Stars, Stripes, and Silicon: Unravelling the ChatGPT's All-American, Monochrome, Cis-centric Bias

Stars, Stripes, and Silicon: Unravelling the ChatGPT's All-American, Monochrome, Cis-centric Bias

Federico Torrielli[0000−0001−8037−8828]

University of Torino, Department of Computer Science, Torino, Italy

Abstract. This paper investigates the challenges associated with bias, toxicity, unreliability, and lack of robustness in large language models (LLMs) such as ChatGPT. It emphasizes that these issues primarily stem from the quality and diversity of data on which LLMs are trained, rather than the model architectures themselves. As LLMs are increasingly integrated into various real-world applications, their potential to negatively impact society by amplifying existing biases and generating harmful content becomes a pressing concern. The paper calls for interdisciplinary efforts to address these challenges. Additionally, it highlights the need for collaboration between researchers, practitioners, and stakeholders to establish governance frameworks, oversight, and accountability mechanisms to mitigate the harmful consequences of biased LLMs. By proactively addressing these challenges, the AI community can harness the enormous potential of LLMs for the betterment of society without perpetuating harmful biases or exacerbating existing inequalities.

Keywords: Bias · LLM · ChatGPT.

1 Introduction

Bias, toxicity, unreliability and lack of robustness are interrelated issues that plague large language models (LLMs). Given that LLMs are utilised in various real-world applications, including language translation [51,13,45,34], search engines [2,1], and scientific literature summarisation [50], it is crucial that production-ready LLMs exhibit minimal bias and do not generate harmful content. However, the current state of language models faces significant challenges in this regard.

In this work, we analyse the issues that affect state-of-the-art language models such as ChatGPT, an RLHF-augmented [5] chatbot based on GPT-3.5 [11]. These problems stem less from the model architectures themselves and more from the fact that the models are trained on massive collections of uncurated data from the Internet [54]. While LLMs gain much of their knowledge and capabilities from the scale of data, we argue for the use of high-quality, curated datasets over stronger content filters as a solution. We discuss why this superficial approach is problematic.

Addressing problems like bias is crucial not only to ensure theoretical soundness but also to implement practically before LLMs become integrated into daily

technologies used by the general public. Their widespread adoption could amplify the negative societal effects of model flaws on vulnerable groups [57]. Techniques including data curation, model interpretability, and adversarial testing can help, but fully mitigating these issues will require ongoing collaboration between researchers and practitioners. Overall, the development of fair, ethical and trustworthy AI systems must be an interdisciplinary effort prioritised in the development of advanced technologies like LLMs.

2 The Bias Bazaar

LLMs generate responses with a coherent and fluent natural language structure, creating an illusion of authority and credibility. This presentation format implies an intelligence that encourages users to accept the outputs at face value, exacerbating the human tendency to trust autonomous systems that reduce cognitive load [49]. These factors could impede the ability to distinguish facts from falsehoods and rational from irrational reasoning in LLM outputs. In this context, accepting fake news, toxic content and bias is easy and seen as normal by a inexperienced user.

It is crucial to note that LLMs like chatbots currently generate all responses as text, though recent models such as GPT-4 experiment with multi-modal approaches [43]. Each step towards humanising these interfaces makes it increasingly difficult to approach them with a critical perspective. Despite this, humans tend to view text as more credible and accurate than other media, given vision's dominance as a sense [53,41].

Biases represent the discrepancy between rational and heuristic behaviour [52,3]. As of 2023, cognitive science has identified innumerable cognitive biases [7,8,32,47,55] which can lead to flawed reasoning, irrationality, and potentially harmful consequences [49]. Prominent biases studied in the literature include cultural, gender, nationality, political, and ethical biases.

While recent work has examined the presence of specific biases in various models [49], the scale and complexity of LLMs today make comprehensive auditing and remediation challenging. As models continue to increase in capability and adoption, governance frameworks, oversight, and accountability are urgently needed. Reliance on biased algorithms and data can directly and negatively impact marginalised groups through unfair treatment, discrimination, or by influencing consequential decisions.

3 ChatGPT Waves the American Flag

Recent progress in LM design has led to increasingly large models that demonstrate strong capabilities in various natural language tasks [9]. However, larger models also introduce and amplify biases present in their training data [56,9]. For models trained primarily on raw data ingestion, the characteristics of the training data have significant influence on model performance and biases.

For example, contemporary language models like GPT are trained on datasets comprised primarily of American English data [36]. This restricted data source limits the diversity of perspectives and linguistic knowledge that can be acquired by the model. As language models become more capable and ubiquitous, it is crucial to consider the ramifications of biases that can be perpetuated and even exacerbated in these systems. Addressing this issue will require developing methods for building models that learn from diverse, high-quality data as well as techniques for identifying and mitigating biases. Overall, the capabilities and limitations of large language models depend strongly on the data used to train them. For a model, language bias is statistical-sampling bias, and ultimately the latter is knowledge bias [36].

A naive approach to addressing lack of diversity in datasets would be to simply increase the size of the training set, with the expectation that a larger dataset would inherently capture greater diversity. However, as noted by Bender et al. [9], increased size alone does not necessarily guarantee increased diversity. The underlying motivation here is that perspective is linked to language itself (as exemplified by the 'elephant' example discussed in [36]), and language represents a powerful and distinguishing feature in how information is filtered and conveyed.

4 The Larger They Get, the Larger Their Shadow is

The representation of viewpoints in large language models (LLMs) is primarily governed by frequency, which is an inherent aspect of their architecture and not easily modifiable [40]. Biases often originate from extensive, unfiltered corpora and can persist even when safety filters are employed in the architecture [43]. The dominance of frequency in LLMs gives rise to a critical issue: the under-representation of less frequent data. Consequently, majority viewpoints tend to overshadow minority perspectives.

For instance, Wikipedia is frequently among the most representative sources in corpora used for LLM training. However, its content is predominantly authored by males, with female contributors constituting less than 15% of the total [6]. Furthermore, training datasets such as CommonCrawl[1] and The Pile [16] have been found to contain high levels of toxic, racist, or sexist content [17]. This underscores the significance of carefully selecting and curating training data to mitigate biases in LLMs.

The widespread adoption of the "*bigger is better*" paradigm in the context of large language models presents both ethical and computational challenges. While it is evident that larger training datasets yield improved performance for LLMs, the necessity of human involvement in dataset curation or generation cannot be ignored. This labour is frequently carried out by crowdworkers who receive inadequate compensation, lack essential protections, and are exposed to harmful content throughout their workday [20,23,28,29]. Moreover, due to the sheer size and continuous growth of these datasets, assessing their quality in terms of bias identification and toxicity presence becomes increasingly difficult.

[1] https://commoncrawl.org/the-data/

An additional concern arises from the primary source of LLM data: private corporations. The majority of LLM research, models, and datasets are developed by these entities, as LLM-related processes are resource-intensive and consequently, amplify corporate influence [48]. This dynamic generates significant implications for the scientific community. Corporations prioritise product development over research, leading to the prevalence of proprietary technology, which inherently obstructs free access to the underlying research and perpetuates their dominance in the field [19].

5 Complete the Sentence: All you Need is... [Violence]

The current trend in LLMs for secure human-chatbot interactions involves reinforcement learning with human feedback (RLHF) [22,14] and standard safety mechanisms. RLHF ensures safety in typical "naive" interactions [57,21], while safety mechanisms have been found to be less reliable against prompt injection [21]. A majority of prompt injection techniques utilise storytelling, an effective method capable of diverting LLMs like ChatGPT from generating innocuous content and instead producing harmful narratives. The underlying motivations behind this phenomenon remain unclear; however, some attribute it to a semiotic-simulation theory known as "*The Waluigi Effect*"[2], wherein the creation of an ideal simulation environment allows the LLM greater freedom to improvise. As larger and more sophisticated models increasingly exhibit a tendency to reproduce human common misconceptions [35], it is anticipated that this issue will continue to exacerbate unless more effective countermeasures are developed and implemented.

Revising the harmful content in primary training datasets is the most effective approach for mitigating the generation of toxic output from large language models. It is widely acknowledged that amidst the vast data sources used for training GPT-x models, harmful and toxic content can be found; these models' initial datasets encompass data from unreliable news sites as well as quarantined and banned subreddits [17]. Even when present in smaller quantities, such data has been shown to be more salient for the model and considerably more challenging to "unlearn" [30,12].

With minimal or no prompting, models have been observed to generate potent and offensive content targeting minority groups and LGBTQIA+ individuals [42], thereby supporting the aforementioned hypothesis.

6 All Bias is Language Bias

A prevalent misconception is that these models exhibit cognitive bias, akin to those found in human decision-making [49]. However, it is essential to clarify that large language models do not possess cognitive bias, as they lack cognitive abilities [37]. Instead, the biases observed in these models stem from language

[2] https://www.lesswrong.com/posts/D7PumeYTDPfBTp3i7/the-waluigi-effect-mega-post

bias inherent in the data they are trained on [36]. Cognitive biases emerge from cognitive processes, which involve conscious and unconscious thinking, perception, memory, and problem-solving. In contrast, large language models, including ChatGPT, function as complex pattern-recognition algorithms that learn to generate text based on statistical correlations within the data they have been trained on, rather than exhibiting any form of cognition or understanding. Language bias arises from the inherent biases present in the training data, which are a reflection of human culture, values, and beliefs. Consequently, these biases can be observed in the generated text, leading to potential misunderstandings about the presence of cognitive bias. To better understand language bias in large language models, the relationship between training data and generated text must be examined. As these models learn from vast amounts of text, they inevitably acquire the biases present in those texts.

7 Words are Powerful: Unintended Consequences of Real-World AI Misadventures

In light of the growing concerns about bias in LLMs like ChatGPT, it is essential that these models are developed to be explainable, transparent, unbiased, fair, verifiable, and accountable for every decision [39,15]. Despite these requirements, many current LLMs are deployed without fully addressing these issues, raising questions about whether our expectations are too high or if these models are not yet ready to be products. The accelerated release of unsafe models by companies may be contributing to the difficulty in refining these models to meet these criteria, necessitating further research.

Furthermore, the extended use of unsafe models in daily life has the potential to jeopardise crucial sectors such as healthcare, medicine, code safety, journalism, online content, and spam prevention, among others [24,26]. Models like GPT-4[3] have been shown to be capable of creating misinformation scenarios and spreading toxic and biased content with ease [43]. It is imperative that the development and deployment of LLMs prioritise safety and responsibility to prevent adverse consequences in these vital areas.

7.1 Healthcare and Medicine

Positive Impacts: ChatGPT can streamline the healthcare industry by providing quick and accurate responses to common medical questions [31,18], thereby saving time for medical professionals. Additionally, it can aid in the analysis of medical records and help identify patterns or trends that might otherwise go unnoticed.

Negative Impacts: If ChatGPT is not properly trained or its knowledge base is outdated, it may provide incorrect or potentially harmful medical advice. This could lead to dangerous consequences for patients and healthcare providers.

[3] Pre-alignment model

7.2 Code Safety

Positive Impacts: ChatGPT can serve as an effective tool for code review [25], detecting potential bugs, and suggesting improvements to existing code. This could improve overall software quality and reduce the likelihood of security vulnerabilities.
Negative Impacts: For now, ChatGPT generates flawed or unsafe code suggestions [27], and it could inadvertently introduce security risks or software bugs in the future, compromising the safety and reliability of the developed software.

7.3 Journalism

Positive Impacts: ChatGPT can assist journalists in drafting articles quickly and efficiently. It can also help in the generation of news summaries, translations, and content personalisation for readers [44].
Negative Impacts: The potential for ChatGPT to generate biased or misleading content poses a risk to journalistic integrity [44,46,33]. If unchecked, it could contribute to the spread of misinformation and undermine public trust in news sources.

7.4 Online Content

Positive Impacts: ChatGPT can help generate engaging and relevant content for websites, blogs, and social media platforms, assisting content creators and marketers in their efforts to reach and captivate audiences.
Negative Impacts: The ease with which ChatGPT can generate content may result in an oversaturation of low-quality or misleading information online. Additionally, it could be used to create and spread fake news, deepfakes, and other manipulative content [46,15,4].

7.5 Spam Prevention

Positive Impacts: ChatGPT can be utilised to develop advanced spam filters, capable of identifying and blocking spam messages more effectively by understanding the semantic meaning of text, rather than relying solely on keywords or patterns.
Negative Impacts: Conversely, ChatGPT can also be employed by malicious actors to generate sophisticated spam messages that bypass existing filters [38,10], leading to an increase in unwanted and potentially harmful content in users' inboxes.

8 Conclusion: the Avalanche Effect

In conclusion, the concerns surrounding bias, toxicity, unreliability, and lack of robustness in large language models (LLMs) such as ChatGPT are multifaceted

and significant. This paper highlights that the primary challenge lies in the diversity of data on which LLMs are trained. As these models become increasingly integrated into daily technologies and real-world applications, their potential to negatively impact society by amplifying existing biases and generating harmful content is intensified.

One particularly noteworthy concern we should raise is the possible *"avalanche effect"*, wherein future LLMs could inadvertently include generated content from previous LLMs in their training data. This effect could result in a self-perpetuating loop of biased and potentially harmful content being propagated across generations of models, exacerbating the issues that already plague these systems. Consequently, it is crucial for researchers and practitioners to develop methods for mitigating these problems and ensuring the development of fair, ethical, and trustworthy AI systems.

References

1. `https://metaphor.systems/`
2. Perplexity ai, `https://www.perplexity.ai/`
3. Ahmad, Z., Ibrahim, H., Tuyon, J.: Institutional investor behavioral biases: syntheses of theory and evidence. Management Research Review **40**, 578–603 (2017)
4. Alkaissi, H., McFarlane, S.I.: Artificial hallucinations in chatgpt: implications in scientific writing. Cureus **15**(2) (2023)
5. Bai, Y., Jones, A., Ndousse, K., Askell, A., Chen, A., et al.: Training a helpful and harmless assistant with reinforcement learning from human feedback. ArXiv **abs/2204.05862** (2022)
6. Barera, M.: Mind the gap: Addressing structural equity and inclusion on wikipedia (2020)
7. Basta, C., Costa-jussà, M.R., Casas, N.: Evaluating the underlying gender bias in contextualized word embeddings. ArXiv **abs/1904.08783** (2019)
8. Beltagy, I., Lo, K., Cohan, A.: Scibert: A pretrained language model for scientific text. In: Conference on Empirical Methods in Natural Language Processing (2019)
9. Bender, E.M., Gebru, T., McMillan-Major, A., Shmitchell, S.: On the dangers of stochastic parrots: Can language models be too big? Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (2021)
10. Borji, A.: A categorical archive of chatgpt failures. arXiv preprint arXiv:2302.03494 (2023)
11. Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., et al.: Language models are few-shot learners (2020)
12. Carlini, N., Liu, C., Erlingsson, Ú., Kos, J., Song, D.X.: The secret sharer: Evaluating and testing unintended memorization in neural networks. In: USENIX Security Symposium (2018)
13. Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., et al.: Palm: Scaling language modeling with pathways. ArXiv **abs/2204.02311** (2022)
14. Christiano, P., Leike, J., Brown, T.B., Martic, M., Legg, S., et al.: Deep reinforcement learning from human preferences (2023)
15. van Dis, E.A.M., Bollen, J., Zuidema, W., van Rooij, R., Bockting, C.L.H.: Chatgpt: five priorities for research. Nature **614**, 224–226 (2023)

16. Gao, L., Biderman, S., Black, S., Golding, L., Hoppe, T., et al.: The pile: An 800gb dataset of diverse text for language modeling (2020)
17. Gehman, S., Gururangan, S., Sap, M., Choi, Y., Smith, N.A.: Realtoxicityprompts: Evaluating neural toxic degeneration in language models. In: Findings (2020)
18. Gilson, A., Safranek, C.W., Huang, T., Socrates, V., Chi, L., et al.: How does chatgpt perform on the united states medical licensing examination? the implications of large language models for medical education and knowledge assessment. JMIR Medical Education **9** (2023)
19. Goetze, T.S., Abramson, D.: Bigger isn't better: The ethical and scientific vices of extra-large datasets in language models. 13th ACM Web Science Conference 2021 (2021)
20. Gray, M.L., Suri, S.: Ghost work: How to stop silicon valley from building a new global underclass (2019)
21. Greshake, K., Abdelnabi, S., Mishra, S., Endres, C., Holz, T., et al.: More than you've asked for: A comprehensive analysis of novel prompt injection threats to application-integrated large language models. ArXiv **abs/2302.12173** (2023)
22. Griffith, S., Subramanian, K., Scholz, J., Isbell, C.L., Thomaz, A.L.: Policy shaping: Integrating human feedback with reinforcement learning. In: NIPS (2013)
23. Hara, K., Adams, A., Milland, K., Savage, S., Callison-Burch, C., et al.: A data-driven analysis of workers' earnings on amazon mechanical turk. Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (2017)
24. Harrer, S.: Attention is not all you need: the complicated case of ethically using large language models in healthcare and medicine. eBioMedicine **90** (2023)
25. He, J., Vechev, M.T.: Large language models for code: Security hardening and adversarial testing (2023)
26. Jones, E., Steinhardt, J.: Capturing failures of large language models via human cognitive biases. ArXiv **abs/2202.12299** (2022)
27. Khoury, R., Avila, A.R., Brunelle, J., Camara, B.M.: How secure is code generated by chatgpt? ArXiv **abs/2304.09655** (2023)
28. Kittur, A., Nickerson, J.V., Bernstein, M.S., Gerber, E., Shaw, A., et al.: The future of crowd work. Proceedings of the 2013 conference on Computer supported cooperative work (2013)
29. Kneese, T., Rosenblat, A., Boyd, D.: Understanding fair labor practices in a networked age (2014)
30. Koh, P.W., Liang, P.: Understanding black-box predictions via influence functions. ArXiv **abs/1703.04730** (2017)
31. Kung, T.H., Cheatham, M., Medenilla, A., Sillos, C., Leon, L.D., et al.: Performance of chatgpt on usmle: Potential for ai-assisted medical education using large language models. PLOS Digital Health **2** (2022)
32. Kurita, K., Vyas, N., Pareek, A., Black, A.W., Tsvetkov, Y.: Measuring bias in contextualized word representations. ArXiv **abs/1906.07337** (2019)
33. Leiter, C., Zhang, R., Chen, Y., Belouadi, J., Larionov, D., et al.: Chatgpt: A meta-analysis after 2.5 months. arXiv preprint arXiv:2302.13795 (2023)
34. Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., et al.: Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In: Annual Meeting of the Association for Computational Linguistics (2019)
35. Lin, S.C., Hilton, J., Evans, O.: Truthfulqa: Measuring how models mimic human falsehoods. In: Annual Meeting of the Association for Computational Linguistics (2021)

36. Luo, Q., Puett, M.J., Smith, M.D.: A perspectival mirror of the elephant: Investigating language bias on google, chatgpt, wikipedia, and youtube. ArXiv **abs/2303.16281** (2023)
37. Mahowald, K., Ivanova, A.A., Blank, I.A., Kanwisher, N.G., Tenenbaum, J.B., et al.: Dissociating language and thought in large language models: a cognitive perspective. ArXiv **abs/2301.06627** (2023)
38. Mansfield-Devine, S.: Weaponising chatgpt. Network Security **2023**(4) (2023)
39. Meo, R., Nai, R., Sulis, E.: Explainable, interpretable, trustworthy, responsible, ethical, fair, verifiable ai... what's next? In: Symposium on Advances in Databases and Information Systems (2022)
40. Mills, M., Whittaker, M.: Disability, bias, and ai (2019)
41. Moore, T., Zirnsak, M.: Neural mechanisms of selective visual attention. Annual Review of Psychology **68**(1), 47–72 (2017). https://doi.org/10.1146/annurev-psych-122414-033400, https://doi.org/10.1146/annurev-psych-122414-033400, pMID: 28051934
42. Nozza, D., Bianchi, F., Lauscher, A., Hovy, D.: Measuring harmful sentence completion in language models for lgbtqia+ individuals. In: LTEDI (2022)
43. OpenAI: Gpt-4 technical report (2023)
44. Pavlik, J.V.: Collaborating with chatgpt: Considering the implications of generative artificial intelligence for journalism and media education. Journalism & Mass Communication Educator p. 10776958221149577 (2023)
45. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., et al.: Language models are unsupervised multitask learners (2019)
46. Rudolph, J., Tan, S., Tan, S.: Chatgpt: Bullshit spewer or the end of traditional assessments in higher education? Journal of Applied Learning and Teaching **6**(1) (2023)
47. Sheng, E., Chang, K.W., Natarajan, P., Peng, N.: The woman worked as a babysitter: On biases in language generation. ArXiv **abs/1909.01326** (2019)
48. Stallman, R.M.: Free software, free society: Selected essays of richard m. stallman (2009)
49. Talboy, A.N., Fuller, E.: Challenging the appearance of machine intelligence: Cognitive bias in llms. ArXiv **abs/2304.01358** (2023)
50. Taylor, R., Kardas, M., Cucurull, G., Scialom, T., Hartshorn, A.S., et al.: Galactica: A large language model for science. ArXiv **abs/2211.09085** (2022)
51. team, N., Costa-jussà, M.R., Cross, J., cCelebi, O., Elbayad, M., et al.: No language left behind: Scaling human-centered machine translation. ArXiv **abs/2207.04672** (2022)
52. Tversky, A., Kahneman, D.: Judgment under uncertainty: Heuristics and biases: Biases in judgments reveal some heuristics of thinking under uncertainty (1978)
53. Wang, H.C., Lu, S., Lim, J.H., Pomplun, M.: Visual attention is attracted by text features even in scenes without text. Cognitive Science **34** (2012)
54. Wei, C., Wang, Y.C., Wang, B., Kuo, C.C.J.: An overview on language models: Recent developments and outlook. ArXiv **abs/2303.05759** (2023)
55. Zhang, H., Lu, A.X., Abdalla, M., McDermott, M.B.A., Ghassemi, M.: Hurtful words: quantifying biases in clinical contextual word embeddings. Proceedings of the ACM Conference on Health, Inference, and Learning (2020)
56. Zhao, W.X., Zhou, K., Li, J., Tang, T., Wang, X., et al.: A survey of large language models. ArXiv **abs/2303.18223** (2023)
57. Zhuo, T.Y., Huang, Y., Chen, C., Xing, Z.: Exploring ai ethics of chatgpt: A diagnostic analysis. ArXiv **abs/2301.12867** (2023)