qi, Maarten Sap, Graham Neubig, Yonatan Bisk, and Hao Zhu. 2024. Sotopia-pi: Interactive learning of socially intelligent language agents. *arXiv preprint arXiv:2403.08715*.

Yufan Wu, Yinghui He, Yilin Jia, Rada Mihalcea, Yulong Chen, and Naihao Deng. 2023. Hi-tom: A benchmark for evaluating higher-order theory of mind reasoning in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10691–10706.

Hainiu Xu, Runcong Zhao, Lixing Zhu, Jinhua Du, and Yulan He. 2024. Opentom: A comprehensive benchmark for evaluating theory-of-mind reasoning capabilities of large language models. *arXiv preprint arXiv:2402.06044*.

Rongwu Xu, Brian S Lin, Shujian Yang, Tianqi Zhang, Weiyan Shi, Tianwei Zhang, Zhixuan Fang, Wei Xu, and Han Qiu. 2023. The earth is flat because...: Investigating llms' belief towards misinformation via persuasive conversation. *arXiv preprint arXiv:2312.09085*.

D Zha, KH Lai, Y Cao, S Huang, R Wei, J Guo, and X Rlcard Hu. 2019. A toolkit for reinforcement learning in card games. *arXiv preprint arXiv:1910.04376*.

Wenqi Zhang, Ke Tang, Hai Wu, Mengna Wang, Yongliang Shen, Guiyang Hou, Zeqi Tan, Peng Li, Yueting Zhuang, and Weiming Lu. 2024a. Agentpro: Learning to evolve via policy-level reflection and optimization. *arXiv preprint arXiv:2402.17574*.

Yadong Zhang, Shaoguang Mao, Tao Ge, Xun Wang, Yan Xia, Man Lan, and Furu Wei. 2024b. K-level reasoning with large language models. *arXiv preprint arXiv:2402.01521*.

Xuhui Zhou, Hao Zhu, Leena Mathur, Ruohong Zhang, Haofei Yu, Zhengyang Qi, Louis-Philippe Morency, Yonatan Bisk, Daniel Fried, Graham Neubig, et al. 2023. Sotopia: Interactive evaluation for social intelligence in language agents. *arXiv preprint arXiv:2310.11667*.

Caleb Ziems, Jane Dwivedi-Yu, Yi-Chia Wang, Alon Halevy, and Diyi Yang. 2023. Normbank: A knowledge bank of situational social norms. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7756–7776.

Appendix

A Related Works

Ego-centric (First-person Perspective) Research In the fields of computer vision and robotics, there has already been considerable research on a first-person perspective. For example, Cheng et al. (2023) explored whether vision-language models can "Think from a First-person Perspective?" Huang et al. (2023) proposes the construction of embodied agents in a 3D world, which involves acquiring and processing first-person perspective images. Huang et al. (2024) built a bridge between third-person and first-person perspectives at the action level, while Dou et al. (2024) proposed a method designed to transform exocentric video-language data for egocentric video representation learning. However, research on first-person perspectives in the field of natural language processing remains unexplored.

Datasets Related to Social Intelligence Sap et al. (2022) proposed SocialIQA and used it to evaluate LLMs. SocialIQA contains many questions related to social commonsense. Ziems et al. (2023) introduced NormBank, a large repository of social norms knowledge, which can be used to assess social norm-related tasks. Li et al. (2024) reorganized and classified existing datasets related to social intelligence. Xu et al. (2023) studied LLMs' understanding of the world and explored how different persuasion strategies could modify LLMs' worldviews.

Previous evaluations for the ToM of LLMs primarily focus on testing models using narrative stories, also referred to as reading comprehension scenarios. Specifically, Le et al. (2019) proposed the ToMi benchmark based on the classic Sally-Anne test. Wu et al. (2023) introduced the HI-ToM benchmark, which focuses on higher-order belief reasoning and sets up scenarios where agents can communicate with each other. Gandhi et al. (2023) proposed BigToM, which presents a framework for designing a ToM benchmark from synthetic templates for evaluating different aspects of LLMs' ToM capabilities. Xu et al. (2024) introduced OpenToM, which assigns personalities to agents in the stories and ensures that the storylines are more reasonable and logical. Chen et al. (2024b) proposed ToMBench, which systematically evaluates LLMs across all dimensions of ToM capabilities. Unlike the above methods that require LLMs to read stories and answer related questions, some studies evaluate LLMs' performance by inputting dialogues to them. Kim et al. (2023) proposed FanToM, which tests LLMs on their ability to infer the mental states of characters in everyday conversations. Chan et al. (2024) introduced NegotiationToM, which restricts the dialogue content to negotiation scenarios.

For the study of LLMs' behaviors and interaction capabilities, (Agapiou et al., 2022) proposed Melting 2.0, which encompasses various environments such as cooperation and gaming, originally designed for research in multi-agent reinforcement learning. (Zhou et al., 2023) introduced an interactive dialogue environment for large language models under a social goal-driven framework. (Chen et al., 2024a) proposed a game-like environment where different LLMs are paired for competitive interactions.

Strategy Enhancement in Interactive Scenarios Some work focuses on designing interaction strategies to enable LLMs to gain more benefits during interactions. For example, Zhang et al. (2024a) proposed Agent-pro, Zhang et al. (2024b) introduced K-level reasoning, and Guo et al. (2023) put forward the Suspicion-Agent. Additionally, Li et al. (2023) explored Multi-LLM collaboration by informing LLMs of task rules through prompts. Park et al. (2023) introduced generative agents that can simulate human behavior. Bianchi et al. (2024) explored the social behavior of LLMs in negotiation scenarios. Fu et al. (2023) show LLMs can improve each other in a negotiation scenario. Fan et al. (2024) examined the capability of LLMs to make rational decisions in game theoretic scenarios.

B Case——Limit Texas Hold'em

As illustrated in Figure 6.

C Belief Dynamic Evolution in G0.8A Scenario

The following three tables correspond to the dynamic evolution data of beliefs for various LLMs under Cognitive Levels 3, 2, and 1.

[{'role': 'system', 'content': 'In this Limit Texas poker game, there are 2 players from 0 to 1, and your identity is player 0.'}, {'role': 'user', 'content': "Now your hand is ['HK', 'H2'], and the community cards is []. The number of chips all players have invested is [1, 2]. the actions you can choose are ['call', 'raise', 'fold']. Currently, Preflop. Please provide your results in the form of {'action': ''}. You must choose one from ['call', 'raise', 'fold'] as your answer. Just output the dictionary, don't use any other text."}]

[{'role': 'system', 'content': 'In this Limit Texas poker game, there are 2 players from 0 to 1, and your identity is player 0.'}, {'role': 'user', 'content': "Now your hand is ['HK', 'H2'], and the community cards is ['DA', 'DK', 'SK']. The number of chips all players have invested is [4, 4]. the actions you can choose are ['raise', 'fold', 'check']. Currently, Preflop. Player 0 raises. Player 1 calls. Flop. Please provide your results in the form of {'action': ''}. You must choose one from ['raise', 'fold', 'check'] as your answer. Just output the dictionary, don't use any other text."}]

[{'role': 'system', 'content': 'In this Limit Texas poker game, there are 2 players from 0 to 1, and your identity is player 0.'}, {'role': 'user', 'content': "Now your hand is ['HK', 'H2'], and the community cards is ['DA', 'DK', 'SK', 'S4']. The number of chips all players have invested is [8, 16]. the actions you can choose are ['raise', 'fold', 'check']. Currently, Preflop. Player 0 raises. Player 1 calls. Flop. Player 0 raises. Player 1 raises. Turn. Please provide your results in the form of {'action': ''}. You must choose one from ['raise', 'fold', 'check'] as your answer. Just output the dictionary, don't use any other text."}]

[{'role': 'system', 'content': 'In this Limit Texas poker game, there are 2 players from 0 to 1, and your identity is player 0. '}, {'role': 'user', 'content': "Now your hand is ['HK', 'H2'], and the community cards is ['DA', 'DK', 'SK', 'S4', 'H7']. The number of chips all players have invested is [32, 32]. the actions you can choose are ['raise', 'fold', 'check']. Currently, Preflop. Player 0 raises. Player 1 calls. Flop. Player 0 raises. Player 1 raises. Turn. Player 0 raises. Player 1 calls. River. Please provide your results in the form of {'action': ''}. You must choose one from ['raise', 'fold', 'check'] as your answer. Just output the dictionary, don't use any other text."}]

Question: Your opponent's hand is ['CQ', 'S9']. Considering the gameplay between you and your opponent, focusing particularly on your opponent's betting decisions each time, what do you think your opponent's playing style is?

Answer: (A) Aggressive (B) Conservative

Figure 6: A Case for Limit Texas Hold'em.

Model	Round 1	Round 2	Round 3	Round 4	Round 5	Round 6	Round 7	Round 8	Round 9	Round 10	Accuracy
GPT-4-Turbo	50 ✓	45	40	35	30	25	22	17	15	13	0.1
GPT-3.5-Turbo	40	20	60	55	70	90	60	45	75	85	0
GPT-4o	50 ✓	40	30	20	15	10	8	6	5	4	0.1
o1-preview	1	20 ✓	✓	✓	✓	✓	✓	✓	✓	✓	0.9
Claude-3-5-Sonnet-20240620	65	45	35	28	20 ✓	17	14	10 ✓	7.5 ✓	5.6 ✓	0.4
Llama3-8b-chat-hf	67	67	67	67	67	67	67	67	67	67	0
Llama3-70b-chat-hf	50 ✓	45	43	30	25	19	15	12	11	7	0.1
Llama3.1-405b-Instruct-Turbo	50 ✓	40 ✓	35	29	23	19	14.5	11.5	9.5	7.5	0.2

Model	Round 1	Round 2	Round 3	Round 4	Round 5	Round 6	Round 7	Round 8	Round 9	Round 10	Accuracy
GPT-4-Turbo	50 ✓	45✓	48	42	36	33	28	22	18	12	0.2
GPT-3.5-Turbo	40	20	60	35✓	70	50	45	60	45	40	0.1
GPT-4o	50✓	40	40 ✓	30	25	20	15	10	10 ✓	5 ✓	0.4
o1-preview	1	✓	✓	✓	✓	✓	✓	✓	✓	✓	0.9
Claude-3-5-Sonnet-20240620	65	45✓	35	25	20	15	12	8	5	8	0.1
Llama3-8b-chat-hf	67	67	67	67	67	67	67	67	67	67	0
Llama3-70b-chat-hf	50✓	45✓	38	32	28	24	21	19	16	11	0.1
Llama3.1-405b-Instruct-Turbo	50✓	40	35	30	28	25✓	22	18	15	10	0.2

Model	Round 1	Round 2	Round 3	Round 4	Round 5	Round 6	Round 7	Round 8	Round 9	Round 10	Accuracy
GPT-4-Turbo	50✓	45	48	47	48	49	48	47	46	45	0.1
GPT-3.5-Turbo	40	35	70	30	80	40	55	60	50	30	0.1
GPT-4o	50✓	40	30	40	35	45	45	45	45	45	0.1
o1-preview	1	✓	✓	✓	✓	✓	✓	✓	✓	✓	0.9
Claude-3-5-Sonnet-20240620	65	45	35	25	20	50✓	50✓	50✓	50✓	50✓	0.5
Llama3-8b-chat-hf	67	67	67	67	67	67	67	67	67	67	0
Llama3-70b-chat-hf	50✓	48	52	53	54	55	54	56	57	58	0.1
Llama3.1-405b-Instruct-Turbo	50✓	33	45	50✓	50✓	50✓	50✓	50✓	50✓	50✓	0.8