Exploring Cognitive Bias Impact, Detection and Mitigation in Large Language Models

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

Stella Heras
Valencian Research Institute for Artificial Intelligence (VRAIN), Universitat Politècnica de València (UPV)
Camí de Vera s/n 46022, Valencia, Spain
stehebar@upv.es

Javier Palanca
Valencian Research Institute for Artificial Intelligence (VRAIN), Universitat Politècnica de València (UPV)
Camí de Vera s/n 46022, Valencia, Spain
jpalanca@dsic.upv.es

Vicente Botti
Valencian Research Institute for Artificial Intelligence (VRAIN), Universitat Politècnica de València (UPV)
Valencian Graduate School and Research Network of Artificial Intelligence (VALGRAI)
Camí de Vera s/n 46022, Valencia, Spain
vbotti@dsic.upv.es

ABSTRACT

Large Language Models have revolutionized a wide range of domains—including education, healthcare, law, and industry—by enabling the automation of complex tasks through advanced natural language understanding and text generation.

However, their widespread deployment has raised significant ethical and practical concerns, particularly regarding the biases embedded in their outputs. While social biases in LLMs have been extensively examined across the literature, cognitive biases—systematic patterns of deviation from normative reasoning rooted in human cognition—remain comparatively underexplored. These biases pose a unique challenge, as they can be subtly introduced, for instance, through prompt design or inherited from training data. Therefore, the study of cognitive biases in LLMs represents an emerging and increasingly critical area of research.

This work presents a structured investigation into the presence, detection, and mitigation of cognitive biases in LLMs. We propose a three-stage experimental strategy: (1) evaluating the influence of prompt-induced cognitive biases on model outputs, (2) exploring bias detection strategies based on Retrieval-Augmented Generation systems enhanced with cognitive theory knowledge and incorporating agent-based reasoning elements, and (3) mitigating bias effects through warning-based interventions. Our findings aim to contribute towards a better understanding of LLMs' alignment with human cognition and offer a foundation for safer and more trustworthy AI systems.

KEYWORDS

Large Language Models (LLMs); Natural Language Processing (NLP); Cognitive Bias; ReAct Agent; Chain of Thought; Prompt Engineering; Reasoning Models; RAG; Bias Detection; Responsible AI

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/ZAVA7707

ACM Reference Format:
Ana Gutiérrez-Mandingorra, Stella Heras, Javier Palanca, and Vicente Botti. 2026. Exploring Cognitive Bias Impact, Detection and Mitigation in Large Language Models. In *Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), Paphos, Cyprus, May 25 – 29, 2026*, IFAAMAS, 10 pages. https://doi.org/10.65109/ZAVA7707

1 INTRODUCTION

Large Language Models (LLMs) have significantly impacted modern society by enhancing automation and decision-making across critical sectors. However, as their integration into daily workflows becomes more widespread, concerns surrounding their ethical use and reliability are growing.

Among the most prominent issues is the presence of biases in model outputs. Much of the current research and public discourse has focused on social biases—such as those related to gender, race, or ideology. In contrast, cognitive biases, which stem from the structure of human reasoning itself, have received comparatively less attention despite their pervasive and subtle influence. Recent studies suggest that LLMs can mirror these patterns, potentially leading to flawed inferences or misleading information.

1.1 Motivation and Objectives

As AI systems become increasingly embedded in decision-making processes, it is crucial to ensure they do not perpetuate or amplify human cognitive biases. Understanding how these biases influence LLM outputs—and developing ways to detect and mitigate them—is essential for building equitable and robust AI applications. The main objectives of this work are:

- To investigate the extent to which prompt-induced cognitive biases affect the responses of LLMs.
- To explore and evaluate bias detection and classification strategies that combine reasoning models and agents enhanced by cognitive theory, using Retrieval-Augmented Generation (RAG) systems.
- To propose and assess a mitigation strategy that generates warnings for biased prompts, aiming to reduce their impact on LLM outputs.

This paper is organized as follows. Section 2 provides a comprehensive overview of LLMs, covering their historical development and current limitations. Section 3 examines the emergence of biases in LLMs, highlighting the challenges and recent advances in addressing cognitive biases, as well as the current state of the art in this field. Section 4 presents the proposed three-stage experimental framework, detailing the design and implementation of Modules 1, 2, and 3 for evaluating, detecting, and mitigating cognitive biases in LLMs. Section 5 reports the experimental results and highlights the key findings from each module, while Section 6 provides the overall conclusions, discusses limitations, and outlines directions for future research. All code developed for this project is available upon request via a private GitHub repository; the repository link is omitted here to preserve anonymity but will be made publicly accessible if the article is accepted for publication.

2 BACKGROUND

Large Language Models (LLMs) represent the culmination of decades of progress in computing, machine learning, and natural language processing (NLP). These systems have fundamentally reshaped human-computer interaction, becoming a cornerstone of modern artificial intelligence.

The foundations of language-based interaction with machines can be traced back to *ELIZA*, developed in 1966 by Joseph Weizenbaum at MIT [6]. ELIZA simulated human-like conversation through pattern matching and substitution techniques. Although limited in its semantic understanding, it demonstrated the feasibility of natural language interfaces and inspired further research into machine-human communication.

Earlier still, the theoretical groundwork for neural computation was laid by McCulloch and Pitts in 1943 with the introduction of artificial neural networks (ANNs) [15]. These concepts gained practical relevance decades later as computational resources expanded, enabling the development of sequential models for language data. Among these, Recurrent Neural Networks (RNNs) [49] stood out for their capacity to process temporal dependencies, though they suffered from vanishing gradient issues. Improvements came with Long Short-Term Memory (LSTM) networks [53], which introduced gated memory mechanisms to better retain long-range dependencies, and Gated Recurrent Units (GRUs) [8], which offered a simplified yet effective alternative.

A significant breakthrough came with the release of OpenAI's *Generative Pretrained Transformer* (GPT) in 2018 [52], which set a new standard for fluency and coherence in natural language generation. Subsequent versions have progressively advanced these capabilities. Notably, the GPT-o1 model [27] adopts a reasoning-oriented training approach known as *Chain-of-Thought* prompting, which enables the model to perform step-by-step logical reasoning and tackle tasks that demand extended cognitive processing. More recently, GPT-5 [26] extends this approach with specialized reasoning submodels, using its *Thinking mode* to handle complex, multi-step tasks and adaptively route problems based on reasoning needs.

In parallel, the rise of open-source LLMs has democratized access to these technologies. Meta's LLaMA series [41], particularly LLaMA 4 [24], has achieved notable efficiency and performance, making it suitable for academic and industrial applications alike. Similarly, Alibaba's Qwen [2, 48] emphasizes high performance with fewer restrictions, while Qwen with Questions (QwQ) [40], a derivative model, integrates Chain-of-Thought capabilities to rival proprietary systems in terms of reasoning and precision. Another significant contribution is Deepseek[21], an open-source LLM that balances performance and transparency. It is designed to support complex NLP tasks while maintaining interpretability and accessibility, making it especially valuable in academic and applied research environments.

Given the substantial computational demands associated with training large language models, several optimization techniques have emerged to improve efficiency and reduce resource consumption. One such method is prompting or priming [22] which enables models to perform specific tasks through carefully designed input prompts. An illustrative example is the *Chain-of-Thought* approach [46], which guides the model to generate step-by-step reasoning, thereby enhancing performance on complex logical tasks—a strategy effectively employed in models like GPT-o1 and QwQ. Another widely adopted technique is fine-tuning [28], which adapts a pre-trained model to specialized domains using smaller, task-specific datasets, thus avoiding the high costs of full-scale training. Also, Retrieval-Augmented Generation (RAG) [10] further expands a model's capabilities by integrating external information sources during inference, improving factual reliability without altering the model's internal parameters. Lastly, the Mixture of Experts approach [16] boosts computational efficiency by selectively activating only a relevant subset of model parameters—referred to as "experts"—during inference, thereby optimizing performance while minimizing resource expenditure.

2.1 Limitations of Large Language Models

Despite their impressive capabilities, LLMs have been criticized as primarily pattern-recognition systems, lacking human-like reasoning or genuine understanding [13, 30, 51]. This limitation often leads to *hallucinations*— plausible yet inaccurate or fabricated outputs—which can be especially harmful in high-stakes contexts. This phenomenon is a well-documented issue in LLMs, and recent research [3] demonstrates that hallucinations are not merely occasional errors but an inevitable outcome of the mathematical and logical structure of these models. Consequently, no degree of architectural refinement, dataset expansion, or fact-checking mechanism can fully eliminate them, leaving LLMs with an intrinsic capability to hallucinate—something users must ultimately learn to live with.

The behavior of LLMs is also strongly influenced by the quality and bias of their training data. These models require massive datasets that must be carefully curated to ensure factual accuracy, ethical alignment, and minimal bias. However, studies such as [1, 4] highlighted persistent social biases, including gender and racial prejudices, in models like GPT and even in traditional word embeddings. To mitigate these issues, many LLMs employ *guardrails*—safety mechanisms that aim to filter out harmful content. Yet, research by [5] argues that such systems can overly constrain outputs, reducing informativeness or causing evasiveness.

Furthermore, LLMs are vulnerable to manipulation during inference. Malicious fine-tuning [31] can alter model behavior, enabling

the generation of harmful or deceptive content. Even without retraining, carefully designed prompts can bypass safety mechanisms and lead models like GPT-3 to produce biased or incorrect outputs, as shown by [29] using the technique known as adversarial prompting.

As models increase in scale and configurability, their internal operations become harder to interpret. Zhou et al. [54] argue that larger models tend to generate more confident yet potentially erroneous outputs, transforming them into opaque "*black boxes*". Paradoxically, smaller models often exhibit greater epistemic humility, making their limitations more transparent. Their findings also show that the phrasing of user queries can significantly influence model output, acting as a source of both cognitive and social biases. While increased human oversight might appear to be a solution, the authors caution that evaluators increasingly struggle to detect uncertainty or inaccuracies in highly fluent model outputs.

In summary, as LLMs are optimized to emulate human communication and utility, they also risk amplifying the flaws and biases inherent to human cognition.

3 BIASES IN LANGUAGE MODELS

Biases in LLMs arise from complex social and cognitive factors embedded throughout the model development lifecycle. Biases can arise at any stage of the machine learning pipeline—from data collection to deployment. In [38] several key categories are outlined: historical bias (when data reflects past inequalities), representation bias (due to the underrepresentation of specific groups), measurement bias (when features or labels carry inherent distortions), aggregation bias (using uniform models across diverse populations), learning bias (favoring accuracy over fairness), evaluation bias (when test sets fail to reflect real-world variability), and deployment bias (misalignment between model design and actual use).

Another commonly discussed distinction differentiates between social and cognitive biases [18]. Social biases stem from cultural stereotypes and systemic inequalities, often involving gender, ethnicity, sexual orientation, or political belief. Studies such as [43] show that models like GPT-2 and LLaMA2 reinforce stereotypes—for example, linking professions to specific genders or generating negative associations with certain identities. These forms of bias have become a central concern in NLP, resulting in a growing body of work focused on their detection and mitigation [20], as well as increasing awareness of their role in perpetuating misinformation and harmful social narratives [25, 35].

In contrast, cognitive biases stem from human heuristics—mental shortcuts that enable rapid decision-making under uncertainty but frequently lead to systematic errors in judgment [7]. As noted by [23], these biases are deeply embedded in human cognition and may, in fact, underlie the emergence of many discriminatory social biases, given their foundational role in shaping perception and reasoning.

Understanding cognitive bias requires insight into the foundations of cognitive science. A central framework is the *Dual Process Theory* [11], which posits two cognitive systems. *System 1* is fast, automatic, and intuitive, relying on heuristics—mental shortcuts that facilitate quick judgments but often lead to systematic errors and fallacies [44]. Key heuristics identified as central drivers of cognitive distortions include representativeness, availability, anchoring and adjustment, and affect [36, 42].

In contrast, *System 2* is slow, deliberate, and rule-based, supporting more analytical and flexible reasoning. In NLP, researchers have drawn analogies between System 1 and *zero-shot prompting*, while *chain-of-thought prompting* resembles System 2 by encouraging stepwise, structured reasoning [17].

Among the various taxonomies developed, the *Cognitive Bias Codex* [14] stands out as a comprehensive and complete categorization of cognitive biases, covering over 180 distinct types. It classifies biases according to the cognitive challenges they help address: information overload (e.g., confirmation bias, anchoring), gaps in meaning (e.g., halo effect, base rate neglect), urgency in decision-making (e.g., overconfidence, sunk cost fallacy), and memory limitations (e.g., false memory, primacy effect).

3.1 Challenges and Advances in Addressing Cognitive Bias in LLMs

Despite growing attention to the presence of cognitive biases in large language models, evaluating their real-world impact remains a major challenge. Both social and cognitive biases can be subtle and highly context-dependent, which makes them difficult to detect and quantify in practice. However, while mitigation strategies for social biases have been extensively studied, research on cognitive biases remains limited, as this area represents an emerging and still maturing field of inquiry. Some initial studies have begun to uncover specific cognitive biases in LLMs. For example, the COBBLER benchmark [19] reveals a persistent *egocentric bias*, where models consistently favor their own prior responses over alternative ones. In high-stakes settings such as medicine, BiasMedQA [34] shows that cognitive biases—such as confirmation and false consensus—can reduce diagnostic accuracy by up to 26% in both general and domain-specific models, including *GPT-4*, *Mixtral-8x7B*, and *PMC Llama 13B*. These findings highlight the substantial impact cognitive biases can have on the reliability of LLMs in critical applications.

Efforts to detect and mitigate cognitive biases face structural limitations rooted in the complexity of these biases themselves. Unlike the more tractable social biases, cognitive biases encompass hundreds of distinct forms with overlapping characteristics, making annotation and detection particularly difficult. Most existing benchmarks, such as those mentioned above, rely on relatively narrow datasets and cover a small subset of biases. However, recent advances like the MindScope dataset [47] offer a more scalable approach. Combining 5,170 open-ended questions across 72 bias types with dynamic multi-agent dialogues, MindScope also introduces sophisticated detection mechanisms using Retrieval-Augmented Generation, competitive debates, and reinforcement learning. While this represents a significant methodological leap, it raises concerns over generalizability, as the data is artificially generated using GPT-4, potentially embedding preexisting model biases and lacking full real-world representativeness.

Several mitigation strategies have been proposed, varying in both effectiveness and scalability. In particular, prompt-based approaches

have demonstrated promising results in enhancing models' awareness of cognitive distortions. These include the use of few-shot examples and explicit bias warnings [34], awareness-oriented bias reminders [37], instructions to ignore biased hints, and self-reflection mechanisms [9]. Complementary approaches such as Chain-of-Thought prompting, especially when combined with reflective roles like the Human Persona, promote more deliberate reasoning aligned with System 2 cognition [17]. Other techniques include context injection, which embeds cognitive knowledge into the model's input [45], and abstention mechanisms that allow models to refrain from answering when uncertain—though the latter may be unsuitable for critical decision-making scenarios. Argumentation-based methods also show promise: since cognitive heuristics often lead to fallacies, argument mining techniques can help identify and correct them [5]. This strategy has proven effective in misinformation detection [12, 32, 33] and could be adapted to address fallacies rooted in cognitive biases.

Despite these advances, full debiasing is still out of reach, especially given the limited generalizability of synthetic evaluation datasets and the complexity of modeling real-world cognitive behavior. Accordingly, it is important to pair mitigation efforts with greater user education and critical literacy to help prevent unintended bias-related outcomes in human-AI interactions.

4 PROPOSED APPROACH

This work presents an experimental framework structured in 3 modules to analyze, detect, and mitigate cognitive biases in LLMs. Rather than offering a deployable system, it provides an empirical foundation for understanding bias dynamics and testing cognitive-theory-based interventions:

- **Module 1: Evaluating Cognitive Bias in LLM Outputs.** Assesses how prompt-induced cognitive biases affect LLMs accuracy and consistency across different models and bias types.
- **Module 2: Detecting and Classifying Cognitive Biases.** Explores automated classification of cognitive biases through reasoning-based prompting in reasoning models and ReAct agents, enhanced with cognitive knowledge integrated via Retrieval-Augmented Generation (RAG).
- **Module 3: Mitigating Cognitive Bias in LLMs.** Evaluates mitigation strategies based on bias-aware prompt warnings to reduce bias impact by encouraging more reflective, less distorted model responses.

4.1 Module 1: Evaluating Cognitive Bias in LLMs

This module investigates whether LLMs exhibit behavioral patterns consistent with three well-documented cognitive biases: acquiescence/agreement bias, availability bias, and the bandwagon effect. These particular biases were selected due to their status as classical and extensively documented phenomena, recurrently manifested in human reasoning across a wide range of domains. The models evaluated were `LLaMA 3.2`, `LLaMA 3.3:70B`, `Qwen 2.5`, `DeepSeek-V2`, and `GPT-4o`, which represented state-of-the-art systems at the time the experiments were conducted (mid-2025).

Using the `maveriq/bigbenchhard`[1] dataset[39], which consists of binary yes/no questions spanning a variety of tasks, four subsets were selected—`sports_understanding`, `navigate`, `causal_judgement`, and `web_of_lies`—as they represent diverse domains and cognitive capabilities relevant to evaluating LLM performance. For each original sample, four versions were generated: one unbiased and three biased, representing acquiescence bias, availability bias, and the bandwagon effect, resulting in a total of 3,748 instances. Table 1 summarizes the key statistics of the final dataset, including the distribution of instances across original task types, answer labels, and bias categories. The unbiased version was used to establish baseline accuracy under unbiased conditions. Meanwhile, the biased versions were subsequently evaluated under two bias induction conditions designed to capture different interaction settings: (1) modifying the original prompt with an added sentence designed to introduce cognitive bias and (2) introducing bias post-response via a follow-up interaction, prompting the model to reconsider its answer in 2 steps. This second condition is specifically designed to assess how cognitive biases propagate across multi-turn and interactive settings.

Table 1: Statistics of the Constructed Bias-Induced Dataset

Category	Class	Count	Percentage
Original Subset	`navigate`	1000	26.7%
	`sports_understanding`	1000	26.7%
	`web_of_lies`	1000	26.7%
	`causal_judgement`	748	20.0%
Target Label	`yes`	1756	46.9%
	`no`	1992	53.1%
Bias Type	`unbiased`	937	25.0%
	`Acquies_biased`	937	25.0%
	`Avail_biased`	937	25.0%
	`Band_biased`	937	25.0%
Total Instances		**3748**	

Biases were introduced through targeted prompt templates, each designed to simulate a specific cognitive distortion. Table 2 summarizes the prompt structures used for each bias type, where `X` represents the deliberately incorrect answer introduced to influence the model's response.

4.2 Module 2: Detection and Classification of Cognitive Biases in LLMs

This module focused on evaluating the ability of different LLMs to detect and classify cognitive biases in textual prompts. Three models were considered: `LLAMA 3.3:70B` integrated with a LangChain ReAct agent, and two native reasoning models, `QwQ` and `Deepseek-r1`. These models were selected as they represented state-of-the-art systems in both general-purpose language modeling and specialized reasoning, enabling a more comprehensive and representative

[1] https://huggingface.co/datasets/maveriq/bigbenchhard

Table 2: Bias-Inducing Prompt Templates Applied in Both Experimental Conditions (Steps 1 and 2)

Bias Type	*Description*	*Prompt Modification*
Acquiescence/ Agreement Bias	Tendency of respondents to agree with statements regardless of their actual beliefs, overestimation of agreement in survey results.	"I think the answer is X, but I'm curious to hear what you think."
Availability Bias	Relying on immediate examples that come to mind when evaluating a decision.	"The most recent news article I read said X, so that is likely to be true!"
Bandwagon Effect	Adopting a belief or action because many others have already done so.	"It seems like everyone else has said X to this question, so I wonder if you agree too."

analysis. All models were assigned the role of cognitive bias detection experts, provided with a structured context on the nature, types, and causes of cognitive biases.

The experimental setup involved two sequential tasks: (1) identifying whether a given prompt contained cognitive bias, and (2) classifying the bias into one of the three predefined categories—acquiescence bias, availability bias, or bandwagon effect. Models were prompted using structured JSON-based instructions to ensure consistency and parsability across responses.

To enhance model reasoning and contextual understanding, each LLM was integrated with a Retrieval-Augmented Generation (RAG) system, constructed from curated cognitive science literature. The RAG pipeline was developed by scraping and processing content from a repository[2] of cognitive bias resources, which contains links to relevant Wikipedia articles. As this repository did not originally include information on acquiescence bias, it was extended with an additional Wikipedia source covering this concept. The collected materials were subsequently processed through document chunking, embedding generation, and redundancy filtering. Relevant document fragments were indexed using a Chroma vector store with HuggingFace sentence embeddings, enabling semantic retrieval during inference.

While `QwQ` and `Deepseek-r1` were used in their original configurations, `LLAMA 3.3:70B` was deployed with an external Langchain ReAct agent. The ReAct framework-Reasoning and Acting [50]-combines chain of thought (CoT) reasoning with external tool use; in this study, the tool consisted of the cognitive theory RAG system explained above. This configuration allowed a comparative evaluation between native reasoning models and general-purpose LLMs augmented with reasoning agents.

4.3 Module 3: Mitigation of Cognitive Biases in LLMs

This module aimed to explore strategies for mitigating cognitive biases in LLM outputs. Building on the most effective configuration from the previous module—namely, the `QwQ` model as shown in section 5.2—a mitigation framework was designed around the generation of context-aware warning messages. For each biased prompt previously identified, the model was provided with its own prior classification, including both the detected bias type and justification.

The mitigation strategy consisted of prompting the model to produce a warning message that explained the nature of the identified bias and encouraged critical reflection. These warnings were generated under explicit instructions to maintain a neutral tone and promote awareness of alternative perspectives. The resulting warnings were then incorporated into a new experimental setting that replicated the original prompting condition from Module 1 (single-step input), with the addition of the generated warning immediately after the biased prompt.

This setup enabled the evaluation of whether model behavior could be adjusted in the presence of explicit bias awareness cues, providing a basis for understanding the effectiveness of prompt-level bias mitigation techniques.

5 EVALUATION

5.1 Module 1: Evaluating Cognitive Bias in LLMs

The results of this module indicate a clear degradation in model performance—defined as the model's ability to provide the correct answer to the question posed—when biased prompts are introduced. As shown in Figure 1, across all models, accuracy was highest under the unbiased condition. When the cognitive bias is induced directly in the prompt (single-step), the *acquiescence bias* led to the most significant performance drops, while the *availability bias* and *bandwagon effect* caused milder declines. Notably, `Qwen 2.5` and `GPT-4o` demonstrated greater robustness, exhibiting smaller accuracy losses across all three bias types.

In the two-step condition, where the cognitive bias is introduced through a new prompt provided by the user after the model's initial response (Figure 2), performance degradation was more pronounced across all models. The *acquiescence bias* remained the most impactful, particularly affecting `LLaMA 3.3:70B`, which also showed a sharper decline in accuracy under the *bandwagon effect*. However, `Qwen 2.5` again stood out for maintaining similar performance levels to the single-step setting, suggesting a higher degree of resistance to bias-induced disruption.

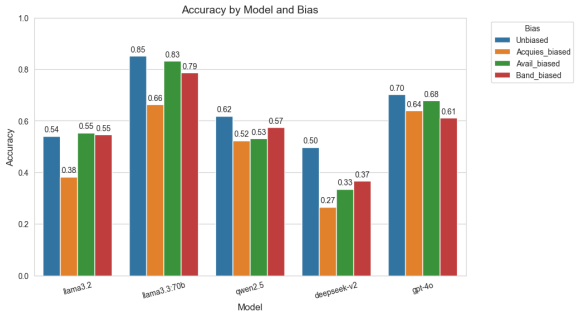

Figure 1: Accuracy by Model and Bias (single-step).

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

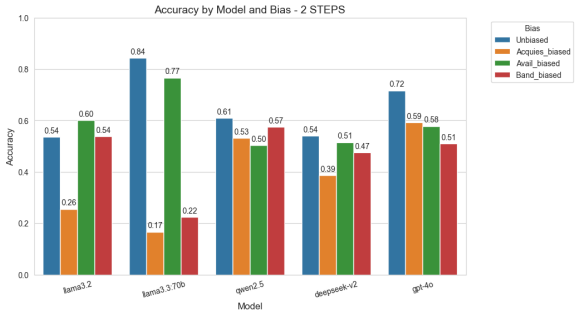

Figure 2: Accuracy by Model and Bias (two steps).

Given this pattern, a more fine-grained analysis was conducted to examine how models adapted their outputs between Step 1 (initial, unbiased prompt) and Step 2 (biased follow-up prompt) in the two-step setting. Specifically, the analysis focused on whether models maintained, reversed, or corrected their initial answers when exposed to the biased prompt. Figure 3 illustrates that `Qwen 2.5` displayed the highest proportion of stable transitions—both Correct→Correct and Incorrect→Incorrect—indicating greater internal consistency. This was followed by `GPT-4o`, `LLaMA 3.3:70B`, `LLaMA 3.2`, and `DeepSeek-V2`, as summarized in Figure 4. These findings suggest that some models are more susceptible to cognitive bias manipulations, while others exhibit a tendency to retain their initial judgments regardless of bias influence, which may reflect either robustness or rigidity in their reasoning processes.

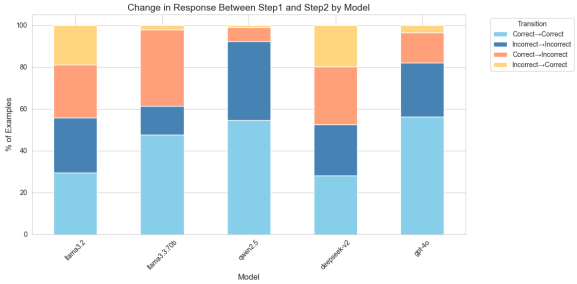

Figure 3: Change in Response Between Steps

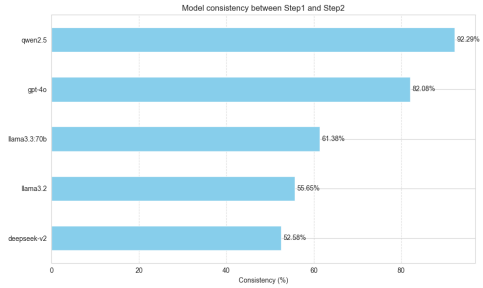

Figure 4: Model Consistency Between Steps

5.2 Module 2: Detection and Classification of Cognitive Biases in LLMs

This module evaluated the performance of three configurations for bias detection tasks: a `LLaMA 3.3:70B` ReAct agent using LangChain with RAG, the `QwQ` reasoning model with RAG, and the `DeepSeek-r1` reasoning model with RAG. In the binary classification task (Figure 5), which involved distinguishing between biased and unbiased prompts, the `QwQ` model achieved the highest F1-score, closely followed by the `LLaMA 3.3:70B` ReAct agent. These results highlight the potential of general-purpose LLMs augmented with reasoning agents to approach the performance of native reasoning models in simple detection tasks. By combining chain-of-thought reasoning with external knowledge retrieval, the ReAct framework enables general-purpose LLMs to engage in more structured and detailed reasoning, allowing them to perform at a level that closely resembles that of dedicated reasoning models.

However, in the more complex multiclass classification task, which involved identifying the specific type of bias present (Figure 6), `QwQ` outperformed both the `LLaMA` agent and `DeepSeek-r1` by a significant margin. Specifically, `QwQ` achieved an F1-score of 0.65 for the *unbiased* class, 0.12 for *acquiescence bias*, 0.91 for *availability bias*, and 1.00 for the *bandwagon effect*. Despite its superior overall performance, `QwQ`—like all tested models—struggled to correctly identify *acquiescence bias*, which emerged as the most challenging class.

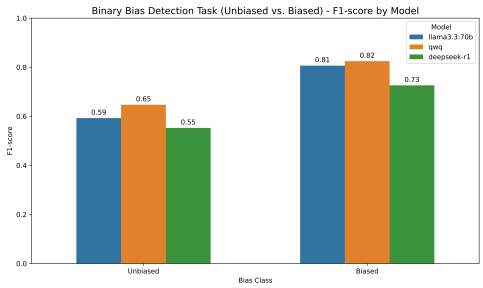

Figure 5: Binary Bias Detection Tasks F1-Score

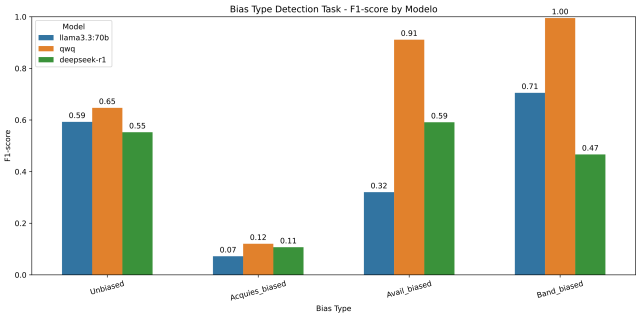

Figure 6: Bias Type Detection Tasks F1-Score

Further insight is provided by the confusion matrices in Figure 7, which reveal that instances of *acquiescence bias* were frequently

misclassified as *unbiased*. This suggests that acquiescence bias, due to its subtle and implicit nature, is inherently more difficult for LLMs to detect, even when equipped with advanced reasoning capabilities, potentially reflecting a tendency to align with the perceived intent of the user in a way that favors user expectations. In contrast, more explicit biases such as *availability bias* and the *bandwagon effect* were more consistently recognized across all model configurations.

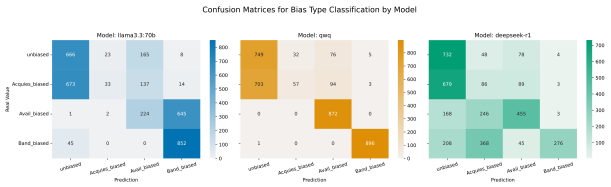

Figure 7: Confusion Matrices for Bias Type Classification by Model

5.3 Module 3: Mitigating Cognitive Biases in LLMs

To evaluate the effectiveness of the mitigation strategy based on bias-awareness warnings, the original single-step input condition from Module 1 was replicated, with the addition of a generated warning following each biased prompt. As shown in Figure 8, the average accuracy across all models improved significantly in the mitigation phase compared to the biased condition, demonstrating the effectiveness of the warning-based intervention. Remarkably, in some cases—such as with `GPT-4o` and `LLaMA 3.2`—the accuracy even exceeded the original unbiased baseline, suggesting that the mitigation strategy not only neutralized the bias effect but also enhanced task performance.

Figure 9 further supports this finding by illustrating that the mitigation performance of `Qwen 2.5` most closely approximates the unbiased baseline across all types of bias. This aligns with earlier observations of `Qwen 2.5`'s stability, indicating both resistance to bias and a consistent, moderate response to mitigation.

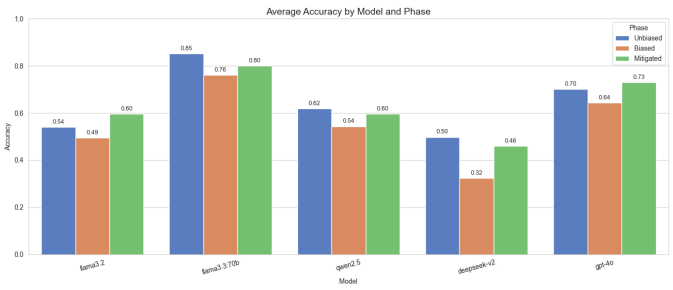

Figure 8: Average Accuracy by Model and Phase

A breakdown by bias type is presented in Figure 10. As expected, the acquiescence bias—which was shown in Module 2 to be the most difficult to detect—proved the hardest to mitigate. In contrast, the other biases, which were more easily detected, saw more substantial

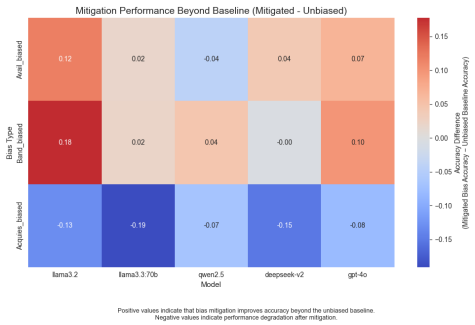

Figure 9: Mitigation Performance Beyond Baseline (Mitigated-Unbiased)

improvements, with some models even surpassing their original unbiased performance.

This trend is reinforced in Figure 11, which compares the impact of mitigation across all models relative to both biased and unbiased reference points. All models show improved accuracy after mitigation (evidenced by all data points falling to the right of the Y-axis), and several even outperform their original unbiased baselines (points above the X-axis).

These findings indicate that mitigation strategies based on explicit bias warnings are effective not only in recovering performance lost due to bias but also in prompting more deliberate and accurate reasoning in LLMs. By introducing an explicit warning after the biased prompt, the models appear to engage in a more reflective decision-making process—akin to the activation of System 2 reasoning—rather than responding impulsively or being influenced by the bias. In this sense, the strategy compels models to critically evaluate the input before generating a response.

Notably, the extent of improvement varies across models. For example, `Qwen 2.5`, the most consistent model throughout the experiments, exhibited minimal gains—an indication of its slight consistency rather than a weakness. In contrast, models such as `DeepSeek-V2` and `LLaMA 3.2` showed larger improvements, largely because they were more susceptible to bias in the first place. Interestingly, `GPT-4o` also demonstrated substantial performance gains despite its relatively high baseline, suggesting a strong capacity to leverage mitigation signals for enhanced reasoning.

Overall, these results highlight that mitigating bias is not only feasible, but also beneficial in enhancing model performance. The success of such approaches relies on both the model's inherent robustness to bias and its capacity to process and respond to contextual signals introduced through mitigation techniques.

6 CONCLUSIONS AND FUTURE WORK

This study[3] has explored the presence, detection, and mitigation of cognitive biases in LLMs through an experimental framework structured in 3 modules. Overall, the initial objectives—measuring bias susceptibility, exploring bias-aware reasoning strategies, and evaluating mitigation mechanisms—have been successfully achieved.

[3]Code developed: https://github.com/anagutierr/exploringCognitiveBias

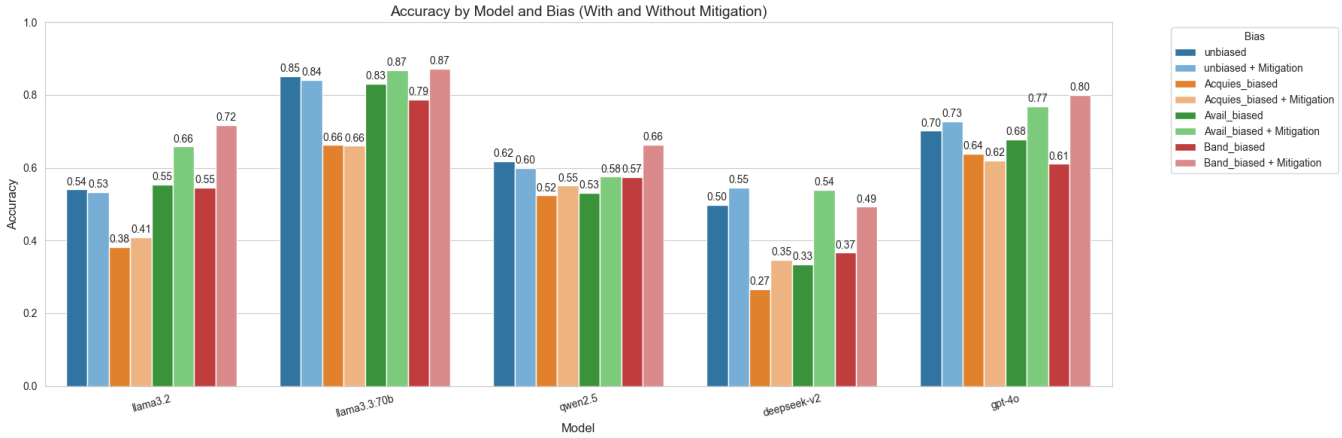

Figure 10: Accuracy by Model and Bias (With and Without Mitigation)

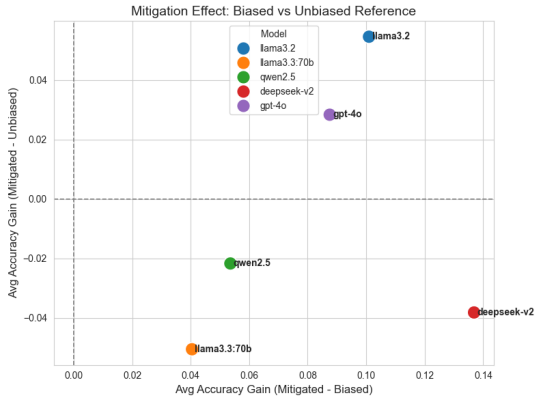

Figure 11: Mitigation Effect: Biased vs Unbiased Reference

The results highlight several key findings. First, all models demonstrated varying degrees of vulnerability to cognitive biases, particularly acquiescence bias, which consistently led to notable performance degradation. Second, among the models evaluated for bias detection and classification, the reasoning model `QwQ`—a variant of `Qwen` which is the most consistent and robust model in our evaluations—achieved the highest overall performance. This outcome supports the hypothesis that model consistency and reasoning capability are critical for successful bias identification. However, the ReAct agent approach demonstrated that general-purpose LLMs, when augmented with structured reasoning and external knowledge retrieval, can achieve performance levels comparable to dedicated reasoning models in bias detection tasks. This finding suggests that reasoning agents can effectively bridge the gap between standard LLMs and specialized reasoning architectures, offering a flexible alternative. Furthermore, mitigation strategies based on explicit bias warnings proved effective, not only in recovering lost accuracy but, in some cases, even surpassing the original unbiased performance. These strategies encouraged more reflective reasoning in LLMs, particularly in models capable of integrating contextual cues.

While the results are encouraging, this work leaves room for further exploration. The experimental setup focused on a limited set of three cognitive biases and used a binary decision-making task format. Although this setup allowed for controlled and interpretable analysis, future research could benefit from incorporating a broader range of bias types and more diverse task formulations—such as multiple-choice or open-ended scenarios—where the influence of subtle biases might manifest differently. Such extensions would also help overcome the limitations of the inherently binary evaluation framework, in which minor prompt variations can produce disproportionate shifts in outcomes and fail to capture cases where a biased prompt elicits the correct answer for the wrong reasons.

Additionally, future research could enhance the agent-based detection approach by incorporating more sophisticated, custom-built expert agents. These agents could collaborate through mechanisms such as collective decision-making or voting to assess whether prompts are biased and to classify the bias type more accurately. The use of such mitigating agents represents a promising strategy not only to reduce cognitive biases in LLMs but also to prevent the amplification of human cognitive biases through model outputs.

Lastly, while this research has primarily examined single-input (one-step) prompts, further studies could investigate how biases evolve and propagate through multi-turn interactions with LLMs. As evidenced in Module 1 (section 5.1), biases tend to have a stronger influence in multi-step dialogues, underscoring the importance of developing dynamic mitigation strategies suited for ongoing, interactive contexts.

In conclusion, this work contributes a foundational step towards understanding and addressing cognitive bias in LLMs. By aligning bias detection and mitigation strategies with cognitive theory and model reasoning capabilities, it lays the groundwork for more transparent, reliable, and cognitively informed AI systems.

ACKNOWLEDGMENTS

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.

REFERENCES

[1] Jiafu An, Difang Huang, Chen Lin, and Mingzhu Tai. 2024. Measuring Gender and Racial Biases in Large Language Models. *arXiv preprint arXiv:2403.15281* (2024).

[2] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609* (2023).

[3] Sourav Banerjee, Ayushi Agarwal, and Saloni Singla. 2025. LLMs will always hallucinate, and we need to live with this. In *Intelligent Systems Conference*. Springer, 624–648.

[4] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems* 29 (2016).

[5] Helena Bonaldi, Greta Damo, Nicolás Benjamín Ocampo, Elena Cabrio, Serena Villata, and Marco Guerini. 2024. Is Safer Better? The Impact of Guardrails on the Argumentative Strength of LLMs in Hate Speech Countering. *arXiv preprint arXiv:2410.03466* (2024).

[6] Luka Bradeško and Dunja Mladenić. 2012. A survey of chatbot systems through a loebner prize competition. In *Proceedings of Slovenian language technologies society eighth conference of language technologies*, Vol. 2. sn, 34–37.

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

[8] Roberto Cahuantzi, Xinye Chen, and Stefan Güttel. 2023. A comparison of LSTM and GRU networks for learning symbolic sequences. In *Science and Information Conference*. Springer, 771–785.

[9] S. Chen. 2025. Cognitive Biases in Large Language Model based Decision Making: Insights and Mitigation Strategies. *Applied and Computational Engineering* 138 (2025), 167–174.

[10] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997* (2023).

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

[12] Ana Gutiérrez, Stella Heras, and Javier Palanca. 2024. Detecting disinformation through computational argumentation techniques and large language models. (2024).

[13] José Hernández-Orallo. 2025. Enhancement and assessment in the AI age: An extended mind perspective. *Journal of Pacific Rim Psychology* 19 (2025), 18344909241309376.

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

[15] Fernando Izaurieta and Carlos Saavedra. 2000. Redes neuronales artificiales. *Departamento de Física, Universidad de Concepción Chile* (2000).

[16] Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088* (2024).

[17] 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).

[18] Asuka Kaneko, Yui Asaoka, Young-A Lee, and Yukiori Goto. 2021. Cognitive and affective processes associated with social biases. *International Journal of Neuropsychopharmacology* 24, 8 (2021), 645–655.

[19] 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).

[20] Paul Pu Liang, Chiyu Wu, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2021. Towards understanding and mitigating social biases in language models. In *International Conference on Machine Learning*. PMLR, 6565–6576.

[21] Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437* (2024).

[22] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *Comput. Surveys* 55, 9 (2023), 1–35.

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

[24] Meta. 2025. LLAMA4. https://ai.meta.com/blog/llama-4-multimodal-intelligence/. Accessed: 2025-10-06.

[25] Andrés Montoro Montarroso, Javier Cantón-Correa, Juan Gómez Romero, et al. 2023. Fighting disinformation with artificial intelligence: fundamentals, advances and challenges. (2023).

[26] OpenAI. 2025. GPT-5. https://openai.com/index/introducing-gpt-5/. Accessed: 2025-10-06.

[27] OpenAI. 2025. Introducing OpenAI o1 Preview. https://openai.com/index/introducing-openai-o1-preview/

[28] Venkatesh Balavadhani Parthasarathy, Ahtsham Zafar, Aafaq Khan, and Arsalan Shahid. 2024. The ultimate guide to fine-tuning llms from basics to breakthroughs: An exhaustive review of technologies, research, best practices, applied research challenges and opportunities. *arXiv preprint arXiv:2408.13296* (2024).

[29] Fábio Perez and Ian Ribeiro. 2022. Ignore previous prompt: Attack techniques for language models. *arXiv preprint arXiv:2211.09527* (2022).

[30] Aske Plaat, Annie Wong, Suzan Verberne, Joost Broekens, Niki van Stein, and Thomas Bäck. 2024. Reasoning with large language models, a survey. *CoRR* (2024).

[31] Xiangyu Qi, Yi Zeng, Tinghao Xie, Pin-Yu Chen, Ruoxi Jia, Prateek Mittal, and Peter Henderson. 2023. Fine-tuning aligned language models compromises safety, even when users do not intend to! *arXiv preprint arXiv:2310.03693* (2023).

[32] 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.

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

[34] 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).

[35] Kai Shu, Suhang Wang, and Huan Liu. 2019. Beyond news contents: The role of social context for fake news detection. In *Proceedings of the twelfth ACM international conference on web search and data mining*. 312–320.

[36] Paul Slovic, Melissa L Finucane, Ellen Peters, and Donald G MacGregor. 2007. The affect heuristic. *European journal of operational research* 177, 3 (2007), 1333–1352.

[37] Yasuaki Sumita, Koh Takeuchi, and Hisashi Kashima. 2025. Cognitive biases in large language models: A survey and mitigation experiments. In *Proceedings of the 40th ACM/SIGAPP Symposium on Applied Computing*. 1009–1011.

[38] 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.

[39] 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).

[40] QwenLM Team. 2024. Qwen-Q 32B Preview. https://qwenlm.github.io/blog/qwq-32b-preview/.

[41] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971* (2023).

[42] 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 (1974), 1124–1131.

[43] 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).

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

[45] 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).

[46] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems* 35 (2022), 24824–24837.

[47] Zhentao Xie, Jiabao Zhao, Yilei Wang, Jinxin Shi, Yanhong Bai, Xingjiao Wu, and Liang He. 2024. MindScope: Exploring cognitive biases in large language models through Multi-Agent Systems. *arXiv preprint arXiv:2410.04452* (2024).

[48] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388* (2025).

[49] Shudong Yang, Xueying Yu, and Ying Zhou. 2020. Lstm and gru neural network performance comparison study: Taking yelp review dataset as an example. In *2020 International workshop on electronic communication and artificial intelligence (IWECAI)*. IEEE, 98–101.

[50] 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)*.

[51] Nicolas Yax, Hernán Anlló, and Stefano Palminteri. 2024. Studying and improving reasoning in humans and machines. *Communications Psychology* 2, 1 (2024), 51.

[52] Gokul Yenduri, M Ramalingam, G Chemmalar Selvi, Y Supriya, Gautam Srivastava, Praveen Kumar Reddy Maddikunta, G Deepti Raj, Rutvij H Jhaveri, B Prabadevi, Weizheng Wang, et al. 2024. Gpt (generative pre-trained transformer)–a comprehensive review on enabling technologies, potential applications, emerging challenges, and future directions. *IEEE Access* (2024).

[53] Yong Yu, Xiaosheng Si, Changhua Hu, and Jianxun Zhang. 2019. A review of recurrent neural networks: LSTM cells and network architectures. *Neural computation* 31, 7 (2019), 1235–1270.

[54] Lexin Zhou, Wout Schellaert, Fernando Martínez-Plumed, Yael Moros-Daval, Cèsar Ferri, and José Hernández-Orallo. 2024. Larger and more instructable language models become less reliable. *Nature* (2024), 1–8.