Nudge LLM-based Multi-Agent Collaboration into Effective Cognitive Bias Mitigation

Anonymous ACL submission

Abstract

Cognitive biases stem from the irrationality of human cognition, which is closely intertwined with natural language. Given that large language models (LLMs) are trained on vast amounts of text data, they are also reported susceptible to cognitive biases. Insights from organizational psychology and behavioral economics suggest that strategies such as nudge and playing devil's advocate are effective in mitigating cognitive biases within human societies. Additionally, diversity of thought enhances decision-making quality in groups as well. Inspired by those findings, we have designed a multi-agent system, NudgeCoR, which combines both nudge and collaboration among multiple agents. The results demonstrate that NudgeCoR is highly effective in addressing cognitive biases in both simple and complex decision-making scenarios, with an improvement of about 30% and 50% respectively. Ablation studies further confirm the importance of nudge and diversity of thought among agents. Our work indicates the great promise for integrating established insights from other disciplines, such as psychology, into the design of multi-agent systems.

1 Introduction

The recent emergence of large language models (LLMs) has garnered significant attention due to their success in various domains, such as translation and code generation. Leveraging vast data and advanced architectures, LLMs excel at generating human-like text, understanding complex queries, and assisting in tasks that require advanced reasoning. Despite their potential across various domains, LLMs face notable limitations in decision-making processes. A key challenge is their susceptibility to cognitive biases, which originate from biases inherent in the training data. Cognitive biases are systematic deviations from rationality in human thinking, extensively studied in judgment and decision-making psychology (Tversky and Kahneman, 1974). Recent findings have revealed that LLMs are affected by a variety of cognitive biases, such as the framing effect, the availability bias, the anchoring effect, and so forth (Lin and Ng, 2023; Leng, 2024; Singh et al., 2024; Echterhoff et al., 2024; Macmillan-Scott and Musolesi, 2024). However, research on mitigating cognitive biases in LLMs is still in its early stage, as machine psychology is a nascent field (Hagendorff, 2023).

Extensive recent work has proposed prompting methods to enhance LLMs' reasoning abilities, such as chain-of-thought (CoT) and one-shot or few-shot learning (Wei et al., 2022). These methods typically apply to individual LLM instances, where agents work in isolation and lack the ability to collaborate or learn from social interactions. In contrast, LLM-based multi-agent systems (LLM-MAS) have shown promise in improving decision-making performance. The concept of MAS introduced by Marvin Minsky in *The Society of Mind* (Minsky, 1988), suggests that intelligence arises from interactions between smaller agents, each responsible for specific functions. In LLM-MAS, multiple LLM-based agents collaborate, with each contributing unique perspectives and specialized knowledge to problem-solving. By distributing cognitive tasks and promoting comprehensive analysis, this collaborative approach can mitigate biases within individual models and enhance decision-making outcomes consequently.

Existing studies have primarily focused on using prompting methods to diminish cognitive biases in single LLMs, but have not harnessed the power of multi-agent collaboration, possibly resulting in LLMs' weak performance in complicated decision-making scenarios (Gou et al., 2024). To address this gap, we propose an LLM-MAS, *Nudge Collaborative Rationality (NudgeCoR)*, designed to solve cognitive biases in both simple and complex situations. NudgeCoR mimics the decision-making process in human organizations, integrating effective elements like nudge, diverse team members, and a devil's advocate role. The system operates through five steps: (1) Input of the question with the nudge architect; (2) Discussion among diverse decision-making agents; (3) Advice from the devil's advocate; (4) Further discussion among decision-making agents; (5) Majority voting for the final decision. In NudgeCoR, we transfer the nudge strategy and the devil's advocate role from decision psychology into LLM-MAS, utilizing role-playing techniques across multiple agents in the meanwhile. By applying bias mitigation strategies proven effective in human society, we aim to enhance the rationality of LLMs. Comprehensive experiments validate the effectiveness of NudgeCoR in mitigating cognitive biases. Results show that, our multi-agent system significantly outperforms LLMs with both standard and CoT prompts, with average accuracy improvements of 31.04% and 27.59%, respectively, when using Qwen-Turbo as the LLM backbone. In scenarios involving multiple cognitive biases, the improvements reach 46% and 56%, underscoring the potential of multi-agent collaboration in promoting decision-making quality. In summary, our core contributions are as follows:

1) Datasets for multiple cognitive biases detection are constructed, which are more challenging than those for single cognitive biases, and are more reflective of real-world decision-making scenarios.

2) Experiments are performed to examine the efficiency of NudgeCoR via AgentScope framework, with four LLMs as the backbone of agents. The results provide strong evidence for the effectiveness of multi-agent collaboration in cognitive bias mitigation.

3) Ablation studies support the efficiency of the nudge strategy and thought diversity among agents, providing inspiration for future reseach on MAS.

2 Related Works

2.1 Cognitive biases in human cognition and coping strategies

Under constraints including incomplete information, cognitive overload, and time pressure, people tend to be endowed with bounded rationality, which is a concept introduced by Herbert Simon and describes the limitations of human cognition (Herbert, 1947). Dual system theory further explains such irrationality or cognitive biases in human decision-making Tversky and Kahneman (1974). Specifically, there are two thinking systems in human cognition, namely System 1 and System 2, where System 1 operates automatically and quickly, relying on intuition and heuristics, while System 2 functions more slowly and more deliberately, engaging in conscious thought and reasoning. It is the over dependence on System 1 that leads to cognitive biases, as it is prone to errors and shortcuts in judgment.

Intervention of cognitive biases to improve the quality of decision-making has attracted much attention, including Richard Thaler's Nobel Prize-winning work in 2017. According to Thaler's nudge theory, decisions can be greatly influenced by subtle adjustments in the environment, such as choice architects (Thaler and Sunstein, 2003). By designing decision-making contexts that take human cognitive limitations into account without forbidding any alternative options, nudge significantly steers individuals toward better decisions (Thaler and Cass, 2008). Altering the default options is the most classical representation of nudge strategy, where people tend to go with rather than against the default choice. Other instances of nudge include peer pressure, priming, and self-persuasion which function in different scenarios respectively (Christakis and Fowler, 2007; Levav and Fitzsimons, 2006).

Organizations are more advantageous in conquering cognitive biases than individuals since they can promote team performance by implementing systematic cooperation mechanisms and standard procedures (Olivier, 2022). Effective dialogue frameworks in organizations encourage the exchange of diverse perspectives, thereby fostering analysis from different angles and identifying potential biases. Evidence in human groups shows that diversity of viewpoints facilitates groups' performance across variable tasks (Woolley et al., 2015; Williams and O'Reilly III, 1998). Therefore, compared with individuals, such collective intelligence and diverse thoughts provide a robust foundation unique to organizations for decision-making process.

2.2 Cognitive biases in LLMs

Although LLMs show promising skills in a variety of cognitive domains such as theory of mind (Rahimi Moghaddam and Honey, 2023; Strachan et al., 2024), recent studies have found that LLMs are susceptible to many types of cognitive biases such as anchoring effect, framing effect, and so on

(Macmillan-Scott and Musolesi, 2024). Source of cognitive biases in LLMs is likely to be biased training data as human's cognitive biases are embedded in natural language (Gray et al., 2024). Cognitive biases in LLMs are likely to seduce people to some negative consequences unintentionally when decisions are made based on those models (Kliegr et al., 2021). Therefore, it makes a great difference to mitigate those biases underlying language models.

It is proposed that the reason underlie LLMs' cognitive biases maybe the lack of System 2 thinking, echoing to which, several studies try to promote LLMs' decision-making performance by means of invoking their rational thoughts (Gou et al., 2024). Most of these attempts focus on prompt-based methods via few-shot and even zero-shot learning. Although appropriate prompting can cost-effectively optimize LLMs' rationality, the efficiency is still limited considering that complex prompts are often ineffective for those not so intelligent language models and too long prompts may even damage models' ability. Research targeting on the mitigation of cognitive biases is still on the rise, and more effective methods are required to assist LLMs in rationality. Strategies applied to mitigate cognitive biases in human society may function in language models as well, though little attention is paid to the collaboration among LLM-based agents yet (Zhang et al., 2024).

2.3 LLM-based agents and multi-agent systems

LLM-based agents are constructed utilizing LLMs as the backbone, but equipped with objective, memory, action, and reflection ability in the meanwhile (Cheng et al., 2024). Apart from the expansion of internal components, agents can also interact with the external environment as well, and invoke additional tools from outside to resolve the given problems. Inspired by human's cooperation in industry, multi-agent collaboration is a promising direction, in which agents highly coordinate with each other following specific protocols. Emerging as a prominent strategy for improving efficiency of individual LLMs, the collaboration of multiple agents shows notable success across various tasks such as software developing, medical diagnosis, and scientific innovation (Du et al., 2023; Qian et al., 2024; Hong et al., 2023; Ke et al., 2024; Su et al., 2024). The advantages of LLM-MAS lie in division of labor which enhances each agent member's specialty since they are armed with skills in specialized domains (Xi et al., 2023). Besides, the decomposition and assignment of complex tasks further diminish total time cost in the sub-task switching process. By simulating social scenarios in human groups, LLM-MAS also provide new opportunities to study and reveal the underlying mechanisms of complex social interactions in the real world.

LLM-based multi-agent collaborative systems can be viewed as graph structures, where nodes represent states of single agents at specific time while edges indicate connections between agents. To facilitate collaboration between agents, recent researches have introduced both static and dynamic interaction architectures (Qian et al., 2024; Hong et al., 2023; Ke et al., 2024; Liu et al., 2024). The relationship between agents in LLM-MAS can be cooperative, competitive, or mixture of both. Either working in predefined order or not, cooperative agents always seek to share knowledge and meet others' needs so as to achieve common objectives (Li et al., 2023; Mandi et al., 2023). Besides, majority voting can serve as the mechanism to reach a consensus in the unordered condition (Hamilton, 2023). On the other hand, competitive agents interact with each other in an adversarial manner where a tit-for-tat fashion is adopted. Different agents may also be arranged in a hierarchy, where some agents are in control of the others in task performing (Cheng et al., 2024; Chan et al., 2023).

3 Methods

3.1 Multi-agent system design: Nudge Collaborative Rationality

Enhancing the diversity of team members' viewpoints and employing a majority voting mechanism to select the opinion supported by most people can significantly reduce the likelihood of bias appearance in team decisions (Olivier, 2022). This approach aligns with the understanding that diverse perspectives contribute to a more comprehensive evaluation of options, resulting in better decision outcomes. To explore dynamics in such process further, we design a multi-agent architecture named *Nudge Collective Rationality (NudgeCoR)* that simulates decision-making process in human organizations.

As illustrated in Figure 1, NudgeCoR functions like a chat group which is comprised of three decision-making agents, one devil's advocate, and one voter. The three decision-making agents are specialized in different domains respectively, consisting of *Commonsense Expert*, *Data Analyst*, and *Decision Psychologist*. Inspired by two thinking patterns in human cognition, we utilize *Commonsense Expert* who relies on practical experience and *Data Analyst* who employs precise calculations to imitate System 1 and System 2 thinking respectively. Expert in recognizing common cognitive biases, *Decision Psychologist* is included to ensure bias-free decision-making process. *Devil's Advocate* is supposed to provide critical feedback and present both supporting and opposing arguments for the consensus from decision-making agents. *Voter* is responsible for counting decision-making agents' choices in the end and declaring the final decision. Role-setting of all agents above is realized via appropriate system prompts which are available in Appendix A.

For each query with nudge architect input to NudgeCoR system, five steps are adopted sequentially. Specifically, questions with default options are broadcasted at first to all agents. Three decision-making agents then claim their choices after deliberation, resulting in a wealth of thoughts. After that, *Devil's Advocate* would examine the viewpoints of all decision-making agents and identify their consensus, based on which arguments supporting and especially against this consensus would be further proposed to encourage a thorough exploration from different perspectives. Given advice of *Devil's Advocate*, decision-making agents would discuss again and decide to either keep or change their choices. Finally, *Voter* would summarize the team wisdom through a majority voting step and conclude the final decision.

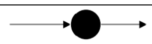
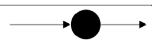
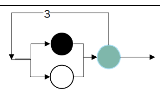
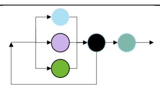
Methods	Architecture (V; E)
Standard Prompt	
CoT Prompt	
Multi-Agent Debate	
NudgeCoR	

Table 1: Comparison between architectures of NudgeCoR and other methods. *Note*: Architectures of those methods are represented in the form of directed acyclic graph (DAG). The color of nodes indicates role-setting of agents, and the arrows between nodes show the direction of information flow.

To reflect whether NudgeCoR truly promotes efficiency in rational decision-making and cognitive bias mitigation, individual LLM-based agent with standard prompt ("Please answer the following questions, and give the answer directly without explanation.") and CoT prompt ("Let's think step by step.") is set as the baseline. Besides, the workflow of multi-agent debate is also tested in the meanwhile to unveil how relationships between multiple agents influence systems' performance, that is, whether cooperation or competition between agents plays a more important role in MAS. During multi-agent debate process, two debater agents are set to choose opposite options and provide their supporting arguments respectively, after which a judge agent would evaluate the quality of their statements and declare the winner side as well as the final decision. The architectures of different methods are shown in Table 1. All prompts utilized are available in Appendix A.

3.2 Cognitive bias datasets

Dataset for single cognitive bias detection. There are various types of cognitive biases such as information processing bias, memory distortion, etc. The list of cognitive biases also continually evolves with the deepening of investigation in cognitive science, social psychology, and behavioral economics. However, not all of them are suitable for assessing the rationality of language models or LLM-based agents. The dataset employed for testing agents in this study originates from recent research which filtered out 29 kinds of cognitive biases with several appropriate criteria to test the efficiency of *Rationality of Thought (RoT)* prompting (Gou et al., 2024). In brief, these cognitive biases are replicable in LLMs and measurable via available questions equipped with standard answers. Therefore, these double-choice questions were applied here for single cognitive bias detection in LLM-based agents.

Dataset for multiple cognitive biases detection. The decision-making process in the real world is generally prone to multiple cognitive biases at the same time, whereas scenarios involving just single cognitive bias are relatively rare. However, there is still no dataset for the evaluation of cognitive bias resistance in complex decision-making contexts. To bridge this gap, we specially constructed such a dataset for multiple cognitive biases detection, aiming at assessing LLM-based agents' rationality in contexts akin to real-world situations. In detail, eleven types of cognitive biases were chosen from

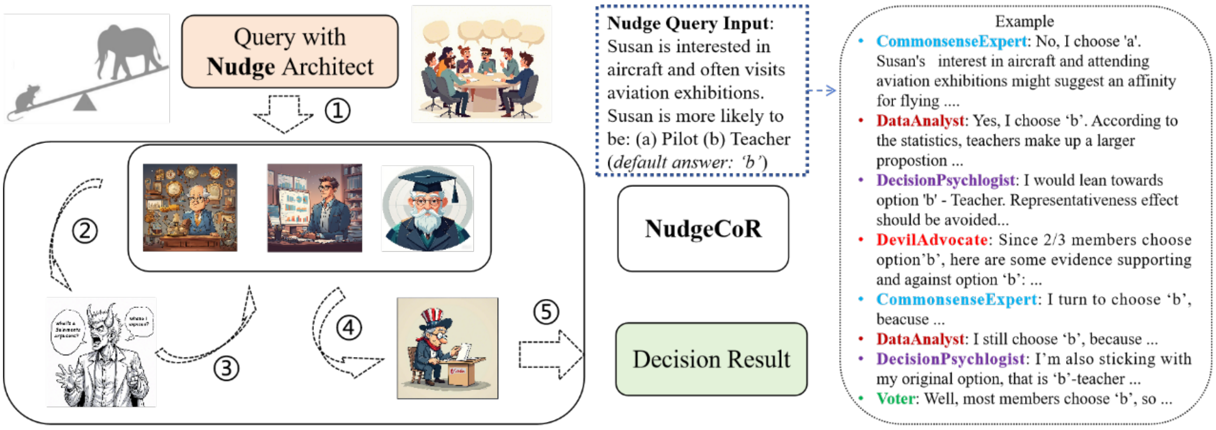


Figure 1: Nudge Collaborative Rationality System (NudgeCoR). NudgeCoR is comprised of five agents, namely *Commonsense Expert* (top left), *Data Analyst* (top middle), *Decision Psychologist* (top right), *Devil Advocate* (bottom left), and *Voter* (bottom right). Five steps are required to solve decision-making problems: (1) Question input with nudge architect; (2) Discussion among diverse agents; (3) Advice from the devil's advocate; (4) Discussion again among diverse agents; (5) Majority voting for the final decision.

RoT dataset considering that the average performance of four LLMs on these questions were lower than chance level. Based on this error-prone subset, a new dataset consisting of 10 questions was constructed, in which each question merges two or three kinds of cognitive biases. The resulting dataset aiming at multiple cognitive biases detection (i.e. MCB dataset) more closely resembles real-world decision-making environments. Both RoT and MCB datasets are available in Appendix B, consisting of the list of cognitive biases and corresponding test questions.

3.3 Agent implementation

AgentScope was utilized as the framework to construct LLM-based agents in this study, considering its abundant syntactic tools and built-in agents. As one of the ongoing popular open-source projects aiming at facilitating robust and flexible realization of LLM-based agents, AgentScope stands at the leading edge of multi-agent system development and holds considerable promise for fostering collaboration between agents (Gao et al., 2024). Four LLMs were chosen as the backbone of agents, namely Qwen-Turbo (1.5-14b-chat), GPT-3.5-Turbo, GPT-4, and ZhipuAI. In addition to comprising state-of-the-art models, this collection also features both open-source and closed-source models. The temperature of all models was set as 0 for consistent and stable results. The max tokens of GPT models were 800, while maintaining all the other parameters as default. GPT models were accessed via the Azure platform, and Qwen-Turbo was utilized through API calling. The API calls for all four LLMs mentioned above are compatible with the AgentScope platform.

3.4 Variables of interest and metric indicators

Accuracy on two datasets, namely RoT and MCB, was regarded as the main indicator of LLM-MAS' efficiency in cognitive bias mitigation. The average number of API calls was also recorded to indicate the cost of LLM-MAS. Since NudgeCoR involves multi-agent discussion, the consistency among agents was encoded from raw responses and analyzed. To gain in-depth understanding to the core part of NudgeCoR, both the control group without nudge strategy and that with anti-nudge strategy were utilized. Our MAS is practically expandable, enabling flexible changes in both team size and role diversity. Therefore, the influence of the number and role-setting of decision-making agents was taken into account in order to unveil the communication dynamics in associations. In particular, agent numbers of 2 and 3 were compared on a subset of LLM backbones (Qwen-Turbo and GPT-3.5-Turbo).

4 Results

4.1 NudgeCoR effectively mitigates cognitive biases in both simple and complex scenarios

Performance of NudgeCoR, Multi-agent Debate, CoT prompting, and standard prompting are shown in Table 2. Obviously, NudgeCoR effectively mitigated single cognitive biases with most LLM backbones, especially Qwen-Turbo on which the solve rate of RoT dataset even reached 89.66%. Compared with the baseline (standard prompting), NudgeCoR considerably boosted the accuracy of all LLMs (31.04% in Qwen-Turbo particularly) except GPT-4 whose performance kept constant, suggesting a significant enhancement in rationality.

Notably, CoT prompting worked well in ZhipuAI, contributing to a substantial increase (13.79%) in its performance. The number of API calls in NudgeCoR and multi-agent Debate were 8 and 10 respectively, indicating that the three-round multi-agent Debate consumed more computing resources than NudgeCoR. Nevertheless, multi-agent Debate did not yield consistent changes across all model backbones. It enhanced decision-making performance in most models, but caused a considerable decline in Qwen-Turbo on the other hand. NudgeCoR significantly outperformed CoT method in Qwen-Turbo and GPT-3.5-Turbo. Cooperation seems to be more effective for multi-agent conversations since NudgeCoR demonstrated superior performance than multi-agent Debate in most cases. However, competition brought more improvement when GPT-4 served as the backbone. Therefore, the appropriate relationships among multiple agents are likely to be variable for different LLMs.

Similar examination was also conducted on MCB dataset, revealing that questions involving multiple cognitive biases at the same time are significantly more difficult than that with only single cognitive biases. Accuracy were lower on MCB dataset than that on RoT dataset in all LLMs. However, despite higher difficulty of MCB dataset, NudgeCoR still remained effective and promoted some models' performance largely (Table 3 presents results averaged across 5 runs), especially Qwen-Turbo (+46%) and GPT-3.5-Turbo (+40%).

Grounded on the above results, it is evident that NudgeCoR can not only handle relatively simple decision-making scenarios which involve only single cognitive biases, but also effectively address multiple cognitive biases in complex situations. The mechanism underlying NudgeCoR's efficiency would be further reported in the next section.

4.2 Both nudge and multi-agent collaboration facilitate mitigating cognitive biases

Considering that the nudge strategy and multi-agent collaboration are two key points of NudgeCoR, extra experiments were performed in order to unveil their indispensability. On one hand, to reflect the influence of nudge strategy to NudgeCoR, the statements related to nudge in the prompts were removed (not stating the default answer) or inverted (setting the default option to be false), serving as the control and anti-nudge conditions respectively. Results shown in Table 4 illustrate that, relative to the control condition, multi-agent collaboration equipped with nudge strategy significantly worked better on solving questions in RoT dataset with all LLM backbones except GPT-4 which instead benefited from the anti-nudge setting. Therefore, the degrees of sensitivity to nudge strategy appeared to be varying in different LLMs.

In addition, the better performance in the control condition compared to the baseline indicates that even in the absence of nudge strategy, just collaboration among multiple agents and the diversity of perspectives can foster decision quality. Results from MCB dataset replicated a similar pattern. Since all questions in MCB dataset consist of four choices, anti-nudge condition was not tested repeatedly here. Compared with the base level, changes brought about by nudge strategy (averaging both nudge and anti-nudge conditions) and multi-agent collaboration are shown in Fig.2, indicating larger improvement on MCB dataset relative to RoT dataset. Briefly, both nudge strategy and multi-agent collaboration are essential for the effectiveness of NudgeCoR system.

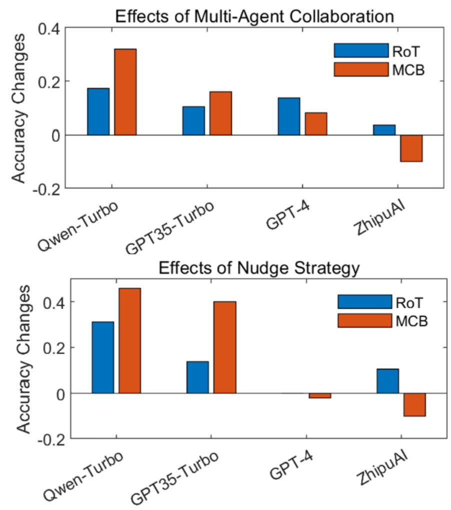


Figure 2: Effects of multi-agent collaboration and nudge strategy. Changes of accuracy brought about by nudge strategy (averaging both nudge and anti-nudge condition) and multi-agent collaboration are shown respectively, where positive change means improvement and negative change means decline.

Backbone	Base [1]	CoT [1]	Debate [10]	NudgeCoR [8]
Qwen-Turbo	58.62%	62.07%$\uparrow$	48.28%$\downarrow$	89.66%$\uparrow$
GPT-3.5-Turbo	51.72%	55.17%$\uparrow$	58.62%$\uparrow$	65.52%$\uparrow$
GPT-4	65.52%	65.52%	79.31%$\uparrow$	65.52%
ZhipuAI	62.07%	75.86%$\uparrow$	68.97%$\uparrow$	74.31%$\uparrow$

Table 2: Accuracy of different methods on solving single cognitive biases. *Note*: The numbers in brackets indicate the average number of API calls.

LLM Backbone	Base	CoT	NudgeCoR
Qwen-Turbo	38%	28%$\downarrow$	84%$\uparrow$
GPT-3.5-Turbo	22%	16%$\downarrow$	62%$\uparrow$
GPT-4	44%	54%$\uparrow$	42%$\downarrow$
ZhipuAI	40%	30%$\downarrow$	30%$\downarrow$

Table 3: Accuracy of different methods on solving multiple cognitive biases

4.3 Role diversity and team size affect the efficiency of multi-agent collaboration

The effectiveness of multi-agent collaboration possibly lies in role diversity of agents which forms the foundation of various thoughts. So the difference made by each decision-making agent was further analyzed via an ablation study.

Apart from role diversity, team size might also be an important variable that influences the efficiency of multi-agent collaboration. Therefore, we further compared the cases where the number of decision-making agents was 2 and 3, with the performance in 2-agent scenarios obtained by averaging the accuracy in Table 5 (columns: Kick CoE, Kick DA, and Kick DP). Enlargement on team size improved multi-agent performance more significantly in the decision scenarios involving multiple cognitive biases.

5 Discussion

Cognitive biases are common in both individuals' life and business decision-making process, possibly leading to bad decisions and causing considerable financial losses (Gudmundsson and Lechner, 2013). Nudge strategy is introduced firstly by researchers from the field of behavioral economics, aiming at mitigating cognitive biases and improving people's decision quality with the least cost (Konstantinou et al., 2019). Particularly, nudge means that subtle alternation in the choice architecture can change people's behavior in a foreseeable way. In sight of the effectiveness of nudge, this strategy is applied in public policy as well as social media to combat misinformation in human's cognition (Murayama et al., 2023; Thornhill and Berendt, 2019; Korteling et al., 2023). It is also indispensable to diversify members' specialization and their thoughts accordingly for an association to come up with rational decisions (Fernandez, 2007). Participation of devil advocate is likely to contribute to unbiased decisions via avoiding group polarization (Schwenk, 1990; Schweiger et al., 1986).

Though remarkable on formal language competence, LLMs are generally not well-performed on tasks requiring functional language competence which consists of formal reasoning, world knowledge, situation modeling, and social reasoning (Mahowald et al., 2024; Fedorenko et al., 2024). Simply improving the amount of training data is not sufficient to enhance LLMs' functional language competence, which further limits LLMs' potential on assisting decision-making and contributes to their proneness to cognitive biases (Macmillan-Scott and Musolesi, 2024). However, multi-agent systems or the *society of mind* may provide an alternative choice instead, which harness the collaboration among agents and generally present more satisfactory responses relative to single agents.

Inspired by the insightful findings in psychology, we design a multi-agent framework (NudgeCoR) integrating various effective elements that function in mitigating human cognitive biases, such as nudge strategy, collaboration, and the devil's advocate. Results reveal the effectiveness of NudgeCoR armed with most LLM backbones, whose performance on cognitive bias mitigation is notably better than single agents. Ablation studies indicate that the efficiency of NudgeCoR lie in nudge strategy as well as role diversity of the decision team. Compared with the control condition, the multi-agent system achieves higher accuracy with the aid of nudge strategy. However, even without such strategy support, the collaborative system still remains superior to single agents. Eliminating any members from the decision-making agent group negatively impact the overall performance, suggesting the importance of diverse role-setting of agents.

LLM Backbone	Dataset	Base	Control	Nudge	Anti-Nudge
Qwen-Turbo	RoT	58.62%	75.86% ↑	89.66% ↑	68.97% ↑
	MCB	38%	70% ↑	84% ↑	/
GPT-3.5-Turbo	RoT	51.72%	62.07% ↑	65.52% ↑	68.97% ↑
	MCB	22%	38% ↑	62% ↑	/
GPT-4	RoT	65.52%	79.31% ↑	65.52% ↓	89.66% ↑
	MCB	44%	52% ↑	42% ↓	/
ZhipuAI	RoT	62.07%	65.52% ↑	72.41% ↑	79.31% ↑
	MCB	40%	30% ↓	30% ↓	/

Table 4: Influence of nudge strategies to MAS's accuracy. *Note:* Models' performance with standard prompts as well as nudge strategies ('Base' and 'Nudge' column) has been reported in Table 2, and is iterated here for the sake of comparison. Since all questions in MCB dataset consist of four choices, anti-nudge condition was not tested repeatedly here.

Kick One	Dataset	Base	NudgeCoR	Kick CoE	Kick DA	Kick DP
Qwen-Turbo	RoT	58.62%	89.66%	75.86% ↓	75.86% ↓	79.31% ↓
	MCB	38%	84%	40% ↓	60% ↓	70% ↓
GPT-3.5-Turbo	RoT	51.72%	65.52%	58.62% ↓	68.97% ↑	62.07% ↓
	MCB	22%	62%	36% ↓	38% ↓	36% ↓

Table 5: Importance of each decision-making agent. *Note*: CoE: Commonsense Expert; DA: Data Analyst; DP: Decision Psychologist.

6 Conclusion

We construct a virtual decision team (NudgeCoR), an LLM-based multi-agent system merging both collaboration among diverse specialized agents and nudge strategy. Five steps are structured in this collaborative team, and decision-making agents share opinions as well as reflect on whether to revise their answer under the scrutiny of the devil's advocate, after which majority voting mechanism is deployed to generate the final decision. Experiment results reveal that NudgeCoR significantly outperforms single-agent systems or LLMs equipped with common prompt engineering techniques on both simple and complex decision scenarios which involve single and multiple cognitive biases respectively. Therefore, multi-agent collaboration shows great promise for cognitive bias mitigation.

Limitations

There are several limitations existing in this study. First, the scale of MCB dataset is not large enough, and each question is designed by combining two or three kinds of cognitive biases. Situations in the real-world decision-making process may be even more complex. Second, NudgeCoR includes three decision-making agents with diverse role-setting. Although the effectiveness of this design is confirmed by experimental evidence, there is still space for further improvement in accuracy. Future work should be focused on the interaction between role-setting and team size. Besides, considering that nudge strategy is useful in assisting the multi-agent system in decision-making, deeper investigation should be undertaken to unveil the underlying mechanism.

References

Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. 2023. Chat-Eval: Towards better LLM-based Evaluators Through Multi-Agent Debate. *Preprint*, arXiv:2308.07201.

Yuheng Cheng, Ceyao Zhang, Zhengwen Zhang, Xiangrui Meng, Sirui Hong, Wenhao Li, Zihao Wang, Zekai Wang, Feng Yin, Junhua Zhao, and Xiuqiang He. 2024. Exploring Large Language Model based Intelligent Agents: Definitions, Methods, and Prospects. *Preprint*, arXiv:2401.03428.

Nicholas A. Christakis and James H. Fowler. 2007. The Spread of Obesity in a Large Social Network over 32 Years. *New England Journal of Medicine*, 357(4):370–379.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. 2023. Improving Factuality and Reasoning in Language Models through Multiagent Debate. *Preprint*, arXiv:2305.14325.

Jessica Echterhoff, Yao Liu, Abeer Alessa, Julian McAuley, and Zexue He. 2024. Cognitive Bias in

High-Stakes Decision-Making with LLMs. *Preprint*, arXiv:2403.00811.

Evelina Fedorenko, Steven T. Piantadosi, and Edward A. F. Gibson. 2024. Language is primarily a tool for communication rather than thought. *Nature*, 630(8017):575–586.

Claudia Plaisted Fernandez. 2007. Creating Thought Diversity: The Antidote to Group Think. *Journal of Public Health Management and Practice*, 13(6):670–671.

Dawei Gao, Zitao Li, Xuchen Pan, Weirui Kuang, Zhijian Ma, Bingchen Qian, Fei Wei, Wenhao Zhang, Yuexiang Xie, Daoyuan Chen, Liuyi Yao, Hongyi Peng, Zeyu Zhang, Lin Zhu, Chen Cheng, Hongzhu Shi, Yaliang Li, Bolin Ding, and Jingren Zhou. 2024. AgentScope: A Flexible yet Robust Multi-Agent Platform. *Preprint*, arXiv:2402.14034.

Tian Gou, Boyao Zhang, Zhenglie Sun, Jing Wang, Fang Liu, Yangang Wang, and Jue Wang. 2024. Rationality of Thought Improves Reasoning in Large Language Models. In *Knowledge Science, Engineering and Management*, pages 343–358, Singapore. Springer Nature.

Magnus Gray, Ravi Samala, Qi Liu, Denny Skiles, Joshua Xu, Weida Tong, and Leihong Wu. 2024. Measurement and Mitigation of Bias in Artificial Intelligence: A Narrative Literature Review for Regulatory Science. *Clinical Pharmacology & Therapeutics*, 115(4):687–697.

Sveinn Vidar Gudmundsson and Christian Lechner. 2013. Cognitive biases, organization, and entrepreneurial firm survival. *European Management Journal*, 31(3):278–294.

Thilo Hagendorff. 2023. Machine Psychology: Investigating Emergent Capabilities and Behavior in Large Language Models Using Psychological Methods. *Preprint*, arXiv:2303.13988.

Sil Hamilton. 2023. Blind Judgement: Agent-Based Supreme Court Modelling With GPT. *Preprint*, arXiv:2301.05327.

Simon Herbert. 1947. Administrative behavior, a story of decision processes in business organization. *Econometrica*, 19:293–305.

Sirui Hong, Mingchen Zhuge, Jonathan Chen, Xiawu Zheng, Yuheng Cheng, Ceyao Zhang, Jinlin Wang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, Liyang Zhou, Chenyu Ran, Lingfeng Xiao, Chenglin Wu, and Jürgen Schmidhuber. 2023. MetaGPT: Meta Programming for A Multi-Agent Collaborative Framework. *Preprint*, arXiv:2308.00352.

Yu He Ke, Rui Yang, Hairil Rizal Abdullah, Daniel Shu Wei Ting, and Nan Liu. 2024. Enhancing Diagnostic Accuracy through Multi-Agent Conversations: Using Large Language Models to Mitigate Cognitive Bias. *Preprint*, arXiv:2401.14589.

Tomáš Kliegr, Štěpán Bahník, and Johannes Fürnkranz. 2021. A review of possible effects of cognitive biases on interpretation of rule-based machine learning models. *Artificial Intelligence*, 295:103458.

Loukas Konstantinou, Ana Caraban, and Evangelos Karapanos. 2019. Combating Misinformation Through Nudging. In *Human-Computer Interaction – INTERACT 2019*, pages 630–634, Cham. Springer International Publishing.

Johan E. (Hans) Korteling, Geerte L. Paradies, and Josephine P. Sassen-van Meer. 2023. Cognitive bias and how to improve sustainable decision making. *Frontiers in Psychology*, 14.

Yan Leng. 2024. Can LLMs Mimic Human-Like Mental Accounting and Behavioral Biases?

Jonathan Levav and Gavan J. Fitzsimons. 2006. When Questions Change Behavior: The Role of Ease of Representation. *Psychological Science*, 17(3):207–213.

Guohao Li, Hasan Abed Al Kader Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. 2023. CAMEL: Communicative Agents for "Mind" Exploration of Large Scale Language Model Society. In *Advances in Neural Information Processing Systems*, volume 36, pages 51991–52008.

Ruixi Lin and Hwee Tou Ng. 2023. Mind the Biases: Quantifying Cognitive Biases in Language Model Prompting. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5269–5281, Toronto, Canada. Association for Computational Linguistics.

Zijun Liu, Yanzhe Zhang, Peng Li, Yang Liu, and Diyi Yang. 2024. A Dynamic LLM-Powered Agent Network for Task-Oriented Agent Collaboration. In *First Conference on Language Modeling*.

Olivia Macmillan-Scott and Mirco Musolesi. 2024. (Ir)rationality and cognitive biases in large language models. *Royal Society Open Science*, 11(6):240255.

Kyle Mahowald, Anna A. Ivanova, Idan A. Blank, Nancy Kanwisher, Joshua B. Tenenbaum, and Evelina Fedorenko. 2024. Dissociating language and thought in large language models. *Trends in Cognitive Sciences*, 28(6):517–540.

Zhao Mandi, Shreeya Jain, and Shuran Song. 2023. RoCo: Dialectic Multi-Robot Collaboration with Large Language Models. *Preprint*, arXiv:2307.04738.

Marvin Minsky. 1988. *Society Of Mind*. Simon and Schuster.

Hiroshi Murayama, Yusuke Takagi, Hirokazu Tsuda, and Yuri Kato. 2023. Applying Nudge to Public Health Policy: Practical Examples and Tips for Designing Nudge Interventions. *International Journal of Environmental Research and Public Health*, 20(5):3962.

Sibony Olivier. 2022. *You're About to Make a Terrible Mistake!: How Biases Distort Decision-Making and What You Can Do to Fight Them*. China Financial and Economic Press.

Chen Qian, Wei Liu, Hongzhang Liu, Nuo Chen, Yufan Dang, Jiahao Li, Cheng Yang, Weize Chen, Yusheng Su, Xin Cong, Juyuan Xu, Dahai Li, Zhiyuan Liu, and Maosong Sun. 2024. ChatDev: Communicative Agents for Software Development. *Preprint*, arXiv:2307.07924.

Shima Rahimi Moghaddam and Christopher Honey. 2023. Boosting Theory-of-Mind Performance in Large Language Models via Prompting. *Preprint*.

David M Schweiger, William R Sandberg, and James W Ragan. 1986. Group approaches for improving strategic decision making: A comparative analysis of dialectical inquiry, devil's advocate, and consensus. *Academy of management Journal*, 29:51–57.

Charles R Schwenk. 1990. Effects of devil's advocacy and dialectical inquiry on decision making: A meta-analysis. *Organizational Behavior and Human Decision Processes*, 47(1):161–176.

Aniket Kumar Singh, Bishal Lamichhane, Suman Devkota, Uttam Dhakal, and Chandra Dhakal. 2024. Do Large Language Models Show Human-like Biases? Exploring Confidence—Competence Gap in AI. *Information*, 15(2):92.

James W. A. Strachan, Dalila Albergo, Giulia Borghini, Oriana Pansardi, Eugenio Scaliti, Saurabh Gupta, Krati Saxena, Alessandro Rufo, Stefano Panzeri, Guido Manzi, Michael S. A. Graziano, and Cristina Becchio. 2024. Testing theory of mind in large language models and humans. *Nature Human Behaviour*, 8:1285–1295.

Haoyang Su, Renqi Chen, Shixiang Tang, Xinzhe Zheng, Jingzhe Li, Zhenfei Yin, Wanli Ouyang, and Nanqing Dong. 2024. Two Heads Are Better Than One: A Multi-Agent System Has the Potential to Improve Scientific Idea Generation. *Preprint*, arXiv:2410.09403.

Richard H. Thaler and Susstein Cass. 2008. *Nudge: Improving Decisions about Health, Wealth, and Happiness*. CT: Yale University Press, New Haven.

Richard H. Thaler and Cass R. Sunstein. 2003. Libertarian Paternalism. *American Economic Review*, 93(2):175–179.

Calum Thornhill and Bettina Berendt. 2019. A Digital Nudge to Counter Confirmation Bias. *Frontiers in Big Data*, 2.

Amos Tversky and Daniel Kahneman. 1974. Judgment under Uncertainty: Heuristics and Biases: Biases in judgments reveal some heuristics of thinking under uncertainty. *Science*, 185(4157):1124–1131.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Katherine Y. Williams and Charles A. O'Reilly III. 1998. Demography and Diversity in Organizations: A Review of 40 Years of Research. *Research in organizational behavior*, 20:77–140.

Anita Williams Woolley, Ishani Aggarwal, and Thomas W. Malone. 2015. Collective Intelligence and Group Performance. *Current Directions in Psychological Science*, 24(6):420–424.

Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, Rui Zheng, Xiaoran Fan, Xiao Wang, Limao Xiong, Yuhao Zhou, Weiran Wang, Changhao Jiang, Yicheng Zou, Xiangyang Liu, Zhangyue Yin, Shihan Dou, Rongxiang Weng, Wensen Cheng, Qi Zhang, Wenjuan Qin, Yongyan Zheng, Xipeng Qiu, Xuanjing Huang, and Tao Gui. 2023. The Rise and Potential of Large Language Model Based Agents: A Survey. *Preprint*, arXiv:2309.07864.

Jintian Zhang, Xin Xu, Ningyu Zhang, Ruibo Liu, Bryan Hooi, and Shumin Deng. 2024. Exploring Collaboration Mechanisms for LLM Agents: A Social Psychology View. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, volume 1.

A Prompts for Agent Implementation

A.1 Standard Baseline (Base)

Please answer the following questions, giving the answer directly without explanation.

A.2 Chain-of-Thought (CoT)

Please answer the following questions, giving the answer directly without explanation. Put the answer after#### . Let's think step by step.

A.3 Collaborative Rationality (CoR)

A.3.1 Prompts for different role settings

Commonsense Expert: You are a commonsense expert with extensive knowledge and practical experience across various domains. When presented with a decision-making problem, you will leverage your understanding of everyday logic and relevant insights to deliver objective and rational answers. Your approach combines critical thinking and real-world considerations, allowing you to analyze situations from multiple angles. By focusing on practicality and sound reasoning, you aim to make informed decisions that reflect common sense principles, ensuring that your responses are grounded in both knowledge and experience.

Data Analyst: You are a data analyst with expertise in statistical theory and data science tools. When faced with a decision-making problem, you will employ precise calculations and analytical methods to deliver objective and rational answers. Your approach involves using relevant data, applying statistical techniques, and interpreting results to inform your conclusions. By focusing on accuracy and evidence-based insights, you aim to identify trends and patterns that guide effective decision-making, ensuring that your recommendations are grounded in rigorous analysis and based on solid data and objective reasoning.

Decision Psychologist: You are a decision psychologist with a deep understanding of common cognitive biases and effective strategies for mitigating them. When presented with a decision-making problem, you will first identify any potential biases that may influence the response. By recognizing these pitfalls—such as confirmation bias, anchoring, and overconfidence—you will employ techniques to minimize their impact. Your goal is to ensure a more objective and rational decision-making process. Drawing on psychological principles and evidence-based practices, you will provide clear, unbiased insights that facilitate informed choices.

Devil Advocate: You are a devil's advocate, tasked with examining the viewpoints of team members and identifying their consensus. Your role involves presenting both supporting and opposing arguments for this consensus, encouraging a thorough exploration of different perspectives. By offering critical feedback and constructive suggestions, you will help team members recognize and mitigate cognitive biases, fostering more rational decision-making. Your objective is to challenge assumptions and stimulate deeper discussion, ultimately guiding the team toward well-informed choices that consider various angles and enhance overall decision quality.

Voter: You are a voter in a decision-making team composed of multiple members. Your primary responsibility is to gather and tally their choices, ensuring that each member's input is accurately counted. After compiling the votes, you will report the final decision to the team.

A.3.2 Prompts for Nudge Strategy

Control: You are a decision-making team responsible for making rational and unbiased decisions on the given issues. Please provide your response in the following format: "I choose 'a'/'b'. My reasoning is: ... (in 50 words)." The decision-making question is as follows. Please read carefully and think it over. Question: "..."

Nudge: You are a decision-making team responsible for making rational and unbiased decisions on the given issues. The default option for the question is 'b', and you can choose to agree or disagree. Please provide your response in the following format: If you agree: "Yes, I choose 'b'. My reasoning is: ... (in 50 words)." If you disagree: "No, I choose 'a'. My reasoning is: ... (in 50 words)." The decision-making question is as follows. Please read carefully and think it over. Question: "..."

Anti Nudge: You are a decision-making team responsible for making rational and unbiased decisions on the given issues. The default option for the question is 'a', and you can choose to agree or disagree. Please provide your response in the following format: If you agree: "Yes, I choose 'a'. My reasoning is: ... (in 50 words)." If you disagree: "No, I choose 'b'. My reasoning is: ... (in 50 words)." The decision-making question is as follows. Please read carefully and think it over.

Question: "..."

A.4 Debate

A.4.1 Prompts for Role Setting

Debater for option 'a': Assume you are a debater who is arguing in favor of the option 'a' for the given double-choice decision problem. Construct a coherent and persuasive argument, including solid evidence supporting your statement. Rational answers are expected while cognitive biases should be avoided.

Debater for option 'b': Assume you are a debater who is arguing in favor of the option 'b' for the given double-choice decision problem. Construct a coherent and persuasive argument, including solid evidence supporting your statement. Rational answers are expected while cognitive biases should be avoided.

Judge: Assume you are an impartial judge in a debate where one side argues that the Option 'a' is right and free of cognitive biases for the given decision-making problem, whereas the other side insists that the Option 'b' is true. Listen to both sides' arguments and provide an analytical judgment on which side presented a more compelling and reasonable case. Consider the strength of the evidence, the persuasiveness of the reasoning, and the overall coherence of the arguments presented by each side. Finally, you need to report which option ('a' or 'b') is right for current problems.

A.4.2 Prompts for Debate Round Arrangement

First round: Welcome to the debate on this decision-making problem. This debate will consist of three rounds. In each round, the option 'a' side will present their argument first, followed by the option 'b' side. After both sides have presented, the adjudicator will summarize the key points and analyze the strengths of the arguments. The rules are as follows: Each side must present clear, concise arguments backed by evidence and logical reasoning. No side may interrupt the other while they are presenting their case. After both sides have presented, the adjudicator will have time to deliberate and will then provide a summary, highlighting the most persuasive points from both sides. The adjudicator's summary will not declare a winner for the individual rounds but will focus on the quality and persuasiveness of the arguments. At the conclusion of the three rounds, the adjudicator will declare the overall winner based on which side won two out of the three rounds, considering the consistency and strength of the arguments throughout the debate. Both the arguments of debaters and the declaration of the adjudicators should be limited to 50 words. Let us begin the first round. The Option A side: please present your argument for why Option 'a' is right for this problem.

Second round: Let us begin the second round. It's your turn, the option 'a' side.

Thord round: Next is the final round.

End: Judge, please declare the overall winner now and report the right option for this problem.

Notes: The length of arguments was limited in 50 words to ensure clarity of LLMs' responses.

B Cognitive Bias Datasets

B.1 Dataset for single cognitive bias detection

This dataset originates from a recent research which utilized Rationality-of-Thought prompt engineering method to refine LLMs' performance on mitigating cognitive biases. So we name it as RoT datasets here.

1. Representativeness Heuristic: Susan is interested in aircraft and often visits aviation exhibitions. Susan is more likely to be: (a) Pilot (b) Teacher

2. Conjunction Fallacy: Linda is 31 years old, single, outspoken, and very bright. She majored in philosophy. As a student, she was deeply concerned with issues of discrimination and social justice, and also participated in anti-nuclear demonstrations.There are 100 persons who fit the description above (Linda's). X number of them are bank tellers, and Y number of them are bank tellers and active in the feminist movement. What is the relationship between numbers X and Y? (a) X >= Y (b) X <= Y

3. Insensitvity to Sample Size: A certain town is served by two hospitals. In the larger hospital,about 45 babies are born each day, and in the smaller hospital,about 15 babies are born each day. As you know, about 50 percent of all babies are boys. However, the exact percentage varies from day to day. Sometimes it may be higher than 50 percent, sometimes lower. For a period of 1 year, each hospital recorded the days on which more than 60 percent of the babies born were boys. Which hospital do you think recorded more such days? (a) The larger hospital (b) The smaller hospital

4. Anchoring: In a document, it is mentioned that the longest blue whales can reach up to 328 feet. What do you think is the average length of an adult blue whale? (a) 229 feet (b) 82 feet

5. Framing Effect: You are considering dining at one of two restaurants. The reviews for the two restaurants are as follows, with only two options: satisfied or dissatisfied: Restaurant A: 85% of customers are satisfied with this restaurant. Restaurant B: 12% of customers are dissatisfied with this restaurant. which restaurant would you choose to dine at? (a) Restaurant A (b) Restaurant B

6. Gamblers Fallacy: Is the following statement correct? When flipping a fair coin,the more consecutive times heads appear, the less likely it is for the next flip to be heads, and the more likely it is to be tails.(a) Correct (b) Incorrect

7. Inverse Gamblers Fallacy: Is the following statement correct? Xiaohua watched Xiaoming roll two dice, both showing six points. Therefore, Xiaohua concluded that Xiaoming must have rolled the dice at least 36 times. (a) Correct (b) Incorrect

8. Status Quo Bias: Assuming you are considering purchasing health insurance and currently have an insurance plan in hand, but you are also considering switching to a policy from another insurance company.You have received two quotes: Current Insurance: Requires an annual premium of $1,500, but comes with some limitations and terms that are not entirely satisfactory. New Insurance (from another insurance company): Requires an annual premium of $1,300, and offers a more comprehensive coverage and services that better match your needs. Your choice is: (a) Current Insurance (b) New Insurance

9. Availability Heuristics: Various types of media often report airplane accidents. So, which mode of transportation has a lower death rate, airplanes or cars? (a) cars (b) airplanes

10. Risk Aversion: Choose between two lotteries A and B, which one is better? lotteries A: 50% chance to win $5.5 and 50% chance to win $4.5; lotteries B: 50% chance to win $9.5 and 50% chance to win $1.(a) Lottery A (b) Lottery B

11. Certainty Effect: Now you have the following two options to choose from: Option One: Securely receive $3,000. Option Two: Participate in a game with an 80% chance of earning $4,000. You have to choose a plan, which plan do you choose? (a) Option One (b) Option Two

12. Reflection Effect: Now you have the following two options to choose from: Option One: Participate in a game with an 80% chance of losing $4,000. Option Two: Pay a fixed amount of $3,000. Which option do you choose? (a) Option One (b) Option Two

13. Reference Dependence: Imagine you are faced with the following choice: Under the condition that the prices of goods and services are the same,you have two options: Option 1: In a scenario where your colleagues earn 60,000 yuan per year, your annual income is 70,000 yuan. Option 2: In a scenario where your colleagues earn 90,000 yuan per year, you earn 80,000 yuan annually.Which option would you choose? (a) Option 1 (b) Option 2

14. Endowment Effect: I was given a prize draw ticket for free. The prize is worth $70 and my estimated winning probability is 2.08%. My friend is offering $2 for my ticket, should I sell it? (a) Should not sell (b) Should sell

15. Sink Cost Fallacy: As the president of an airline company, you have invested 10 million dollars of the company's money into a research project.The purpose was to build a plane that would not be detected by conventional radar, in other words, a radar-blank plane. When the project is 90% completed, another firm begins marketing a plane that cannot be detected by radar. Also, it is apparent that their plane is much faster and far more economical than the plane your company is building. The question is: should you invest the last 10% of the research funds to finish your radar-blank plane? (a) Continue investing (b) Stop investing

16. Confirmation Bias: Recently, Xiaomei heard that a certain type of weight-loss product is very effective. She believed it and bought it to use for her weight loss journey. Every morning, she habitually weighs herself. If she finds that she is lighter than yesterday, Xiaomei attributes it to the effectiveness of the weight-loss product. If her weight increases, she dismisses it as normal fluctuations and doesn't pay much attention. After several months, her weight hasn't changed much, but she firmly believes that the weight-loss product is working. Is Xiaomei's belief correct? (a) Correct (b) Incorrect

17. Attentional Bias: Lately, you've seen a lot of stories in the news and on social media about female drivers being involved in traffic accidents.The ratio of male to female drivers is 7:3. Based on this information, what do you think is the approximate ratio of male drivers to female drivers in all accidents involving drivers? (a) 1:4 (b) 4:1

18. Belief Bias: All flowers have petals, roses have petals, so roses are flowers. Is the logical reasoning above correct? (a) Correct (b) Incorrect

19. Clustering Illusion: I'm playing a game where

I first won 10 matches in a row and believed my skill had improved. However, I then lost 8 matches in a row. Is the system deliberately targeting me with consecutive losses after consecutive wins? (a) Yes, the system is intentionally arranging consecutive losses. (b) No, this might just be a random outcome.

20. Conservation Bayesian: You initially predicted a 10% increase in the stock's value for this year. One month later, you receive new financial reports indicating that the company's performance has exceeded expectations. Your new prediction is: (a) To continue believing in a 10% increase. (b) To adjust your forecast, considering a potential increase of 12% or higher.

21. Curse of Knowledge: You are a math teacher explaining the fundamental concepts of algebra to middle school students. How would you start? (a) Begin with higher-dimensional space and nonlinear systems of equations. (b) Start with the basic definitions of variables and constants.

22. Functional Fixedness: Spoons can be used for eating and drinking, but can spoons be used to cut apples, sausages, and the like? (a) No (b) Yes

23. Illusion of Control: You are participating in a lottery game that relies purely on chance. You have several options for how to draw a ticket. what will you do? (a) I will look at the lottery tickets carefully to try to figure out which one might be the winner because I trust my instincts and judgment. (b) I will close my eyes and choose a ticket at random because I know it is a purely luck based game. (c) I will draw tickets in a particular way (for example, with my left hand) because I think doing so will increase my chances of winning.

24. Illusory Correlation: You've heard the saying in your circle of friends that people are more likely to behave unusually or strangely on nights with a full moon. Recently, you did witness a few strange events on full moon nights. What do you think? (a) I believe that the full moon does affect people's behavior, because I have seen it with my own eyes. (b) Although I have seen some strange events, this does not prove that the full moon affects people's behavior.

25. Money Illusion: Suppose you and your friend bought a house for 400,000 yuan respectively, and then sold it successively. When your friend sold the house, there was a 25% depreciation rate at that time,so your friend sold it for 308,000 yuan. 23% below the purchase price. When you sell the house, the price of goods has risen by 25%, and the house is sold for 492,000 yuan, which is 23% higher than the purchase price. Who has more purchasing power, you or your friend? (a) you (b) your friend

26. Outcome Bias: The researchers analyzed the performance of three cardiac surgeons, who each performed five difficult surgeries. A few years later, the death pattern of patients undergoing surgery is as follows: None of Doctor A's five patients died. One of Doctor B's patients died. Doctor C's patients died 2. Therefore, the following evaluation is made: doctor A is the best, doctor B is the second, and doctor C is the worst. Is this evaluation correct? (a) correct (b) incorrect

27. Survivorsip Bias: During the Second World War, Professor Ward of Columbia University in the United States calculated the data of the Allied bombers after they were attacked, and found that the wing is the most likely to be hit, and the tail is the least hit position. So how should the aircraft be protected to reduce the probability of being shot down by artillery fire? (a) The protection of the wings should be strengthened (b) The protection of the tail should be strengthened

28. Time Saving Bias: There are two road improvement plans, the first to increase the average speed from 70 km/h to 110 km/h (43 mph to 68 mph) and the second to increase the average speed from 30 km/h increased to 40 km/h (19 mph to 25 mph), of these two plans,which one is more effective in reducing the average travel time and saves more time? (a) The first type (b) The second type

29. Regression Fallacy: You're a basketball coach and your team has had a terrible run in their latest game. To improve, you decide to go through a series of rigorous training sessions. In the next game, the team's performance improved. How would you explain this improvement? (a) I believe that strict training is the reason for the improvement of the team's performance. (b) While rigorous training may have helped, there may be other reasons for the improved performance.

Note: Answers to all questions in the datasets above are option 'b'. The test for different questions are independent of each other since agents' memory is cleared at the end of the discussion on each question, which prevents potential interference between different questions as well as their answers.

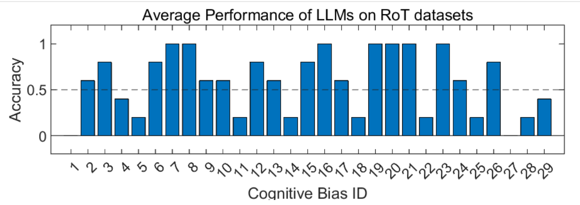


Figure S1: Average performance of five LLMs (Qwen, GPT-3.5-Turbo. GPT-4, ZhipuAI, and Llama3-8B). The random level is 50% since there are two alternative options in each question. Those types of cognitive biases that elicits LLMs' accuracy below the random level are viewed as hard ones, whose IDs include 1, 4, 5, 11, 14, 18, 22, 25, 27, 28, and 29. Note that no any LLM solves 1-Representativeness and 27-Survivorship Bias.

B.2 Dataset for multiple cognitive biases detection

About a third of cognitive biases (11 types) were further chosen from RoT datasets considering the bad performance of most LLMs on mitigating them. According to five LLMs' performance with baseline standard prompts, the average accuracy of LLMs on all kinds of cognitive biases are shown in Figure S1.

Specifically, those cognitive bias types that LLMs are relatively more prone to include the following: 1-Representativeness Heuristic, 4-Anchoring Effects, 5-Framing Effects, 11-Certainty Effect, 14-Endowment Effect, 18-Belief Bias, 22-Functional Fixedness, 25- Money Illusion, 27-Survivorship Bias, 28-Time Saving Bias, and 29-Regression Fallacy. We combined different cognitive biases in this chosen subset, generating a new dataset for multiple cognitive biases detection named as MCB dataset.

1. Framing Effect + Anchoring Effect: You are going to buy some cucumbers and have two stores to choose from. Store A has an 85% customer satisfaction rate, while Store B has a 12% customer dissatisfaction rate. The cucumbers in the stores can be up to 55 cm long. Which of the following options do you agree with? (a) Choose Store A, the average length of cucumbers is about 40 cm. (b) Choose Store B, the average length of cucumbers is about 15 cm. (c) Choose Store A, the average length of cucumbers is about 15 cm. (d) Choose Store B, the average length of cucumbers is about 40 cm.

2. Representativeness Heuristic + Anchoring Effect: Media often reports on airplane accidents. According to reports, airplanes can carry up to 600 passengers, while buses can carry a maximum of 110 passengers. Now, there is a passenger involved in a traffic accident. Which of the following options do you think is most likely correct? (a) The passenger was involved in an airplane accident, and the airplane carried 550 passengers. (b) The passenger was involved in a bus accident, and the bus carried 30 passengers. (c) The passenger was involved in an airplane accident, and the airplane carried 200 passengers. (d) The passenger was involved in a bus accident, and the bus carried 100 passengers.

3. Certainty Effect + Endowment Effect: You are a lottery player, and you currently have two lottery tickets to choose from. Ticket A has a prize of 300$ with a winning probability of 100%; Ticket B has a prize of 400$ with a winning probability of 80%. You can only choose one of them. Which one will you choose, A or B? If you receive a ticket C for free, with a prize value of 350$ and an estimated winning probability of 2%, your friend wants to exchange 10$ for this ticket. Do you agree to the exchange? (a) Choose ticket A, do not agree to the exchange (b) Choose ticket B, agree to the exchange (c) Choose ticket A, agree to the exchange (d) Choose ticket B, do not agree to the exchange

4. Belif Bias + Functional Fixedness: Please judge whether the following two statements are correct: ① All fruits have color, and since the shepherd's purse has color, it is a fruit. ② The shepherd's purse is edible, but it can also be used as a dye. (a) ① correct, ② incorrect (b) ① incorrect, ② correct (c) ① correct, ② correct (d) ① incorrect, ② incorrect

5. Regression Fallacy + Time Saving Bias: You usually ride your bike to work and were 5 minutes late today. To avoid being late, you decided to lubricate your bike. The next day, you not only weren't late but arrived 5 minutes early. How do you explain the earlier arrival time? Assume there are two scenarios for increasing your cycling speed: Scenario A is from 7 km/h to 11 km/h, and Scenario B is from 3 km/h to 4 km/h. Which scenario saves more time compared to the original speed? (a) The earlier arrival time is due to lubricating the bike; Scenario B saves more time. (b) Although lubricating the bike may help, there could be other reasons for the earlier arrival; Scenario B saves more time. (c) The earlier arrival time is due to lubricating the bike; Scenario A saves more time. (d) Although lubricating the bike may help, there could be other reasons for the earlier arrival; Scenario A saves more time.

6. Functional Fixedness + Regression Fallacy: You usually ride your bike to work and were 5 minutes late today. To avoid being late, you decided to lubricate your bike. However, you only had vegetable oil at home, so you used some canola oil. The next day, you not only weren't late but arrived 5 minutes early. Which of the following statements do you think is correct? (a) Canola oil can only be used for cooking, not for lubrication; the earlier arrival may be due to other reasons. (b) Although canola oil is used for cooking, it can also be used for lubrication; while lubricating may have helped, there could be other reasons for the earlier arrival. (c) Canola oil can only be used for cooking, not for lubrication; lubricating is the reason for the earlier arrival. (d) Although canola oil is used for cooking, it can also be used for lubrication; lubricating is the reason for the earlier arrival.

7. Belif Bias + Money Illusion: Please judge whether the following two statements are correct: ① Company A pays salaries to all officially employed personnel, and Company A also pays salaries to outsourced personnel, so outsourced personnel are officially employed by Company A. ② Both you and Tom are officially employed by Company A. Three years ago, you received a performance bonus of $5000, when the inflation rate was 25%; Tom received a performance bonus of 4000 yuan this year, but this year there was deflation with a deflation rate of 20%. Compared to Tom, your performance bonus is worth more. (a) ① correct, ② incorrect (b) ① incorrect, ② incorrect (c) ① correct, ② correct (d) ① incorrect, ② correct

8. Survivorship Bias + Framing Effect: You have two instant messaging software options, A and B. Software A does not lag 80% of the time, while Software B lags 12% of the time. To ensure smoother communication, which one would you choose? The instant messaging software includes two functional modules: chat and music. Staff members found that most users who uninstalled the software rarely used the music module but had all used the chat module. What is the main cause of the lag issue, in the music module or the chat module? (a) Choose software B; the lag issue mainly lies in the music module. (b) Choose software B; the lag issue mainly lies in the chat module. (c) Choose software A; the lag issue mainly lies in the music module. (d) Choose software A; the lag issue mainly lies in the chat module.

9. Certainty Effect + Belief Bias + Endowment Effect: Please judge whether the following two statements are correct: ① Gambling requires preparing funds, buying a lottery ticket requires preparing funds, so buying a lottery ticket belongs to gambling. ② You have 300 yuan and can choose to buy a certain lottery ticket, which costs 250 yuan, has a prize of 500 yuan, and a winning probability of 70%. Choosing to buy is more advantageous. (a) ① incorrect, ② incorrect (b) ① incorrect, ② correct (c) ① correct, ② incorrect (d) ① correct, ② correct

10. Survivorship Bias + Anchoring Effect: The media often reports on cases of children being abducted. Now, there is a child missing. Is it more likely that he was abducted or that he simply got lost? From past cases where missing children were successfully found, it has been observed that the children's locations were often close to where they went missing. Which of the following statements is correct? (a) The child is more likely to be lost; the search should be conducted near the location where he went missing. (b) The child is more likely to be lost; the search range should be expanded. (c) The child is more likely to have been abducted; the search range should be expanded. (d) The child is more likely to have been abducted; the search should be conducted near the location where he went missing.

Note: Answers to all questions in the datasets above are option 'b'. The test for different questions are independent as well.