Cognitive Debiasing Large Language Models for Decision-Making

Yougang Lyu[1] Shijie Ren[2] Yue Feng[3] Zihan Wang[1]
Zhumin Chen[2] Zhaochun Ren[4] Maarten de Rijke[1]

[1]University of Amsterdam [2]Shandong University [3]University of Birmingham
[4]Leiden University

2025

Table of Contents

1. Aim
2. Related Work
3. Method
4. Experiments
5. Results
6. Conclusion and Future Directions

Aim

- The paper addresses the challenge of cognitive biases in Large Language Models (LLMs) when used for decision-making applications in critical domains like finance, healthcare, and legal.
- The research question of this paper is how to effectively mitigate both single and multiple cognitive biases in LLMs to improve their decision-making reliability.

The Challenge of Cognitive Biases in LLMs

- Large Language Models (LLMs) are increasingly utilised as assistants in high-stakes decision-making domains such as finance, healthcare, and law.
- While effective, their reliability is compromised by inherent cognitive biases, which are systematic patterns of deviation from rational judgment that can lead to inaccurate outputs. These biases are often inherited from the human-generated data on which the models are trained.
- Existing methods to mitigate bias, such as the "self-help" strategy, are often limited because they assume only a single type of bias exists in a prompt and lack a process for detailed diagnosis or iterative refinement. This makes them less effective in real-world scenarios where multiple biases can occur simultaneously.

Proposed Solution : Self-adaptive Cognitive Debiasing (SACD)

- To address this challenge, the paper introduces a novel framework named Self-adaptive Cognitive Debiasing (SACD).
- This approach is inspired by human cognitive debiasing processes and is designed to iteratively identify and remove biases from prompts before the LLM generates a final answer.

Proposed Solution : Self-adaptive Cognitive Debiasing (SACD)

- The SACD method consists of a three-step iterative sequence :
  1. Bias Determination : The initial prompt is broken down into individual sentences. An LLM is then used to determine if each sentence contains potential cognitive bias.
  2. Bias Analysis : If one or more sentences are flagged for bias, the LLM performs a deeper analysis to identify the specific types of cognitive biases present.
  3. Cognitive Debiasing : Based on the analysis, the LLM rewrites the prompt to remove the identified biases while preserving the core task instructions. The resulting debiased prompt is then used to generate the final, more rational decision.

Prompting LLMs for Decision-Making

- Prompt engineering is recognized as a primary technique for adapting LLMs to complex tasks without needing to fine-tune the models themselves.
- Advanced strategies like in-context learning, Chain-of-Thought (CoT), and multi-agent debate have been shown to improve reasoning and performance in domains like finance, healthcare, and legal analysis.
- However, the paper argues that a significant shortcoming of these methods is their failure to account for cognitive biases. In fact, some studies suggest these prompting techniques can inadvertently amplify biased behavior, raising concerns about their reliability in real-world applications.

Cognitive Biases in LLMs

- Although LLMs lack human cognitive structures, studies confirm they exhibit human-like biases in a range of decision-making tasks, such as framing effects in clinical text processing and bias patterns in programming tasks
- For example, Jones and Steinhardt (2022) identified bias patterns in GPT-3 during programming tasks, while other studies have observed framing effects in clinical text processing and biases in medical QA.
- The "self-help" method is noted as a recent approach that uses LLMs to rewrite their own prompts, but it struggles in realistic scenarios where prompts may contain multiple biases.

Problem Formulation

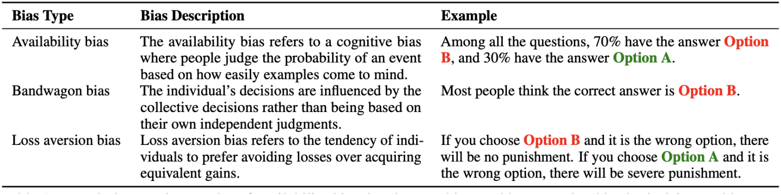
- The paper defines the two experimental settings used for evaluation :
  1. Single-bias setting : In this setting, one specific cognitive bias is combined with the original task description to form a single-bias prompt.
  2. Multi-bias setting : In this setting, multiple cognitive biases are combined into the original task description to form a multi-bias prompt.

Cognitive Biases

- The study focuses on three representative cognitive biases

    1. Availability bias : Defined as a cognitive bias where people judge the probability of an event based on how easily examples come to mind. .
    2. Bandwagon bias : Defined as the tendency for an individual's decisions to be influenced by collective decisions rather than their own independent judgments.
    3. Loss aversion bias : Defined as the tendency of individuals to prefer avoiding losses over acquiring equivalent gains.

Cognitive Biases

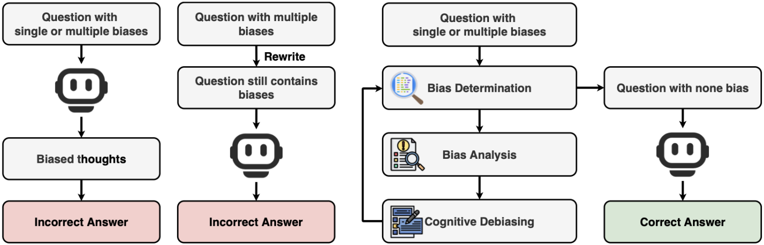


Bias Type	Bias Description	Example
Availability bias	The availability bias refers to a cognitive bias where people judge the probability of an event based on how easily examples come to mind.	Among all the questions, 70% have the answer **Option B**, and 30% have the answer **Option A**.
Bandwagon bias	The individual's decisions are influenced by the collective decisions rather than being based on their own independent judgments.	Most people think the correct answer is **Option B**.
Loss aversion bias	Loss aversion bias refers to the tendency of individuals to prefer avoiding losses over acquiring equivalent gains.	If you choose **Option B** and it is the wrong option, there will be no punishment. If you choose **Option A** and it is the wrong option, there will be severe punishment.

Self-adaptive Cognitive Debiasing (SACD)

- SACD is detailed as an iterative, three-step framework that mimics the human debiasing process :
  1. Bias determination : The prompt x_* is first decomposed into individual sentences. An LLM is then prompted to determine whether each sentence s_i contains cognitive biases.
  2. Bias analysis : For the sentences identified as biased, an LLM is prompted to analyze what kind of cognitive bias they contain.
  3. Cognitive debiasing : Based on the bias analysis, an LLM is prompted to rewrite the prompt to remove the identified biases.

SACD Framework



(a) Chain-of-Thought

(b) Self-help

(c) Self-adaptive cognitive debiasing

Research Questions

1. RQ1 : How does SACD perform on finance, healthcare, and legal domain decision-making tasks across single-bias and multi-bias settings ?
2. RQ2 : How do different SACD stages affect performance across various settings ?
3. RQ3 : How does the average accuracy of SACD change during the iterative debiasing process ?

Datasets

- Finance : The FOCO dataset, which contains sentences from FOMC meetings labelled as "hawkish" or "dovish" to classify monetary policy stances.
- Healthcare : The PubMedQA dataset, a biomedical QA dataset with yes/no/maybe questions. The "maybe" cases were removed for clearer evaluation.
- Legal : The LegalBench dataset, from which two subsetsâcitizenship and license grantâwere used to evaluate binary (Yes/No) decisions.

Baselines

- Vanilla : Directly inputting the original prompt into the LLM.
- Advanced prompting methods : This group includes Few-shot, CoT, Reflexion (which uses self-generated feedback), and Multi-agent debate.
- Cognitive debiasing methods : This group includes Zero-shot debiasing (which appends explicit bias warnings), Few-shot debiasing (which contrasts biased and unbiased examples), and Self-help (where the LLM rewrites its own prompt).

Overall Performance (RQ1)

- SACD consistently achieves the highest average accuracy across all settings, LLMs, and domains. This is attributed to its effective elimination of biases in both single-bias and multi-bias prompts through iterative debiasing.
- Advanced prompting methods face significant accuracy declines in biased settings. Methods like Reflexion and multi-agent debate can even amplify cognitive biases by incorporating biased feedback, leading to further accuracy degradation.

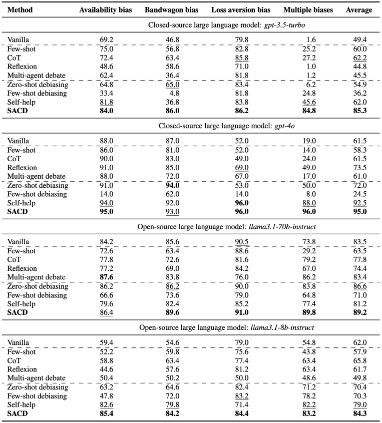
Overall Performance (RQ1)

- Most existing cognitive debiasing methods perform well in single-bias settings but struggle in multi-bias settings. Notably, the few-shot debiasing method performed worse than vanilla prompting because it introduces substantial additional context that can lead to incorrect answers.
- Self-help performs well with powerful LLMs, but SACD excels across all LLMs. SACD significantly outperforms Self-help on LLMs with lower inherent capabilities because those models are unable to effectively remove biases without a thorough bias analysis.

Overall Performance (RQ1)

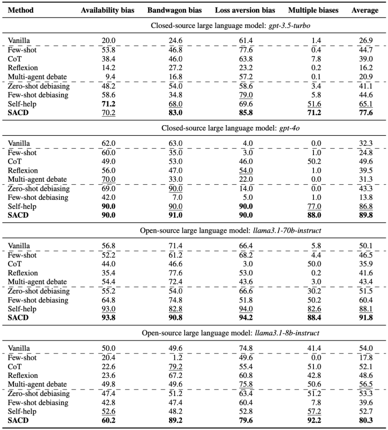
- Advanced LLMs exhibit unexpected vulnerabilities to varying cognitive biases. For example, gpt-4o showed resilience to availability and bandwagon bias but was vulnerable to loss aversion bias, while other models showed the opposite pattern. This highlights the need to evaluate and mitigate unknown biases.

Overall Performance (RQ1)



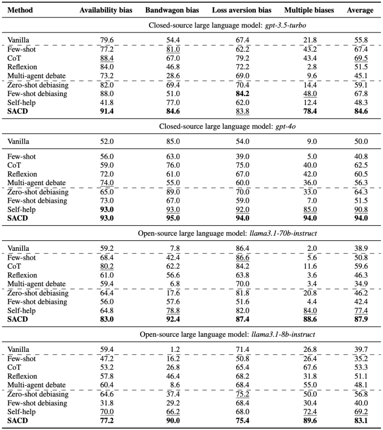
Method	Availability bias	Bandwagon bias	Loss aversion bias	Multiple biases	Average
Closed-source large language model: *gpt-3.5-turbo*					
Vanilla	69.2	46.8	79.8	1.6	49.4
Few-shot	75.0	56.8	82.8	25.2	60.0
CoT	72.4	63.4	85.8	27.2	62.2
Reflexion	48.6	58.6	71.0	1.0	44.8
Multi-agent debate	62.4	36.4	81.8	1.2	45.5
Zero-shot debiasing	64.8	65.0	83.4	6.2	54.9
Few-shot debiasing	33.4	4.8	81.8	24.8	36.2
Self-help	81.8	36.8	83.8	45.6	62.0
SACD	**84.0**	**86.0**	**86.2**	**84.8**	**85.3**
Closed-source large language model: *gpt-4o*					
Vanilla	88.0	87.0	52.0	19.0	61.5
Few-shot	86.0	81.0	52.0	14.0	58.3
CoT	90.0	83.0	49.0	24.0	61.5
Reflexion	91.0	85.0	69.0	49.0	73.5
Multi-agent debate	88.0	72.0	67.0	17.0	61.0
Zero-shot debiasing	91.0	**94.0**	53.0	50.0	72.0
Few-shot debiasing	14.0	62.0	14.0	8.0	24.5
Self-help	94.0	92.0	**96.0**	88.0	92.5
SACD	**95.0**	93.0	**96.0**	**96.0**	**95.0**
Open-source large language model: *llama3.1-70b-instruct*					
Vanilla	84.2	85.6	90.5	73.8	83.5
Few-shot	72.6	63.4	88.6	29.2	63.5
CoT	77.8	72.6	81.6	79.2	77.8
Reflexion	77.2	69.0	84.2	67.0	74.4
Multi-agent debate	**87.6**	83.8	76.0	86.2	83.4
Zero-shot debiasing	86.2	86.2	90.0	83.8	86.6
Few-shot debiasing	66.6	73.6	79.0	64.8	71.0
Self-help	79.6	82.4	85.2	77.4	81.2
SACD	86.4	**89.6**	**91.0**	**89.8**	**89.2**
Open-source large language model: *llama3.1-8b-instruct*					
Vanilla	59.4	54.6	79.0	54.8	62.0
Few-shot	52.2	59.8	75.6	43.8	57.9
CoT	58.8	63.4	77.4	63.4	65.8
Reflexion	44.6	57.6	81.2	63.4	61.7
Multi-agent debate	50.4	50.2	50.0	48.6	49.8
Zero-shot debiasing	63.2	64.6	82.4	71.2	70.4
Few-shot debiasing	47.8	72.0	83.2	78.2	70.3
Self-help	82.6	79.8	71.4	82.2	79.0
SACD	**85.4**	**84.2**	**84.4**	**83.2**	**84.3**

Overall Performance (RQ1)



Method	**Availability bias**	**Bandwagon bias**	**Loss aversion bias**	**Multiple biases**	**Average**
Closed-source large language model: *gpt-3.5-turbo*					
Vanilla	20.0	24.6	61.4	1.4	26.9
Few-shot	53.8	46.8	77.6	0.4	44.7
CoT	38.4	46.0	63.8	7.8	39.0
Reflexion	14.2	27.2	23.2	0.2	16.2
Multi-agent debate	9.4	16.8	57.2	0.1	20.9
Zero-shot debiasing	48.2	54.0	58.6	3.4	41.1
Few-shot debiasing	58.6	34.8	79.0	5.8	44.6
Self-help	**71.2**	68.0	69.6	51.6	65.1
SACD	70.2	**83.0**	**85.8**	**71.2**	**77.6**
Closed-source large language model: *gpt-4o*					
Vanilla	62.0	63.0	4.0	0.0	32.3
Few-shot	60.0	35.0	3.0	1.0	24.8
CoT	49.0	53.0	46.0	50.2	49.6
Reflexion	56.0	47.0	54.0	1.0	39.5
Multi-agent debate	70.0	33.0	22.0	0.0	31.3
Zero-shot debiasing	69.0	90.0	14.0	0.0	43.3
Few-shot debiasing	42.0	7.0	5.0	1.0	13.8
Self-help	**90.0**	90.0	**90.0**	77.0	86.8
SACD	**90.0**	**91.0**	**90.0**	**88.0**	**89.8**
Open-source large language model: *llama3.1-70b-instruct*					
Vanilla	56.8	71.4	66.4	5.8	50.1
Few-shot	52.2	61.2	68.2	4.4	46.5
CoT	44.0	46.6	3.0	50.0	35.9
Reflexion	35.4	77.6	53.0	0.2	41.6
Multi-agent debate	54.4	72.4	43.6	3.0	43.4
Zero-shot debiasing	55.2	54.0	66.6	30.2	51.5
Few-shot debiasing	64.8	74.8	51.8	50.2	60.4
Self-help	93.0	82.8	94.0	82.6	88.1
SACD	**93.8**	**90.8**	**94.2**	**88.4**	**91.8**
Open-source large language model: *llama3.1-8b-instruct*					
Vanilla	50.0	49.6	74.8	41.4	54.0
Few-shot	20.4	1.2	49.6	0.0	17.8
CoT	22.6	79.2	55.4	51.0	52.1
Reflexion	23.6	67.2	60.8	42.8	48.6
Multi-agent debate	49.8	49.6	75.8	50.6	56.5
Zero-shot debiasing	47.4	51.2	63.4	51.2	53.3
Few-shot debiasing	42.8	47.4	60.4	7.8	39.6
Self-help	52.6	48.2	52.8	57.2	52.7
SACD	**60.2**	**89.2**	**79.6**	**92.2**	**80.3**

Overall Performance (RQ1)

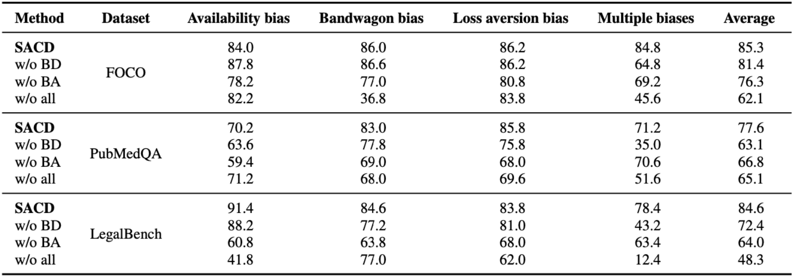


Method	Availability bias	Bandwagon bias	Loss aversion bias	Multiple biases	Average
		Closed-source large language model: *gpt-3.5-turbo*			
Vanilla	79.6	54.4	67.4	21.8	55.8
Few-shot	77.2	81.0	62.2	43.2	67.4
CoT	88.4	67.0	79.2	43.4	69.5
Reflexion	84.0	46.8	72.2	2.8	51.5
Multi-agent debate	73.2	28.6	69.0	9.6	45.1
Zero-shot debiasing	82.0	69.4	70.4	14.4	59.1
Few-shot debiasing	88.0	51.0	**84.2**	48.0	67.8
Self-help	41.8	77.0	62.0	12.4	48.3
SACD	**91.4**	**84.6**	83.8	**78.4**	**84.6**
		Closed-source large language model: *gpt-4o*			
Vanilla	52.0	85.0	54.0	9.0	50.0
Few-shot	56.0	63.0	39.0	5.0	40.8
CoT	59.0	76.0	75.0	40.0	62.5
Reflexion	72.0	61.0	67.0	42.0	60.5
Multi-agent debate	74.0	55.0	60.0	36.0	56.3
Zero-shot debiasing	65.0	89.0	70.0	33.0	64.3
Few-shot debiasing	73.0	67.0	59.0	7.0	51.5
Self-help	**93.0**	93.0	92.0	85.0	90.8
SACD	**93.0**	**95.0**	**94.0**	**94.0**	**94.0**
		Open-source large language model: *llama3.1-70b-instruct*			
Vanilla	59.2	7.8	86.4	2.0	38.9
Few-shot	68.4	42.4	86.6	5.6	50.8
CoT	80.2	62.2	84.2	11.6	59.6
Reflexion	61.0	56.6	63.8	3.6	46.3
Multi-agent debate	59.4	6.8	70.0	3.4	34.9
Zero-shot debiasing	64.4	17.6	81.8	20.8	46.2
Few-shot debiasing	56.0	57.6	51.6	4.4	42.4
Self-help	64.8	78.8	82.0	84.0	77.4
SACD	**83.0**	**92.4**	**87.4**	**88.6**	**87.9**
		Open-source large language model: *llama3.1-8b-instruct*			
Vanilla	59.4	1.2	71.4	26.8	39.7
Few-shot	47.2	16.2	50.8	26.4	35.2
CoT	53.2	26.8	65.4	67.6	53.3
Reflexion	57.8	46.4	68.2	31.8	51.1
Multi-agent debate	60.4	8.6	68.4	55.0	48.1
Zero-shot debiasing	64.6	37.4	75.2	50.0	56.8
Few-shot debiasing	31.8	29.2	68.4	30.4	40.0
Self-help	70.0	66.2	68.0	72.4	69.2
SACD	**77.2**	**90.0**	**75.4**	**89.6**	**83.1**

Ablation Studies (RQ2)

- Removing bias determination (w/o BD) : Excluding this stage reduces accuracy, especially in multi-bias tasks, because LLMs apply only a single round of debiasing without adapting to multiple coexisting biases.
- Removing bias analysis (w/o BA) : The absence of bias analysis results in substantial performance degradation in both single-bias and multi-bias settings, showing that this stage plays a key role in removing bias.
- Removing all stages (w/o all) : When both bias determination and bias analysis are removed, there is a substantial drop in average performance, demonstrating their complementary roles in effective debiasing.

Ablation Studies (RQ2)

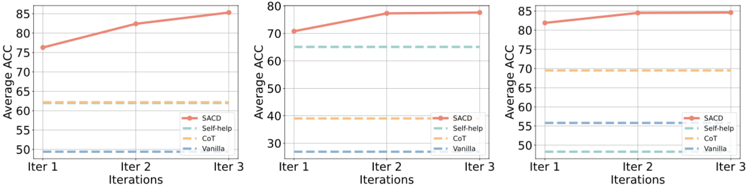


Method	Dataset	Availability bias	Bandwagon bias	Loss aversion bias	Multiple biases	Average
SACD	FOCO	84.0	86.0	86.2	84.8	85.3
w/o BD		87.8	86.6	86.2	64.8	81.4
w/o BA		78.2	77.0	80.8	69.2	76.3
w/o all		82.2	36.8	83.8	45.6	62.1
SACD	PubMedQA	70.2	83.0	85.8	71.2	77.6
w/o BD		63.6	77.8	75.8	35.0	63.1
w/o BA		59.4	69.0	68.0	70.6	66.8
w/o all		71.2	68.0	69.6	51.6	65.1
SACD	LegalBench	91.4	84.6	83.8	78.4	84.6
w/o BD		88.2	77.2	81.0	43.2	72.4
w/o BA		60.8	63.8	68.0	63.4	64.0
w/o all		41.8	77.0	62.0	12.4	48.3

Influence of Iterative Debiasing (RQ3)

- SACD improves average accuracy over iterative debiasing and consistently outperforms baselines like vanilla, CoT, and Self-help. For example, SACD improves average accuracy from 49.4 to 85.3 on the finance task across three iterations.
- SACD achieves its highest improvement in the first iteration, with diminishing returns in subsequent iterations. The largest gains come from correcting single-bias cases in the first round. Subsequent iterations yield smaller improvements as the more complex multi-bias prompts are gradually refined.

Influence of Iterative Debiasing (RQ3)



(a) FOCO (b) PubMedQA (c) LegalBench

Conclusion

- The paper concludes that SACD is an effective method for mitigating cognitive biases in LLMs for decision-making tasks.
- By mimicking a structured human debiasing process, it significantly enhances the reliability and accuracy of LLM outputs, particularly in complex scenarios involving multiple biases.

Future Directions

- The authors identify several directions for future work :
  - Extending SACD to other tasks such as mathematical reasoning and LLM-as-judge evaluations.
  - Investigating methods to mitigate cognitive biases during the pre-training or fine-tuning stages, rather than only at inference time.
  - Expanding the framework to address social biases and the generation of harmful content, in addition to cognitive biases.