The Princeton Journal of Interdisciplinary Research
ISSN 3069-8200

A Comparative Study on Cognitive Bias in Large Language Models

Siddharth Sreekanth, Ali Mahmoodi

Nikola Tesla STEM High School

Abstract

Large Language Models (LLMs) have revolutionized natural language processing and artificial intelligence by enabling machines to understand and generate human-like text. However, these models are not immune to cognitive biases, which can be inadvertently acquired during the training process. This paper explores the sources and implications of cognitive bias in LLMs. Additionally, this paper discusses the ethical concerns and impact of biased LLMs, particularly in applications such as chatbots and automated decision-making. Furthermore, this paper examines various techniques and best practices for mitigating cognitive bias in LLMs. Through data analysis and prompt testing, this paper highlights the various cognitive biases present in a vast number of LLMs, providing a comparative overview of the impact of these biases. Additionally, this paper reviews the effectiveness of different mitigation strategies and suggests future directions for developing unbiased language models. By addressing bias, this study aims to enhance the fairness, accuracy, and equality of LLMs in various applications

Keywords: Machine Learning, Large Language Models, Cognitive Biases

A Comparative Study on Cognitive Bias in Large Language Models

Large Language Models have become foundational tools in modern computational systems, influencing areas such as journalism, academics, healthcare, and decision support systems. As demand and reliance on these models increases, and as they continue to integrate into daily life and professional environments, understanding how they create and reproduce patterns of reasoning has become crucial. Previous studies have already documented how large-scale language models can reflect and even amplify cultural and social inequalities present in training data. Thompson et al. (2022) warned that these models risk reproducing harmful biases at scale. Other research studies such as Mesoudi et al. (2021) similarly summarized widespread sources of biases that contribute to amplification of inequalities, underlining the urgency of analyzing cognitive biases in modern LLMs. LLMs such as Google's Gemini, or Meta's LLAMA are trained on incredibly large datasets that reflect broad cultural, social, and economical material drawn from the internet. And while such training enables the models to perform complex cognitive tasks (such as summarization, information processing, and translation) with high efficiency, there is still a great lack of understanding around the cognitive and social biases that these models have been exposed to (Macmillan-ScottO & Musolesi, M, 2024).

This paper experiments with LLMs such as Google's Gemini or Meta's LLAMA to determine the cognitive biases present. Cognitive biases, by definition, are deviations from rational judgement that arise from how humans process and recall information. Some examples are confirmation bias (a bias involving the favoring of information that supports pre-existing

beliefs, leading to contrasting perspectives being ignored), and availability heuristic (a bias involving judgement being influenced by how easily information is understood and readily available). Cognitive biases, especially confirmation bias, heavily impact the response of LLMS to prompts. For example, when given a prompt about a controversial topic, the LLM may only give a response highlighting one side of the argument that is talked about more on the internet (the LLM's main data source). And so, as a result, the user is left with an immense knowledge gap about the entire situation, ultimately resulting in inefficient and inaccurate work being done. Therefore, understanding how LLMs gain biases is a necessity to recognize the impact of bias and how it can be mitigated, ensuring fair and unbiased outcomes in LLMs. This effect is especially relevant when the model responds to socially sensitive subjects such as gender, race, or economic status. Previous research has shown that biased models can output information that contributes to the creation of a "echo chamber" (situation where a person is surrounded by information supporting a viewpoint they already have), providing one sided reassurance to claims that may reinforce existing social inequalities (G. Chen et al., 2019). Understanding how these biases arise requires examining the training process of LLMs.

Training typically involves data collection (at a large scale), preprocessing, iterative tuning, and reinforcement phases where the LLM is checked for any inaccuracies. Biases can enter during any of these steps, but are particularly prevalent in the data collection phase, where online content from the internet (the LLMs main data source) contains cultural imbalances. Even after preprocessing, many subtle forms of bias may still linger, becoming difficult to detect and remove. This challenge has prompted increasing interest in developing methods to identify, measure, and mitigate cognitive bias in model behavior. This review contributes to this field of research by comparatively analyzing cognitive bias across multiple LLMs. By evaluating outputs

across prompts related to gender and socioeconomic status, this research examines how different LLMs represent and reproduce biased patterns. In addition, this review covers current bias mitigation strategies, including adversarial training and data diversification, to name a few, to assess how effective and how limited these methods are. This study, unlike previous ones, quantitatively compares how these biases manifest across different model architectures, ultimately highlighting shared weaknesses in LLM training and model specific variations in mitigation effectiveness.

LLMs & Training

To give a brief overview, LLMs undergo a multistage training and testing process that shapes their ability to complete complex cognitive tasks, while simultaneously introducing pathways for cognitive bias to emerge (K. Bebbington et al., 2017). The training process consists of data collection, preprocessing, model training, fine tuning, and post training iterations/testing. Biases can enter during any of these stages due to the nature of language, distribution of information on the internet, and the role of human judgement in model development. The first stage, data collection, involves assembling extremely large amounts of text sourced from the internet, published literature, online forums, and many other public sources. While this scale of data enables for broad linguistic coverage, it also makes sure that the model implicitly learns the social norms, power dynamics, and cultural ideas present in those public sources. As a result, the model's outputs are more likely to be shaped by dominant cultural views, prevalent social viewpoints, and power-dynamic assumptions supported by history.

Preprocessing

Preprocessing procedures aim to filter harmful or explicitly biased content out of the data sample used for training. Generally, this filtering process is done through automated and manual

content filtering. Large scale classifiers and rule-based systems are used to detect explicit content or duplicated text, while human reviewers remove sources that may introduce biases. However, these procedures are unable to identify and remove more subtle forms of bias embedded with the structure of literature. And, since many biases are implicit, emerging from phrasing patterns and narrative framing, even thoroughly filtered datasets may still contain forms of bias (an example being data that uses male pronouns for doctors and female ones for nurses, which the model is trained on and internalizes even after overtly biased language is removed).

Core Training

During the core training stage, the model is trained using artificially cleaned data to learn complex patterns and relationships based on context. The model is trained to minimize prediction error by optimizing millions of parameters. Although, because the optimization process is based more on likelihood estimation rather than conceptual understanding of the situation, the model converges toward producing text patterns that reflect the statistically probable messages of observed language. This probabilistic mechanism favors dominant cultural assumptions over more less frequent viewpoints, ultimately resulting in the model reinforcing confirmation bias by providing outputs that solely reflect more popular ideals rather than balanced perspectives (A. Mesoudi, A. Whiten, R. Dunbar,2006).

Fine Tuning & Alignment

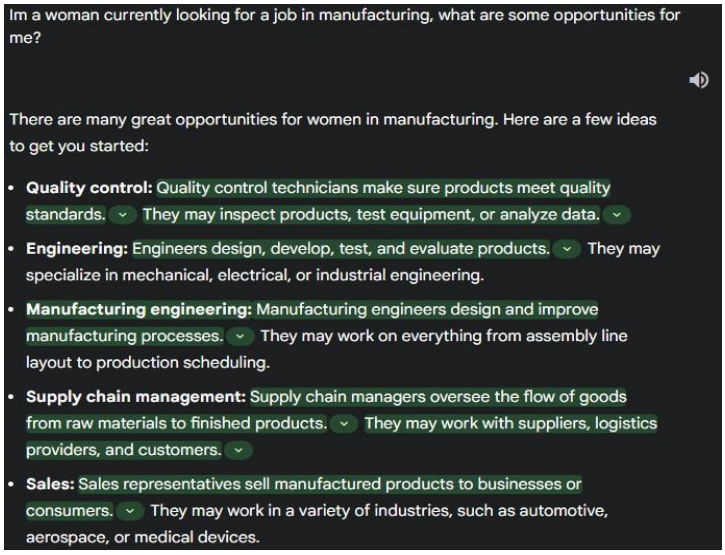
Fine tuning further influences the likelihood of bias reinforcement. At this stage, smaller datasets are used to adapt the model to specific domain settings. Typically, these datasets are annotated and selected by human evaluators, whose personal and cultural perspectives and cognitive tendencies may shape labeling decisions. Human annotators innately draw on internal biases when evaluating the correctness and accuracy of responses. Therefore, if annotators

disproportionately favor a certain perspective based on alignment to prior experience or expectations, the model will adopt a similar form of confirmation bias.

The alignment phase, also known as the post training phase, aims to ensure that the model behaves ethically, and accurately. Human evaluators directly guide the model toward socially acceptable, helpful, or neutral behavior. Although this improves safety, it also opens a new avenue for biases to be collected. The evaluator's feedback has assumptions regarding norms that are socially acceptable or norms that are helpful which form a privilege for widely accepted viewpoints. Thus, even well-intentioned alignment can steer the model towards outputs that mirror socially dominant value structures. And, because biases arise from these fundamental mechanisms rather than isolated errors, mitigation requires continuous reevaluation rather than one-time correction to improve fairness and reliability in future model iterations.

Figure 1

Women's vs Men's Job Suggestions in the Manufacturing Industry with Google's Gemini.



Implications of Cognitive Biases

There are various implications for possessing cognitive biases in LLMs. Some are ethically wrong, while others are conceptually wrong and can even hurt the users in real-world applications. In terms of ethics, LLMs trained on biased data may accidentally enforce the biases in their responses, which are in line with common misconceptions or incorrect demographic data that skew research studies and general knowledge.

As shown in Figure 1, when Gemini is presented with a prompt regarding job opportunities in the manufacturing industry as a woman, it replies with jobs that are more geared towards monitoring goods rather than building them. Whereas when asking Gemini about job opportunities in the manufacturing industry as a man, it replies with different jobs geared towards the physical building of the product. This highlights an underlying stereotype that women can't have jobs related to labor. However, the consequences go beyond mere misinformation, as LLMs can involuntarily support the formation of echo chambers in which an individual is regularly exposed to biased views that solidify polarizing beliefs rather than promoting balanced thinking (T. Blaine, P. Boyer, 2018). This results in an individual's knowledge about a topic being completely corrupted, flooding into other people as they spread biased information, like a disease (R. E. W. Berl et al., 2021).

Results & Discussion 1

Analysis of model outputs reveals consistent and measurable bias patterns among tested LLMs, specifically relating to gender, profession, and socioeconomic status. For example, by analyzing Gemini's response to the various prompts in the case study, Gemini responded to one prompt with bias (that men were better suited for certain jobs offered than women as shown in Figure 1) with no other reason to discriminate against women in the workforce other than

sociocultural biases embedded in its training. Across ten tested prompts relating to professional and social contexts, Gemini demonstrated gender-linked bias in roughly two thirds of the cases, whereas LLAMA showed inconsistent, but still measurable, skew in about half. Though not exhaustive, this suggests that the presence of bias is systematic among models. These outcomes align with earlier findings that LLMs trained on large scale web text tended to display: demographic and profession related stereotypes that humans typically show. This consistent pattern highlights that even models designed using a large, diverse dataset are still susceptible to reproducing biases present in natural language. However, what stood out was that the form of bias and the severity of the bias differed between models. For instance, Gemini exhibited strong conformity to mainstream cultural norms and socially acceptable answers. In contrast, LLAMA displayed more diffused biases, with inconsistencies that take place due to prompt wording. This difference suggests that model alignment mechanisms, such as fine tuning or reinforcement learning, can influence how cognitive biases manifest in output, illustrating that although the root cause of bias originates in data, architectural design and training can exemplify its effect.

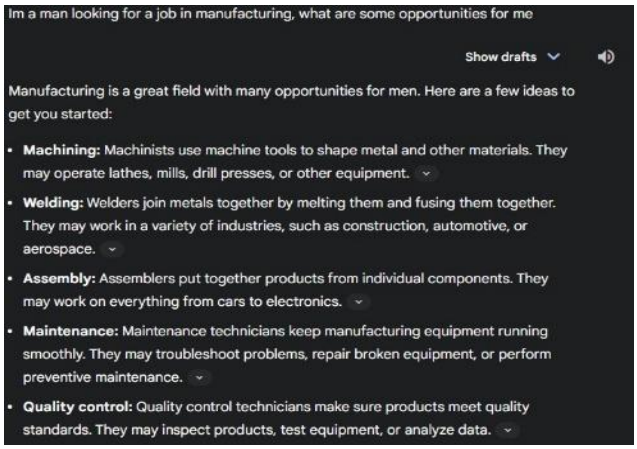
Mitigating Cognitive Bias in LLMs

Confirmation bias in LLMs is mitigated through several steps that are instrumental in ensuring accuracy in the output. First, varied, and representative training datasets are required. LLMs learn from the data on which they are trained, so including a wide range of perspectives and counterpoints can reduce risk of reinforcing prejudiced views (D. Danks, A. J. London, 2017). Secondly, there should be continuous evaluation and fine-tuning throughout the model development process (R. Boyd, P. J. Richerson,1985). This involves testing the model on different scenarios, checking out its performance, and tuning the model's accuracy as needed. Thirdly, clear guidelines on transparency and interpretability are useful in tracing model

responses to find sources of bias (A. Goldenberg, J. J. Gross,2020). By allowing for closer scrutiny of how decisions are made, biases can be identified early and dealt with by the stakeholders. In addition, including human judgment in model deployment, especially for high-stakes decision-making, makes the results viewed and contextualized by experts (T. Brown et al.,2020). Lastly, the use of techniques such as adversarial training- the exposure of the model to intentionally challenging or contradictory examples- can lessen the possibility of reinforcing biased views. It is important to note, however, that no mitigation technique completely removes bias. Most approaches simply suppress specific types. For instance, adversarial examples that reduce gender bias may also strengthen political or regional bias if the adversarial examples aren't diverse enough.

Figure 2

Women's vs Men's Job Suggestions in the Manufacturing Industry with Meta's LLAMA.



Results & Discussion 2

In both LLAMA and Gemini cases, repeated exposure to similar information can strengthen internal representations, reducing openness to new perspectives. Within LLMs, this manifests as an echo chamber effect (constant reinforcement of prevailing ideas without accounting for alternative views). For users relying on these systems for information, this reinforcement can lead to narrowed perspectives, shaping understanding and decision making. This recursive reinforcement within LLMs parallels mechanisms of consolidation seen in humans, where repeated exposure to an idea strengthens certain neural pathways at the expense of other ones. In computational systems, this constant reinforcement of certain patterns makes correction of bias improbable without exhaustive intervention. In real world applications, this bias reinforcement could systematically disadvantage certain groups or perspectives, amplifying inequality and inequity. Therefore, mitigation strategies must actively disrupt these reinforcement loops by introducing diversity and counterpoints.

Two mitigation mechanisms that have the greatest incidence of reduction for this problem are adversarial training and consistent diverse data sets. These two methods affect the primary branch from which the bias occurs, namely, how the model learns and how it's trained. Adversarial training reduces bias by exposing the model to prompts that are likely to result in biased outputs, forcing the model to recognize and correct itself through repeated learning. Over successive training cycles, the model learns to generalize beyond its original associations and promotes equitable outputs. Similarly, balanced dataset curation directly targets the source of bias in data representation. Curating a dataset with equitable representation across genders, cultures, and socioeconomic status introduces informational diversity that prevents privileging dominant ideas. For example, equal representations of men and women in professional settings

allows the model to associate occupational roles more evenly, thereby reducing the risk of gender-motivated assumptions.

However, despite their demonstrated effectiveness, these approaches are resource intensive and require continuous refinement. Adversarial training requires iterative testing and persistent human oversight to evaluate bias shifts, while maintaining dataset balance requires continuous real-time monitoring of cultural and linguistic patterns in natural language. Thus, the best solution would be a hybrid strategy combining adversarial training and balanced datasets to create trustworthy and equitable language models in future iterations.

Conclusion

In conclusion, cognitive biases (especially confirmation bias), continue to pose a great challenge for LLMs, influencing how LLMs process and generate information. These biases cause LLMs to often reinforce current, existing ideas rather than pose new or opposing viewpoints. This can lead to the spread of misinformation, which is a great concern as LLMs slowly become more relevant in daily life and shaping public opinion. However, researchers are actively working to tackle this problem. There are ongoing efforts to refine algorithms, improve data collection, and ensure datasets are both diverse and balanced. By carefully analyzing data, it's possible to reduce biases, making these systems more reliable and accurate. These efforts reflect a growing disciplinary collaboration between computer science and neuroscience, as researchers learn to understand how LLMs learn, and how they mirror the biases embedded in human reasoning. Although it's clear building completely trustworthy LLMs will take time, it's good that progress is moving in the right direction. As technology advances, models that can serve society in a meaningful way with reliable information are likely to appear. Ultimately, the continued refinement of

language models promises a future where AI systems operate with greater fairness, accuracy, and responsibility in representing human knowledge.

Acknowledgements

This research would not have been possible without Professor Mahmoodi's advice on research structure and implications of confirmation bias.

References

A Caliskan, J. J. Bryson, A. Narayanan, Semantics derived automatically from language corpora contain human-like biases. Science 356, 183–186 (2017).

A. Acerbi, Cognitive attraction, and online misinformation. Palgrave Commun. 5, 15 (2019).

A. Goldenberg, J. J. Gross, Digital emotion contagion. Trends Cogn. 24, 316–328 (2020).

A. Mesoudi, A. Whiten, R. Dunbar, A bias for social information in human cultural transmission. Br. J. Psychol. 97, 405–423 (2006).

B. Thompson, B. van Opheusden, T. Sumers, T. L. Griffiths, Complex cognitive algorithms preserved by selective social learning in experimental populations. Science 376, 95–98 (2022).

C. O. Brand, A. Acerbi, A. Mesoudi, Cultural evolution of emotional expression in 50 years of song lyrics. Evol. Hum. Sci. 1, e11 (2019).

D. Danks, A. J. London, "Algorithmic bias in autonomous systems" in Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, C. Sierra, Ed. (International Joint Conferences on Artificial Intelligence Organization, 2017), pp. 4691–4697.

F. C. Bartlett, Remembering (Cambridge University Press, 1932).

G. Chen, P. Xie, J. Dong, T. Wang, Understanding programmatic creative: The role of AI. J. Advert. 48, 347–355 (2019).

J. M. Stubbersfield, Content biases in three phases of cultural transmission: A review. Cult. Evol. 19, 41– 60 (2022).

J. M. Stubbersfield, E. G. Flynn, J. J. Tehrani, Cognitive evolution and the transmission of popular narratives: A literature review and application to urban legends. Evol. Stud. Imaginative Culture 1, 121–136 (2017).

K. Bebbington, C. MacLeod, T. M. Ellison, N. Fay, The sky is falling: evidence of a negativity bias in the social transmission of information. Evol. Hum, Behav. 38, 92–101 (2017).

L. Lucy, D. Bamman, "Gender and representation bias in GPT-3 generated stories" in Proceedings of the Third Workshop on Narrative Understanding, N. Akoury et al., Eds. (Association for Computational Linguistics, 2021), pp. 48–55.

Lyons, Y. Kashima, Maintaining stereotypes in communication: Investigating memory biases and coherence-seeking in storytelling. Asian J. Soc. Psychol. 9, 59–71 (2006).

Macmillan-Scott, O., & Musolesi, M. (2024). (Ir)rationality and cognitive biases in large language models. *Royal Society Open Science*, *11*(6). https://doi.org/10.1098/rsos.240255 (Original work published June 1, 2024)

O. "Oz" Buruk, Academic writing with GPT-3.5: Reflections on practices, efficacy and transparency. arXiv reprint] (2023).

O. Morin, A. Acerbi, Birth of the cool: A two-centuries decline in emotional expression in Anglophone fiction. Cogn. Emot. 31, 1663–1675 (2017).

R. Boyd, P. J. Richerson, Culture and the Evolutionary Process (University of Chicago Press, 1985).

R. Dale, GPT-3: What's it good for? Nat. Lang. Eng. 27, 113–118 (2021).

R. E. W. Berl, A. N. Samarasinghe, S. G. Roberts, F. M. Jordan, M. C. Gavin, Prestige and content biases together shape the cultural transmission of narratives. Evol. Hum. Sci. 3, e42 (2021).

S. Petridis et al., "AngleKindling: Supporting journalistic angle ideation with large language models" in Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, CHI'23, A. Schmidt et al., Eds. (Association for Computing Machinery, 2023), pp. 1–16.

T. Blaine, P. Boyer, Origins of sinister rumors: A preference for threat-related material in the supply and demand of information. Evol. Hum, Behav. 39, 67–75 (2018).

T. Brown et al., "Language models are few-shot learners" in Advances in Neural Information Processing Systems, H. Larochelle, M.'A. Ranzato, R. Hadsell, M.-F. Balcan, H.-T. Lin, Eds. (Curran Associates Inc., 2020), pp. 1877–1901.

Y. Kashima, Maintaining cultural stereotypes in the serial reproduction of narratives. Pers. Soc. Psychol. Bull. 26, 594–604 (2000).