Towards Detecting, Mitigating and Explaining Biased and Fallacious Reasoning in Large Language Models

Doctoral Consortium

Ana Gutiérrez-Mandingorra
Universitat Politècnica de València (UPV)
Camí de Vera s/n 46022, Valencia, Spain
agutman@upv.es

ABSTRACT

The proliferation of generative AI in the era society is currently living intensifies concerns about disinformation, bias and fallacious reasoning. Large Language Models (LLMs), while capable of generating coherent text, may reproduce systematic errors inherent in human cognition, often lacking a necessary logical layer. In this paper, I detailed ongoing research that proposes an interdisciplinary framework integrating artificial intelligence, computational argumentation (CA), cognitive science and ethics to detect, mitigate and explain biased or fallacious reasoning.

KEYWORDS

LLMs; Computational Argumentation; Disinformation; Cognitive Bias; Fallacious Reasoning; Responsible AI

ACM Reference Format:
Ana Gutiérrez-Mandingorra. 2026. Towards Detecting, Mitigating and Explaining Biased and Fallacious Reasoning in Large Language Models: Doctoral Consortium. In *Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), Paphos, Cyprus, May 25 – 29, 2026*, IFAAMAS, 3 pages. https://doi.org/10.65109/GNAS4540

1 INTRODUCTION

The integration of LLMs into critical sectors has established them as core components of modern artificial intelligence. However, their widespread deployment raises concerns about ethics, reliability and societal impact, as they may perpetuate disinformation, bias and fallacious reasoning.

Disinformation does not always take the form of fake news deliberately created to manipulate; it often relies on logical fallacies and rhetorical manipulation, which are themselves rooted in cognitive biases (CBs)—systematic deviations from normative reasoning shaped by human heuristics. Biases in LLMs emerge from complex social and cognitive factors that can appear at any stage of the machine learning pipeline, from data collection to deployment [15]. While social biases (e.g., gender or race) have been widely studied [8, 17], CBs remain comparatively underexplored. These mental shortcuts enable rapid decision-making but frequently lead to systematic reasoning errors [2] and may underlie many social biases [10].

Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), C. Amato, L. Dennis, V. Mascardi, J. Thangarajah (eds.), May 25 – 29, 2026, Paphos, Cyprus. © 2026 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org). https://doi.org/10.65109/GNAS4540

Understanding CBs requires grounding in cognitive science, particularly in the Dual Process Theory [3], which distinguishes between System 1 (fast, heuristic and prone to fallacies [18]) and System 2 (slow, deliberate and analytical) reasoning. NLP researchers have drawn parallels between System 1 and zero-shot prompting, while chain-of-thought prompting reflects System 2 reasoning through explicit, stepwise deliberation [8]. However, LLMs fundamentally rely on pattern recognition rather than genuine understanding; they assess surface structure rather than the logical validity of arguments [4]. As a result, they risk replicating human cognitive flaws characteristic of System 1 and amplifying misinformation, with measurable impacts such as a 26% reduction in diagnostic accuracy in medical LLMs attributed to CBs [14].

Efforts to detect and mitigate CBs remain limited by the inherent complexity of these biases and by the narrow scope of current benchmarks [9, 14], which typically cover only a subset of bias types. Existing approaches—including prompt-based strategies [14], chain-of-thought with Human Persona [8] and context injection [20]—offer partial solutions but lack generalizability.

This doctoral research hypothesizes that strengthening argumentative and explanatory capacities can improve the detection and mitigation of fallacious reasoning. Rather than eliminating biases internally, it focuses on enhancing robustness at the interaction level to foster more critical, responsible and transparent generative systems. Two main research milestones have been achieved.

2 DISINFORMATION THROUGH ARGUMENT SCHEMES

The first research milestone [5], inspired by the work of [11, 12], addressed the challenge of integrating a formal reasoning layer into LLMs to counter disinformation, which frequently manifests through argumentative fallacies. The underlying hypothesis was that CA techniques—particularly the use of Argumentation Schemes (AS) and their associated Critical Questions (CQs)—could guide LLMs to assess the logical soundness and veracity of arguments by questioning their underlying structure. This approach combined CA with LLMs enhanced through external contextualization. A modified version of the NLAS-MULTI corpus [13] was employed, comprising 19 argumentation schemes although an additional "*no scheme*" class was introduced to capture arguments that did not fit any predefined category, improving the system's adaptability to real-world discourse. The implemented web-based tool was structured into two interconnected modules:

Module 1: AS Classifier. This component classified input arguments into one of 20 categories using the open-source RASA [1]

conversational framework. An initial single-layer classifier achieved suboptimal results, leading to the adoption of a two-layer architecture inspired by Walton's taxonomy of AS [19], which groups schemes into Source-Based, Rule-Application and Reasoning arguments. The first layer identified the argument group and the second refined the classification into a specific scheme. This hierarchical approach consistently outperformed the single-layer model, achieving robust performance in the AS classification task, with accuracy and F1-scores ranging between 0.85 and 0.87.

Module 2: Veracity Level Generator and Evaluator System. This module employed a quantized LLAMA 3 70B model augmented with external contextualization. Upon detecting the AS, the system generated its corresponding CQs, executed parallel context searches using external sources (Google, Wikipedia and Bing) and synthesized the retrieved information. The model then acted as an expert assistant in computational argumentation, producing both quantitative and qualitative justifications for each argument's truthfulness. Human evaluation (80 participants aged 18–65), indicated that 83.8% of users rated the system's responses as satisfactory in terms of coherence and adequacy of sources.

3 COGNITIVE BIAS IN LLMS

The second milestone explored the subtle domain of CBs in LLMs, focusing on their impact, detection and mitigation [6]. An experimental framework was developed, structured into three modules.

Module 1: Evaluating CBs in LLM Outputs. This module examined how prompt-induced CBs affect LLM accuracy and consistency. Three well-documented biases—acquiescence bias, availability bias and the bandwagon effect (the latter two documented in the Cognitive Bias Codex [7])—were analyzed across state-of-the-art models (LLaMA 3.2, LLaMA 3.3:70B, Qwen 2.5, DeepSeek-V2 and GPT-4o). Experiments were conducted using the maveriq/bigbench-hard dataset [16], where each original item was expanded into four variants: one unbiased and three biased according to the selected cognitive patterns. Each original item consisted of a question paired with a binary (YES/NO) answer. The biased variants were generated using phrase templates, with each template designed to induce a specific bias. Bias was introduced through two mechanisms: (1) direct prompt modification in a single-step and (2) post-response bias induction via follow-up interaction, prompting the model to reconsider its answer in 2 steps. Results showed consistent performance degradation under biased conditions, especially for acquiescence bias and in the two-step condition. For instance, LLaMA 3.3:70B showed a marked decline in accuracy on acquiescence bias, decreasing from 0.85 to 0.66 in single-step interactions and from 0.84 to 0.16 in two-step interactions.

Module 2: Detection and Classification of CBs. This module compared native reasoning models (QwQ and DeepSeek-V1) with a general-purpose LLM (LLaMA 3.3:70B) configured as a ReAct-Reasoning and Acting-agent [21]. All systems were enhanced with a Retrieval-Augmented Generation (RAG) system grounded in cognitive theory knowledge[1]. In binary classification (biased vs. unbiased), the QwQ model achieved the highest F1-score, closely followed by the LLaMA ReAct agent, confirming that general-purpose LLMs equipped with reasoning modules can approach the performance of dedicated reasoning models. However, in the more complex multiclass setting, QwQ outperformed all models (F1-scores: unbiased 0.65, acquiescence 0.12, availability 0.91, bandwagon 1.00). All models struggled to distinguish acquiescence bias, often misclassifying it as unbiased.

Module 3: Mitigating CBs in LLMs. Based on the best- performing configuration (QwQ), a mitigation framework was designed to generate context-aware warning messages. For each detected bias, the model produced a brief explanation encouraging critical reflection, which was then appended to the biased prompt. Incorporating these warnings into the input led to substantial accuracy improvements across models and in several cases (GPT-4o and LLama 3.2) performance even surpassed the unbiased baseline. These results suggest that explicit bias warnings can trigger more deliberative, System 2-like reasoning in LLMs, enhancing both accuracy and interpretive robustness.

4 FUTURE DIRECTIONS

This research establishes an initial foundational step toward the development of ethical and cognitively informed AI systems. Future work will aim to scale the current experiments to more complex and realistic settings, further refining strategies for detecting and mitigating fallacious reasoning in LLMs. A central direction involves the integration of multi-agent architectures composed of specialized expert agents capable of collective reasoning, voting and ethical deliberation to improve bias detection and veracity assessment. These multi-agent systems will also be employed in simulation environments designed to model complex processes of reasoning and error detection, enabling the study of argumentative dynamics and ethical interactions among agents.

Another promising direction is to extend the use of CA techniques—successfully applied in disinformation detection—to the domain of CBs analysis. AS and CQs could serve as structured mechanisms for identifying or challenging biased reasoning, potentially providing explainable and logically grounded mitigation strategies. Additionally, future experiments will broaden the scope of analysis to a wider range of cognitive biases beyond the initial three, while examining how such biases evolve and propagate within dynamic, multi-turn conversational contexts. As observed in our experiments, bias introduced through multi-step interactions has a stronger influence on LLM behavior, underscoring the need for deeper investigation into conversational bias amplification.

Finally, exploring the relationship between human values and CBs offers a promising avenue for understanding how moral and psychological dimensions shape reasoning in both humans and artificial systems. Such exploration may ultimately contribute to the creation of AI models capable of more reflective, value-aligned and transparent decision-making.

5 AKNOWLEDGMENTS

This work was partially supported by project PID2024-158227NB-C33 funded by MICIU/AEI/10.13039/501100011033/ FEDER, UE, and by the Valencian Government through grant CIPROM/2021/077.

[1] https://github.com/scottleedavis/cognitive-bias-codex/tree/master

REFERENCES

[1] Tom Bocklisch, Joey Faulkner, Nick Pawlowski, and Alan Nichol. 2017. Rasa: Open source language understanding and dialogue management. *arXiv preprint arXiv:1712.05181* (2017).

[2] Alexander Brem and Giorgia Rivieccio. 2024. Artificial Intelligence and Cognitive Biases: A Viewpoint. *Journal of Innovation Economics & Management* 44, 2 (2024), 223–231.

[3] Gerd Gigerenzer and Peter M Todd. 1999. *Simple heuristics that make us smart*. Oxford University Press, USA.

[4] Pierpaolo Goffredo, Mariana Espinoza, Serena Villata, and Elena Cabrio. 2023. Argument-based Detection and Classification of Fallacies in Political Debates. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11101–11112.

[5] Ana Gutierrez, Stella Heras, and Javier Palanca. 2024. Detecting Disinformation through Computational Argumentation Techniques and Large Language Models.. In *CMNA@ COMMA*. 46–51.

[6] Ana Gutiérrez, Stella Heras, Javier Palanca, and Vicente Botti. 2026. Exploring Cognitive Bias Impact, Detection and Mitigation in Large Language Models. 25th International Conference on Autonomous Agents and Multiagent Systems. To appear.

[7] John Manoogian III. 2024. Cognitive Bias Codex - 180+ biases, Wikipedia. https://w.wiki/ByPr

[8] Mahammed Kamruzzaman and Gene Louis Kim. 2024. Prompting techniques for reducing social bias in llms through system 1 and system 2 cognitive processes. *arXiv preprint arXiv:2404.17218* (2024).

[9] Ryan Koo, Minhwa Lee, Vipul Raheja, Jong Inn Park, Zae Myung Kim, and Dongyeop Kang. 2023. Benchmarking cognitive biases in large language models as evaluators. *arXiv preprint arXiv:2309.17012* (2023).

[10] Naroa Martínez, Ujué Agudo, and Helena Matute. 2022. Human cognitive biases present in Artificial Intelligence. *Revista Internacional de los Estudios Vascos* 67, 2 (2022).

[11] Ramon Ruiz-Dolz and John Lawrence. 2023. Detecting argumentative fallacies in the wild: Problems and limitations of large language models. In *Proceedings of the 10th Workshop on Argument Mining*. Association for Computational Linguistics.

[12] Ramon Ruiz-Dolz and John Lawrence. 2025. An explainable framework for misinformation identification via critical question answering. *arXiv preprint arXiv:2503.14626* (2025).

[13] Ramon Ruiz-Dolz, Joaquin Taverner, John Lawrence, and Chris Reed. 2024. NLAS-multi: A Multilingual Corpus of Automatically Generated Natural Language Argumentation Schemes. *arXiv preprint arXiv:2402.14458* (2024).

[14] Samuel Schmidgall, Carl Harris, Ime Essien, Daniel Olshvang, Tawsifur Rahman, Ji Woong Kim, Rojin Ziaei, Jason Eshraghian, Peter Abadir, and Rama Chellappa. 2024. Addressing cognitive bias in medical language models. *arXiv preprint arXiv:2402.08113* (2024).

[15] Harini Suresh and John Guttag. 2021. A framework for understanding sources of harm throughout the machine learning life cycle. In *Proceedings of the 1st ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*. 1–9.

[16] Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V Le, Ed H Chi, Denny Zhou, , and Jason Wei. 2022. Challenging BIG-Bench Tasks and Whether Chain-of-Thought Can Solve Them. *arXiv preprint arXiv:2210.09261* (2022).

[17] Daniel Van Niekerk, María Peréz-Ortiz, John Shawe-Taylor, Davor Orlic, Jackie Kay, Noah Siegel, Katherine Evans, Nyalleng Moorosi, Tina Eliassi-Rad, Leonie Maria Tanczer, et al. 2024. Challenging Systematic Prejudices: An Investigation into Bias Against Women and Girls. (2024).

[18] Douglas Walton. 2010. Why fallacies appear to be better arguments than they are. *Informal logic* 30, 2 (2010), 159–184.

[19] Douglas Walton, Christopher Reed, and Fabrizio Macagno. 2008. *Argumentation schemes*. Cambridge University Press.

[20] Liman Wang, Hanyang Zhong, Wenting Cao, and Zeyuan Sun. 2024. Balancing rigor and utility: Mitigating cognitive biases in large language models for multiple-choice questions. *arXiv preprint arXiv:2406.10999* (2024).

[21] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. React: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*.