UC Merced
Proceedings of the Annual Meeting of the Cognitive Science Society

Title
Balancing Rigor and Utility: Mitigating Cognitive Biases in Large Language Models for Multiple-Choice Questions

Permalink
https://escholarship.org/uc/item/2vr690cx

Journal
Proceedings of the Annual Meeting of the Cognitive Science Society, 47(0)

Authors
zhong, hanyang
Wang, Liman
CAO, WENTING
et al.

Publication Date
2025

Copyright Information
This work is made available under the terms of a Creative Commons Attribution License, available at https://creativecommons.org/licenses/by/4.0/

Peer reviewed

eScholarship.org Powered by the California Digital Library
University of California

Balancing Rigor and Utility: Mitigating Cognitive Biases in Large Language Models for Multiple-Choice Questions

Hanyang Zhong[1][1], Liman Wang[1][1], Wenting Cao[2], Zeyuan Sun[3]
[1]University of York
[2]Friedrich Schiller University Jena
[3]King's College London
{hanyang.zhong,liman.wang}@york.ac.uk, wenting.cao@uni-jena.de, zeyuan.sun@kcl.ac.uk

Abstract

This paper examines the role of cognitive biases in the decision-making processes of large language models (LLMs), challenging the conventional goal of eliminating all biases. When properly balanced, we show that certain cognitive biases can enhance decision-making efficiency through rational deviations and heuristic shortcuts. By introducing heuristic moderation and an abstention option, which allows LLMs to withhold responses when uncertain, we reduce error rates, improve decision accuracy, and optimize decision rates. Using the Balance Rigor and Utility (BRU) dataset, developed through expert collaboration, our findings demonstrate that targeted inspection of cognitive biases aligns LLM decisions more closely with human reasoning, enhancing reliability and suggesting strategies for future improvements. This approach offers a novel way to leverage cognitive biases to improve the practical utility of LLMs across various applications.

Introduction

Bias in LLMs is a critical challenge in AI research. While significant efforts have been made to address social biases embedded in training datasets (Bang, Chen, Lee, & Fung, 2024; Gallegos et al., 2024; Minaee et al., 2025), cognitive biases that emerge during inference processes remain underexplored and problematic (Suri, Slater, Ziaee, & Nguyen, 2024; P. Wang, Xiao, Chen, & Oswald, 2024; Moore, Roberts, Pham, & Fisher, 2024). These biases can mirror human cognitive tendencies, leading to flawed decision-making. Rational deviations, a concept from psychology introduced by Gerd Gigerenzer, suggest that not all biases are inherently harmful. These deviations involve heuristic thinking that simplifies decision-making, especially under uncertainty, but they can also introduce systematic errors (Gerd, 2006; Kruis, Maris, Marsman, Bolsinova, & Maas, 2020; Berthet, 2022). In LLMs, these deviations manifest in design choices that balance performance and efficiency. For example, models like GPT-4 (OpenAI et al., 2024) use context-driven token prediction to generate text, prioritizing coherence but sometimes at the cost of accuracy (Brown et al., 2020). These reasoning shortcuts mimic human cognitive biases, leading to suboptimal outcomes, making their mitigation essential for developing reliable and fair AI systems.

Our research tackles these challenges by proposing a balanced approach that integrates abstention as a rational response alongside cognitive bias mitigation techniques. We introduce heuristic moderation and an abstention mechanism, enabling LLMs to withhold decisions when uncertainty is high, reducing errors and improving accuracy. Using the BRU dataset, we demonstrate that scaling bias inspection and incorporating abstention significantly improve model performance and align LLM decision-making with human reasoning. Our result and dataset is available at: hanyangzhong.github.io/BRU-website.

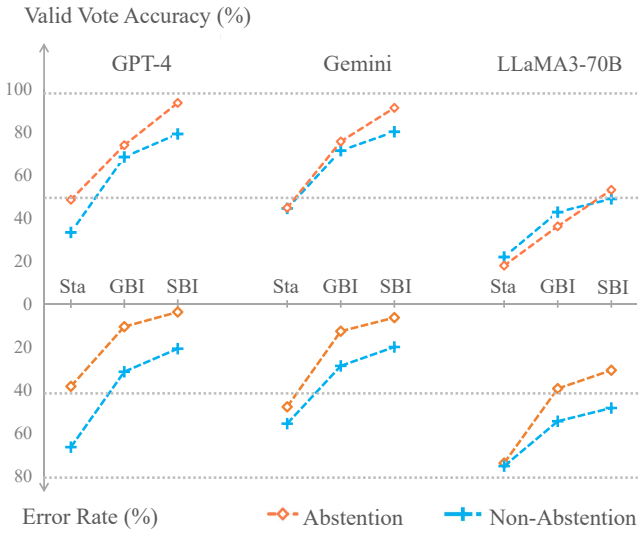

Figure 1: Valid vote accuracy and error rates on the BRU dataset for LLMs balancing rational deviations, both with and without the option to abstain. 'Sta' represents the standard baseline used for comparison, while 'GBI' and 'SBI' denote the proposed prompting strategies, as detailed in Section 4.

Related Works

Research on mitigating cognitive biases and rational deviations in LLMs is extensive. Suri et al. (Suri et al., 2024) identify human-like biases in GPT-3.5, such as anchoring and framing effects, while Bubeck et al. (Bubeck et al., 2023) and Binz and Schulz (Binz & Schulz, 2023) observe similar issues, including the framing effect and conjunction fallacy, in GPT-4 and GPT-3. Wang et al. (P. Wang et al., 2024) show that LLMs often rely on stereotypes over statistical reasoning, reflecting the representativeness heuristic. Advance-

[1]These authors contributed equally to this work and share first authorship.

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society* ©2025 the author(s). This work is licensed under a Creative Commons Attribution 4.0 International License (CC BY).

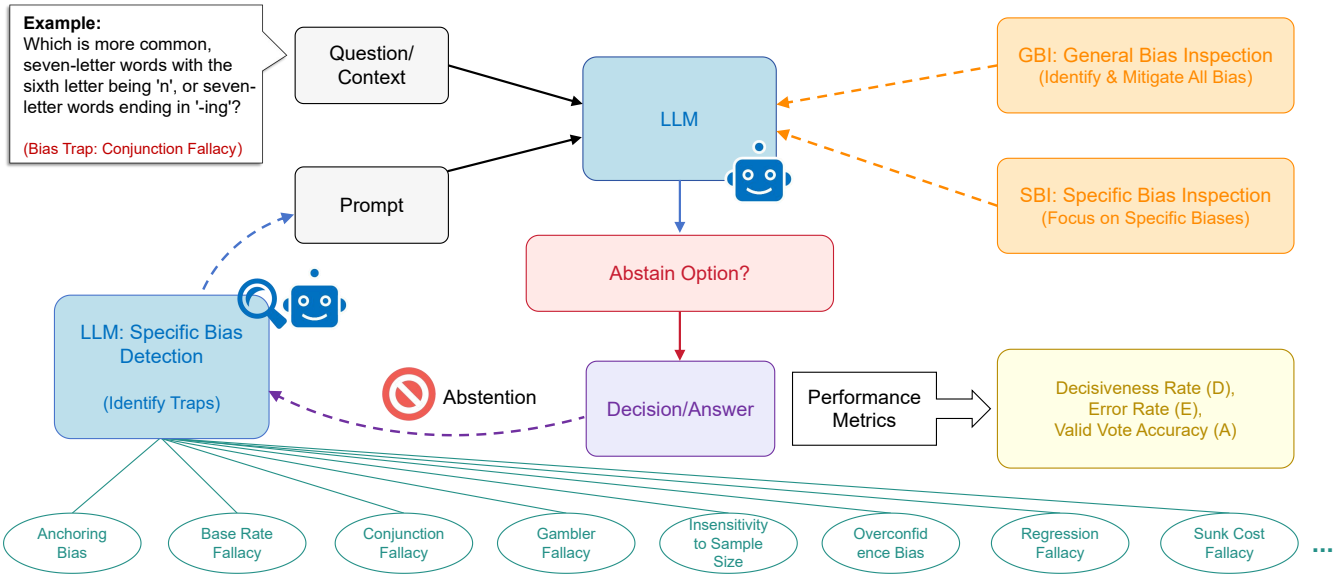

Figure 2: QA examples from GPT-4. The Conjunction Fallacy is a subset of cognitive biases. **Scaling the scope of bias inspection can influence rational deviations,** thereby impacting the outcomes of LLMs' reasoning. To address this, we propose a feedback loop **Bias Detection** module to identify the type of bias and adjust the inspection scope when an abstention from answering is considered. This approach ensures that LLMs provide more accurate responses by systematically addressing biases during decision-making. The detailed demonstration of the whole workflow is shown in Appendix Table 18-21.

ments suggest multi-modal and multi-agent systems (Jiang et al., 2024) as promising solutions for enhancing LLM rationality. Grounding decisions in broader sensory contexts, as demonstrated by Awadalla et al. (Awadalla et al., 2023) and Bai et al. (Bai et al., 2023), helps reduce hallucinations and improve decision-making. Yang et al. (Yang, Chen, Li, Wang, & Yang, 2024) and Wu et al. (Wu, Lu, Sabharwal, & Mottaghi, 2022) highlight external knowledge sources to address model limitations, while Du et al. (Du, Li, Torralba, Tenenbaum, & Mordatch, 2024) and Cohen et al. (Cohen, Hamri, Geva, & Globerson, 2023) emphasize consensus and debate mechanisms in multi-agent systems. Efforts to improve LLM reliability also stress robust evaluation metrics, as discussed by Echterhoff et al. (Echterhoff, Liu, Alessa, McAuley, & He, 2024) and Wang et al. (S. Wang, Long, Fan, Huang, & Wei, 2025). Wang et al. (H. Wang, Zhao, Qiang, Qin, & Liu, 2024) further critique multiple-choice question answering (MCQA) benchmarks and propose the MCQA+ dataset for better performance assessment. Building on this foundation, our research bridges the gap between rational deviations and cognitive biases by integrating heuristic moderation with abstention strategies, improving decision accuracy and robustness in LLM reasoning and prediction.

Methodology

Phenomenon and Method Formation

Evaluating LLMs through multiple-choice questions (MCQs) poses a key challenge: traditional metrics force definitive choices, even under uncertainty, ignoring the nuanced decision-making influenced by cognitive biases. These biases can aid efficiency via heuristics but also cause systematic errors (P. Wang et al., 2024). Accuracy metrics often fail to differentiate between confident correct answers and correct guesses, obscuring a model's true capabilities and skewing performance evaluations. For instance, high accuracy may stem from guessing rather than informed decisions, raising concerns about reliability in real-world use (H. Wang et al., 2024). To address this, we propose new metrics that better capture LLM decision-making complexities:

- **Decisiveness Rate** (D)**:** Captures the model's willingness to commit to an answer, providing insight into how often it prefers to abstain rather than make an uninformed guess.
- **Error Rate** (E)**:** Unlike traditional metrics that simply count correct answers, this metric focuses on the frequency of incorrect choices, particularly under conditions of uncertainty.
- **Valid Vote Accuracy** (A)**:** Measures the accuracy of the model's decisions when it chooses to answer, thereby highlighting its reliability in situations where it has confidence.

These new metrics aim to offer a more accurate and nuanced evaluation of LLM performance, balancing reducing error rates with improving decision accuracy. By allowing models to abstain when uncertain, we can better align LLM decision-making with human-like reasoning, ultimately enhancing their reliability and trustworthiness in practical ap-

plications (Madhusudhan, Madhusudhan, Yadav, & Hashemi, 2024).

Strategic Abstention

Abstaining is crucial for mitigating cognitive biases in LLMs during MCQs, enhancing accuracy and fairness. Forcing answers under uncertainty can lead to biased and erroneous responses, especially with skewed training data. Enabling abstention helps models avoid decisions that exacerbate biases, improving reliability (Madhusudhan et al., 2024). This is particularly important in complex reasoning tasks, where biases are more influential. Traditional metrics rewarding confident answers without considering uncertainty can reinforce these biases. Abstention mechanisms allow models to withhold responses when detecting bias or low confidence (Balabanov & Linander, 2024; Chen et al., 2023). In critical fields like healthcare and finance, abstention reduces risks from biased or harmful answers, enhancing reliability. By focusing on confident responses, LLMs improve task performance. Techniques like Strict Prompting and Chain-of-Thought (CoT) effectively support abstention, balancing rigor and utility (Madhusudhan et al., 2024; Wei et al., 2022). Our experiments show that incorporating abstention improves LLM decision-making in uncertain or ambiguous scenarios, ensuring fairness and trust, especially in high-stakes environments.

Scaling the Inspection Scope

Research shows that providing LLMs with cues about cognitive biases can improve accuracy in biased MCQs (Echterhoff et al., 2024). However, improper scaling of corrective behavior hinders balanced application of rational deviations. Correcting biases is akin to adjusting focus: precision is key, as overcorrection or under-correction reduces effectiveness. Our findings indicate that LLMs become more cautious when defining broad concepts like "Cognitive Bias," often opting to "abstain," which slightly boosts accuracy but reduces decisiveness. In contrast, defining specific concepts like "Conjunction Fallacy" makes the model less cautious and more decisive. As shown in Fig. 2, scaling the self-inspection scope appropriately ensures a balanced approach to managing cognitive biases, improving both accuracy and decisiveness.

Feedback Loop with Bias Detection

As shown in Fig. 2, the "Bias Detection Module" is an additional component designed to enhance the decision-making accuracy of LLMs by identifying cognitive bias traps embedded in MCQs. This module leverages the advanced reasoning capabilities of GPT-4o to detect biases that could influence the model's responses. The detection process is outlined in Algorithm 1.

The process starts with GPT-4o detecting potential biases in the question and prioritizing the subtype most likely to impact decision-making. This requires a robust model like GPT-4o to identify subtle biases effectively. A feedback loop iteratively refines decision-making by integrating bias detection into responses. The loop continues until a bias is found or a maximum number of iterations is reached, enabling dynamic adjustment and uncovering less obvious biases.

Algorithm 1 Feedback Loops with Bias Detection

Initialize: $Bias \leftarrow None$, $LoopCount \leftarrow 0$
$Decision \leftarrow \text{ANSWERMODEL}(MCQ)$
while $LoopCount < MaxLoops$ **and** $Decision = Abstain$ **do**
  $Bias \leftarrow \text{DETECTMODEL}(MCQ)$
  $Decision \leftarrow \text{ANSWERMODEL}(MCQ, \text{Bias})$
  $LoopCount \leftarrow LoopCount + 1$
end while
Output: $Decision$

Misjudgment of Probability	Errors in Judgment
Base Rate Fallacy (40)	Regression Fallacy (35)
Conjunction Fallacy (15)	Anchoring Bias (20)
Insensitivity to Sample Size (30)	Overconfidence Bias (30)
Gambler's Fallacy (20)	Sunk Cost Fallacy (15)

Table 1: Categories and Quantities in the BRU Dataset

A limitation is that "abstention" is never a correct ground truth answer. If the ANSWERMODEL selects "abstention," the loop may continue until the iteration limit, delaying a valid answer. To address this, we restrict the loop to one iteration, allowing "abstention" as an option while refining the output without indefinite looping. Future work could explore multi-iteration loops and nuanced modifications to improve adaptive bias detection.

Dataset and Experimental Setup

Dataset Setup

The BRU dataset used in this study includes 205 MCQs, designed to comprehensively test cognitive biases in language models. Unlike datasets like MMLU (Hendrycks et al., 2021), which cover broad question categories but have limited relevance to cognitive biases, or TruthfulQA (Lin, Hilton, & Evans, 2022) and PIQA (Bisk, Zellers, Bras, Gao, & Choi, 2020), which focus on factual correctness and commonsense reasoning, the BRU dataset addresses a wider range of cognitive biases.

Developed by an experienced psychologist with input from a medical data expert for reliability, the dataset was optimized by NLP specialists for clarity in testing LLMs. Each bias subcategory is well-documented with detailed descriptions, as shown in Table 1 and Appendix A (Tables 9-12, Figs. 1 and 5), ensuring transparency and traceability. These resources provide comprehensive insights into the dataset's composition and question design.

Models and Prompting

In our study, we evaluate the performance of three LLMs: GPT-4 (OpenAI et al., 2024), Gemini 1.0 Pro (Pichai & Hassabis, 2023), and LLaMA3-70B (Meta, 2024). To thoroughly

assess their capabilities and decision-making processes, we employ various prompting techniques to mitigate or encourage heuristic thinking. These techniques are crucial for understanding how LLMs navigate cognitive tasks and biases.

Abstention Prompting Abstention prompting allows the model to refrain from making a decision when uncertain:

> *If you prefer not to make a decisive choice, then select option E.*
> *E: I am not sure which choice is the best to select.*

This reduces incorrect answers by avoiding guesses and improving accuracy by encouraging the model to make decisions only when confident. The option E is designed to aid the final result statistics.

Non-Abstention Prompting This prompting forces the model to make a choice:

> *You can only choose one option.*

This evaluates the LLM's performance under pressure by requiring it to make decisions even when uncertain. It assesses the model's ability to handle forced-choice scenarios, revealing its strategies for making informed guesses despite uncertainty.

General Bias Inspection Consider the set of all cognitive biases, denoted as $B = \{b_1, b_2, b_3, \ldots, b_n\}$, where each element b_i represents a distinct cognitive bias. General Bias Inspection (GBI) involves a comprehensive review of the entire set B to identify and mitigate any cognitive bias in decision-making. This approach ensures that the model is broadly aware of the full spectrum of cognitive biases, enabling it to self-check for potential bias influences. By reflecting on the general concept of cognitive bias, defined as any systematic deviation from rational judgment, GBI promotes overall bias mitigation by considering the union of all possible biases $\cup B$:

> *Please provide a definition of cognitive bias and identify any instances of these biases in the decision-making process.*

Specific Bias Inspection Specific Bias Inspection (SBI) focuses on a particular subset of cognitive biases, denoted as $S \subseteq B$, where S represents a specific group of related biases relevant to the current context or question. For example, if $S = \{b_3\}$ represents biases related to probability misjudgments, such as the Base Rate Fallacy and the Gambler's Fallacy, SBI involves a focused analysis on this subset S. This targeted approach allows the model to concentrate on the most pertinent biases for a given scenario, enhancing accuracy by mitigating the influence of specific biases $\cup S$ rather than the entire set B:

> *Please provide a definition of the Base Rate Fallacy, then identify any instances of this specific bias in the decision-making process.*

This targeted approach enables more precise and context-sensitive bias analysis.

Bias Detection Module To further enhance the model's ability to detect potential "bias traps" in questions, we introduce a dedicated prompt for the Bias Detection Module:

> *Please identify which cognitive bias trap is contained in this question and return the cognitive bias type. The most likely cognitive bias trap is _.*

This prompt enables the model to identify specific cognitive biases embedded in each problem. Although the model may list all relevant bias subtypes, each question is designed to target a particular bias subtype, creating a hierarchy of relevance. In our experiments, we select the highest-priority bias subtype as the output, ensuring that the most significant cognitive bias is accurately identified and addressed.

Evaluation Criteria

In evaluating the reasoning outcomes of LLMs, we categorize the reasoning process and results using the notations TT, TF, FT, and FF. These represent the following scenarios: a correct reasoning process with a correct result (TT), a correct reasoning process with an incorrect result (TF), an incorrect reasoning process with a correct result (FT), and an incorrect reasoning process with an incorrect result (FF). The symbol O denotes instances of "abstention," where the model selects option E. For the statistical analysis, the Decisiveness Rate D is defined as follows:

$$D = \frac{N_{total} - N_O}{N_{total}}$$

where N_O represents the number of abstained questions and N_{total} represents the total number of questions in the BRU dataset. The Error Rate E is defined as:

$$E = \frac{N_{FF} + N_{TF}}{N_{total} - N_O}$$

Here, N_{FF} and N_{TF} denote the number of questions with incorrect results. The Valid Vote Accuracy A is given by:

$$A = \frac{N_{TT} + N_{FT}}{N_{total} - N_O}$$

where N_{TT} and N_{FT} indicate the number of questions with correct results. It should be noted that the ground truth answers are manually annotated, excluding the reasoning process content. The accuracy of the LLM's reasoning is assessed through a manual review of the dialogue context.

Model / Accuracy *A*	+ Non-Abstention			+ Abstention		
	Standard	GBI	SBI	Standard	GBI	SBI
GPT-4	33.2	68.3 (+35.1)	79.0 (**+45.8**)	48.4	73.8 (+25.4)	**93.5** (+45.1)
Gemini 1.0 Pro	44.4	71.2 (+26.8)	**80.0** (+35.6)	44.6	75.5 (+30.9)	91.1 (**+46.5**)
LLaMA3-70B	22.4	43.9 (+21.5)	50.2 (+27.8)	18.3	37.1 (+18.8)	54.8 (+36.5)

Model / Error Rate *E*	+ Non-Abstention			+ Abstention		
	Standard	GBI	SBI	Standard	GBI	SBI
GPT-4	66.8	31.7 (-35.1)	21.0 (**-45.8**)	38.5	10.7 (-27.8)	**3.9** (-34.6)
Gemini 1.0 Pro	55.6	28.8 (-26.8)	**20.0** (-35.6)	47.8	12.7 (-35.1)	6.3 (-41.5)
LLaMA3-70B	77.6	56.1 (-21.5)	49.8 (-27.8)	76.1	40.5 (-35.6)	31.7 (**-44.4**)

Table 2: Prediction accuracy and error rate of GPT-4, Gemini 1.0 Pro, and LLaMA3-70B in Non-Abstention and Abstention experiments (%) on the BRU dataset with different prompting strategies. Bold numbers indicate the relative extrema. Differences between Standard groups with and without abstention are shown with ± values in black.

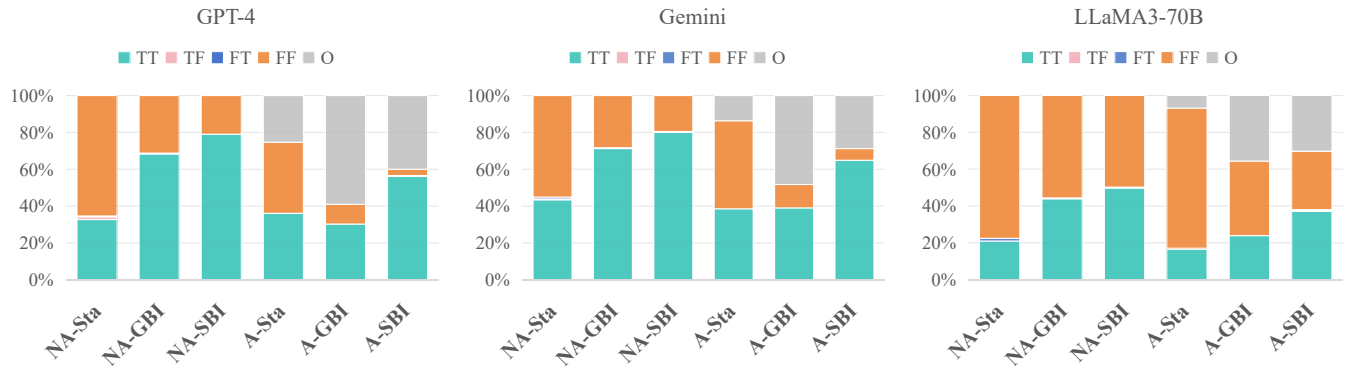

Figure 3: The combination of TT, TF, FT, FF, and O rates for GPT-4, Gemini 1.0 Pro, and LLaMA3-70B on the BRU dataset using different prompting strategies. 'NA-' denotes Non-Abstention, 'A-' denotes Abstention, and 'Sta' represents the Standard used for comparison. The detailed distributions of the TT, TF, FT, FF, and O rates for GPT-4, Gemini 1.0 Pro and LLaMA3-70B are elaborately listed in Appendix Tables 5, 6 and 7.

Experimental Results

Scaling Bias Inspection Effect Findings

In the context of Non-Abstention, with "Standard" serving as the control group, significant improvements in *A* score are observed for GPT-4, Gemini 1.0 Pro, and LLaMA3-70B on the BRU dataset when employing the prompting techniques GBI and SBI, as shown in Table. 2. Notably, the *A* score enhancement is more pronounced with SBI. Specifically, GPT-4, Gemini 1.0 Pro, and LLaMA3-70B achieve *A* score of 79%, 80%, and 50.2%, respectively, when utilizing SBI in Non-Abstention. This underscores the effectiveness of the GBI and SBI prompting strategies. The heightened accuracy with SBI, attributed to its narrower and more targeted cognitive bias scope, supports our hypothesis that specific guidance in bias inspection can significantly boost the accuracy of LLM responses.

Abstention Effect Findings

As shown in Table 2, the abstention effect reveals significant shifts in model performance. (see Appendix Tables 3, 4). Introducing the abstention option generally boosts the *A* score but reduces the decision number, as depicted by the grey bars in Fig. 3. For GPT-4, the *A* score improves from 33.2% to 48.4%, indicating effective avoidance of incorrect answers through abstention. Gemini's *A* score also sees a slight increase from 44.4% to 44.6%. Conversely, LLaMA3-70B's *A* score drops from 22.4% to 18.3%, highlighting its weaker decision-making capabilities. *E* score decreases notably with abstention: GPT-4's drops from 66.8% to 38.5%, and Gemini 1.0 Pro's from 55.6% to 47.8%. LLaMA3-70B experiences a minor reduction from 77.6% to 76.1%. These results suggest that abstention reduces errors and leads to fewer decisions.

Combination Test Findings

Before conducting the combination tests, we assessed whether the Abstention technique offered greater benefits than GBI and SBI. As shown in Table 2, the improvements from abstention (highlighted in green) are generally less significant than those from GBI and SBI. However, GPT-4 gains more from abstention rights than Gemini 1.0 Pro and LLaMA3-70B. Under the Abstention condition, GPT-4 performs best with SBI, while Gemini 1.0 Pro peaks with SBI

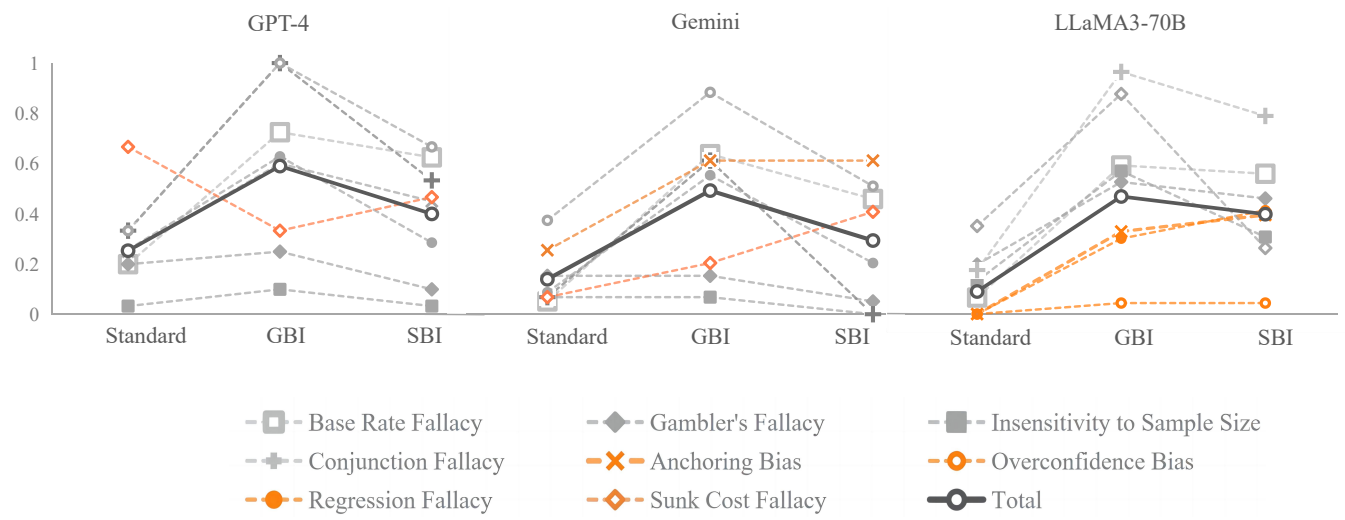

Figure 4: Distribution chart of abstention rates for GPT-4, Gemini 1.0 Pro, and LLaMA3-70B across different question types in the BRU dataset with Abstention enabled and using different prompting strategies.

under the Non-Abstention condition. Fig. 3 also shows that GPT-4 abstains more frequently than Gemini 1.0 Pro, adopting a "fewer decisions, fewer mistakes" strategy. Although this approach indicates emergent intelligence, excessive abstention has potential risks, which will be discussed further. In Fig. 3, the green bars represent the proportion of correct answers, which increases with GBI and SBI but decreases with Abstention, consistent with earlier findings. Conversely, the *E* decreases when Abstention, GBI, or SBI are introduced. Among the tested combinations, **Abstention+SBI** allows all three LLMs to achieve the highest *A* score of **93.5%** and the lowest *E* score of **3.9%**, as shown in Table 2. Notably, GPT-4 and Gemini 1.0 Pro using Abstention+SBI achieve near-zero *E* scores of **3.9%** and **6.3%**, respectively, demonstrating the value of minimizing decision risk.

Bias Detection Loop by GPT-4o

Using GPT-4o for bias detection in 205 questions, the model accurately identified the specific bias subtype 65% of the time. In an additional 15% of cases, it matched a parent category or synonym of the bias subtype (details in Appendix Table 8) , resulting in a cumulative **80%** recognition rate. This demonstrates the module's ability to align closely with actual biases, effectively bridging GBI and SBI methods. This high recognition rate enables dynamic adjustment of the bias detection scope, enhancing decision accuracy and minimizing errors. By refining bias recognition, the model seamlessly transitions from broader GBI to precise SBI, improving *D* scores and reducing *E* scores.

Abstention Patterns for Question Types

When LLMs are allowed to abstain, their abstention rates generally fluctuate across question types, as shown in Fig. 4. However, question types like the Sunk Cost Fallacy, Anchoring Bias, Overconfidence Bias, and Regression Fallacy (highlighted in orange) deviate from this trend. These Errors in Judgment, less common in training data than Misjudgment of Probability concepts, are more likely to be overlooked. Under GBI alone, LLMs may miss these traps, leading to overconfidence and lower abstention rates. With SBI, the model recognizes these biases, shifting from overconfidence to uncertainty, reflected in higher abstention rates. This highlights how identifying specific biases, such as the "Sunk Cost Fallacy," prompts LLMs to reconsider their judgments. The Bias Detection Loop module addresses this by using GPT-4o's reasoning to transition GBI abstention into automatic SBI recognition, boosting both *D* and *A* scores.

Conclusion

This study examines the roles of cognitive biases and rational deviations in LLM decision-making, showing that leveraging certain biases through heuristic moderation and strategic abstention can be beneficial. The BRU dataset reveals that SBI prompting, combined with the option to abstain from uncertain decisions, reduces errors and aligns LLM reasoning with human patterns. The findings demonstrate that integrating abstention with targeted bias inspection, like SBI, significantly enhances accuracy and reliability. Models such as GPT-4 and Gemini 1.0 Pro achieve notable improvements in decision accuracy by abstaining under uncertainty, better replicating human-like judgment. The proposed Bias Detection Loop facilitates a seamless transition from GBI to SBI, combining inspection scopes without over-relying on abstention, further improving accuracy and utility. This research underscores the value of balanced bias management in LLMs, showing that biases, when properly moderated, can enhance practical utility in applications like conversational agents and decision support systems. It establishes a foundation for further exploration of how balancing biases and rational deviations can improve LLM performance, particularly in MCQs.

References

Arkes, H. R., & Blumer, C. (1985). The psychology of sunk cost. *Organizational Behavior and Human Decision Processes*, *35*(1), 124-140.

Awadalla, A., Gao, I., Gardner, J., Hessel, J., Hanafy, Y., Zhu, W., … Schmidt, L. (2023). *Openflamingo: An open-source framework for training large autoregressive vision-language models.* Retrieved from https://arxiv.org/abs/2308.01390

Bai, J., Bai, S., Yang, S., Wang, S., Tan, S., Wang, P., … Zhou, J. (2023). *Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond.* Retrieved from https://arxiv.org/abs/2308.12966

Balabanov, O., & Linander, H. (2024). *Uncertainty quantification in fine-tuned llms using lora ensembles.* Retrieved from https://arxiv.org/abs/2402.12264

Bang, Y., Chen, D., Lee, N., & Fung, P. (2024). Measuring political bias in large language models: What is said and how it is said. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 11142–11159). Stroudsburg, PA, USA.

Berthet, V. (2022). The impact of cognitive biases on professionals' decision-making: A review of four occupational areas. *Frontiers in Psychology*, *12*, 802439. Retrieved from https://doi.org/10.3389/fpsyg.2021.802439

Binz, M., & Schulz, E. (2023). Using cognitive psychology to understand GPT-3. *Proceedings of the National Academy of Sciences*, *120*(6), e2218523120.

Bisk, Y., Zellers, R., Bras, R. L., Gao, J., & Choi, Y. (2020). PIQA: reasoning about physical commonsense in natural language. AAAI Press. Retrieved from https://doi.org/10.1609/aaai.v34i05.6239

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., … Amodei, D. (2020). Language models are few-shot learners. In *Proceedings of the 34th international conference on neural information processing systems.* Curran Associates Inc.

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., … Zhang, Y. (2023). *Sparks of artificial general intelligence: Early experiments with gpt-4.* Retrieved from https://arxiv.org/abs/2303.12712

Chen, J., Yoon, J., Ebrahimi, S., Arik, S., Pfister, T., & Jha, S. (2023). Adaptation with Self-Evaluation to Improve Selective Prediction in LLMs. In *Findings of the Association for Computational Linguistics: Emnlp 2023.* Stroudsburg, PA, USA.

Cohen, R., Hamri, M., Geva, M., & Globerson, A. (2023). LM vs LM: Detecting factual errors via cross examination. In *Proceedings of the 2023 conference on empirical methods in natural language processing* (pp. 12621–12640). Singapore: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2023.emnlp-main.778/

Du, Y., Li, S., Torralba, A., Tenenbaum, J. B., & Mordatch, I. (2024). Improving factuality and reasoning in language models through multiagent debate. In *Proceedings of the 41st international conference on machine learning.* JMLR.org.

Echterhoff, J. M., Liu, Y., Alessa, A., McAuley, J., & He, Z. (2024). Cognitive bias in decision-making with LLMs. In (pp. 12640–12653). Association for Computational Linguistics. Retrieved from https://aclanthology.org/2024.findings-emnlp.739/

Gallegos, I. O., Rossi, R. A., Barrow, J., Tanjim, M. M., Kim, S., Dernoncourt, F., … Ahmed, N. K. (2024). Bias and fairness in large language models: A survey. *Computational Linguistics*, *50*(3), 1097–1179. Retrieved from https://aclanthology.org/2024.cl-3.8/

Gerd, G. (2006). Bounded and rational. In R. Stainton (Ed.), *Contemporary debates in cognitive science* (pp. 115–133). Wiley-Blackwell.

Gilovich, T., Vallone, R., & Tversky, A. (1985). The hot hand in basketball: On the misperception of random sequences. *Cognitive Psychology*, *17*(3), 295-314.

Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., & Steinhardt, J. (2021). Measuring massive multitask language understanding. In *International conference on learning representations.*

Jiang, B., Xie, Y., Wang, X., Su, W. J., Taylor, C. J., & Mallick, T. (2024). Multi-modal and multi-agent systems meet rationality: A survey. In *Icml 2024 workshop on llms and cognition.*

Kahneman, D. (2011). *Thinking, fast and slow*. macmillan.

Kruis, J., Maris, G., Marsman, M., Bolsinova, M., & Maas, H. (2020). Deviations of rational choice: an integrative explanation of the endowment and several context effects. *Scientific Reports*, *10*.

Lin, S., Hilton, J., & Evans, O. (2022). TruthfulQA: Measuring how models mimic human falsehoods. Dublin, Ireland: Association for Computational Linguistics.

Madhusudhan, N., Madhusudhan, S. T., Yadav, V., & Hashemi, M. (2024). *Do llms know when to not answer? investigating abstention abilities of large language models.* Retrieved from https://arxiv.org/abs/2407.16221

Meta. (2024). *Introducing Meta Llama 3: The most capable openly available LLM to date.* https://ai.meta.com/blog/meta-llama-3.

Minaee, S., Mikolov, T., Nikzad, N., Chenaghlu, M., Socher, R., Amatriain, X., & Gao, J. (2025). *Large language models: A survey.* Retrieved from https://arxiv.org/abs/2402.06196

Moore, K., Roberts, J., Pham, T., & Fisher, D. (2024). *Reasoning beyond bias: A study on counterfactual prompting and chain of thought reasoning.* Retrieved from https://arxiv.org/abs/2408.08651

OpenAI, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., … Zoph, B. (2024). *Gpt-4 technical report.*

Retrieved from https://arxiv.org/abs/2303.08774

Pichai, S., & Hassabis, D. (2023). *Introducing gemini: our largest and most capable ai model.* Retrieved from https://blog.google/technology/ai/google-gemini-ai (Accessed: April 9, 2025)

Suri, G., Slater, L., Ziaee, A., & Nguyen, M. (2024). Do large language models show decision heuristics similar to humans? a case study using gpt-3.5. *Journal of Experimental Psychology: General*, *153*, 1066-1075.

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

Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review*, *90*(4), 293–315. doi: 10.1037/0033-295X.90.4.293

Wang, H., Zhao, S., Qiang, Z., Qin, B., & Liu, T. (2024). Beyond the answers: Reviewing the rationality of multiple choice question answering for the evaluation of large language models. *CoRR*, *abs/2402.01349*. Retrieved from https://doi.org/10.48550/arXiv.2402.01349

Wang, P., Xiao, Z., Chen, H., & Oswald, F. L. (2024). Will the real linda please stand up...to large language models? examining the representativeness heuristic in LLMs. In *First conference on language modeling.* Retrieved from https://openreview.net/forum?id=3GhOWfSLrD

Wang, S., Long, Z., Fan, Z., Huang, X., & Wei, Z. (2025). Benchmark self-evolving: A multi-agent framework for dynamic LLM evaluation. In *Proceedings of the 31st international conference on computational linguistics* (pp. 3310–3328). Abu Dhabi, UAE: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2025.coling-main.223/

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., … Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th international conference on neural information processing systems.* Red Hook, NY, USA: Curran Associates Inc.

Wu, J., Lu, J., Sabharwal, A., & Mottaghi, R. (2022). Multimodal answer validation for knowledge-based vqa. *Proceedings of the AAAI Conference on Artificial Intelligence*.

Yang, Z., Chen, G., Li, X., Wang, W., & Yang, Y. (2024). Doraemongpt: toward understanding dynamic scenes with large language models (exemplified as a video agent). In *Proceedings of the 41st international conference on machine learning.* JMLR.org.