Stereotypes in AI Stories: Identifying Societal Bias in Narratives Generated by Large Language Models

Louis Escouflaire

Catholic University of Louvain (UCLouvain), Ruelle de la Lanterne Magique 14, 1348 Ottignies-Louvain-la-Neuve, Belgium

Abstract

Large language models (LLMs) can generate written stories with unprecedented efficiency, but their outputs are highly dependent on their training data. This paper presents a research project in which we plan to analyze biases in AI-generated stories by examining how LLMs replicate societal stereotypes, especially regarding culture and gender inequalities. On a corpus of stories generated on diverse topics, characters, in different genres and languages, we will use content analysis, lexicography, and natural language processing tools, including explainability methods, to uncover and highlight biases in stories generated by LLMs. Findings will contribute to developing ethical AI systems that promote diverse and accurate representation in storytelling, helping to prevent the amplification of harmful stereotypes in AI-generated stories.

Keywords

Large Language Models, Generative AI, Story generation, Bias, Explainability, Lexicography

1. Introduction

Myths, whether ancient or modern, have first been characterized by historian Eliade [1] both as models shaping the ways societies view the world and as a form of collective beliefs that reproduce the ideologies, values and power structures of the group of people who convey them. Because stories evolve with society, mirroring its transformations and challenges, shaped by cultural, historical, and personal contexts [2], they also encode the biases and stereotypes inherent to it [3]. For instance, scholars have argued that all kinds of narratives (e.g. myths, news, children's literature) have historically marginalized or distorted women's voices, reproducing and strengthening patriarchal norms [4, 5].

In recent years, the booming rise of large language models (LLMs) has introduced new dynamics into the world of storytelling. With significant advancements in natural language processing and machine learning, generative AI models like GPT-4 [6] and Llama-3 [7] can produce stories that mimic human writing. The availability of these models and their integration in chatbots such as ChatGPT has offered unprecedented speed and the ability to generate any kind of stories on demand [8]. However, LLMs are trained on enormous datasets, and the stories that they generate are inevitably dependent on the data that they absorbed. Most of the data on which models such as GPT-4 were trained consists of texts written on the Internet since its creation: not only research papers, Wikipedia pages and movie scripts, but also tweets, fanfiction and blog posts [9]. Just as myths are reflections of a society's culture and biases, it is very likely that AI-generated stories deeply capture and display open and latent beliefs of modern Western societies, such as biases related to gender inequalities [10].

AI-generated stories are appearing more and more across various domains, from creative fiction to journalism and interactive entertainment, changing the way we approach narrative creation [11]. Without conscious intervention or mechanisms to detect and mitigate these biases, the use of LLMs in storytelling risks reinforcing existing inequalities and potentially creating new forms of narrative exclusion, where certain voices, especially those belonging to societal minorities, are further silenced or distorted [12]. Scheduled to start in Summer 2025, the research project introduced in this paper aims to address these questions, leveraging methods from corpus linguistics, narratology and natural language processing.

2. Methods

By exploiting open-source LLMs such as BLOOM [13] or Cedille [14], we will build a corpus of stories generated by models with transparent training data and parameters, unlike GPT-4, giving us greater control over the criteria relevant to the analysis. The first step involves generating stories based on defined prompts, engineered to produce diversity across variables such as character gender, profession, setting, and narrative structure to ensure a comprehensive range of story dynamics [15]. The corpus will include stories on a carefully curated sample of topics and characters, which will allow us to evaluate how LLMs handle representation of cultural backgrounds and gender roles. Comparing similar narratives generated in different languages, such as English and French, will help identify whether patterns of bias are linguistically dependent or consistent across languages. The project will also include stories produced in different genres, like journalistic or crime stories. To identify patterns of bias in the corpus, we will use methods from discourse analysis and narratological frameworks, analyzing how elements such as story structure, character agency and narrative voices reflect or deviate from societal stereotypes. The patterns identified will contribute to raise awareness about the potential amplification of societal biases through the dissemination of AI-generated stories.

While LLMs like GPT-4 can produce coherent narratives, they simply generate stories by relying on statistical distributions of words [16]. They lack a deep understanding of plot structures, character dynamics, or narrative progression [17]. In addition to qualitative content analysis of the stories, computational tools relying on lexicography and higher-level representations of the corpus will therefore also be leveraged. Using tools and packages developed in Python for lexicographical analysis of large text corpora will help extend our identification of societal biases. For example, collocation analysis may reveal patterns of word associations, highlighting how specific terms are more frequently used together in some narrative contexts than in others [18]. Sentiment analysis can help quantify emotional biases in the portrayal of characters or themes within AI-generated stories [19]. Topic modeling and clustering algorithms will reveal whether certain perspectives or characters are systematically marginalized or ignored [20]. Using models of word embeddings [21] will also allow us to detect subtle associations in the corpus. For instance, word embeddings can reveal whether certain professions or traits are disproportionately associated with male or female characters or whether non-Western cultural markers are underrepresented or mischaracterized. By comparing embeddings trained in different languages, we can also evaluate whether and how cultural and linguistic nuances impact the biases in AI-generated narratives.

Finally, we will use model explainability techniques recently developed in the field of natural language processing to track how and why certain prompts lead to the presence of societal bias in specific AI stories. Although LLMs are more powerful and accurate than traditional machine learning models in many natural language processing tasks, they suffer from two crucial flaws: their require substantial computational resources [22], and their rapidly increasing number of

parameters (GPT-4 has 1.7 trillion) create a significant lack of transparency in the inner workings of the model [23]. In the context of automated story generation, the use of such black-box models maker it very difficult to know which elements in a user's input led the model to prioritize one narrative pattern over another. Recent research has focused on exploring different methods used to explain the outputs made by LLMs [24], such as model perturbation with LIME [25] and SHAP [26], or attention probing [27]. LLMs rely on the *transformer* neural network architecture, which uses a mechanism called 'attention' to weigh the importance of different words or tokens in a sequence when generating text [28]. In this project, we will leverage the attention mechanism to interpret the weights attributed to textual elements in the prompts provided to the model, which will allow us to highlight which features of the input prompts are most influential in shaping some AI-generated stories over others [29].

3. Conclusion

Through different computational and qualitative methods, our research project aims to critically explore stereotypes and omissions inherent to stories generated by LLMs, such as underrepresentation of societal minorities, and to bring these biases to light. The patterns identified will contribute to raise awareness about the potential amplification of societal biases through the dissemination of AI-generated stories [30]. Ultimately, this research project will contribute to the growing body of knowledge on ethical AI use, especially in storytelling, helping to inform the development of more transparent and fair narrative generation tools that better represent diverse voices and perspectives.

References

[1] M. Eliade, *Mythes, rêves et mystères*. Les Essais, Gallimard, 1961.

[2] J. Bruner, "Narrative, Culture, and Mind", *Telling Stories: Language, Narrative, and Social Life*, *46*, 2010.

[3] J. Gottschall, *The Storytelling Animal: How Stories Make us Human*. Houghton Mifflin Harcourt, 2012.

[4] H. Lovatt, *The Epic Gaze: Vision, Gender and Narrative in Ancient Epic.* Cambridge University Press, 2013.

[5] K. Casey, K. Novick, S. F. Lourenco, "Sixty Years of Gender Representation in Children's Books: Conditions Associated with Overrepresentation of Male versus Female Protagonists", *PLoS One, 16(12)*, 2021.

[6] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, et al., GPT-4 Technical Report, 2023. URL: https://arxiv.org/abs/2303.08774.

[7] A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, et al., *The Llama-3 Herd of Models*, 2024 URL: https://arxiv.org/abs/2407.21783.

[8] P. P. Ray, "ChatGPT: A Comprehensive Review on Background, Applications, Key Challenges, Bias, Ethics, Limitations and Future Scope", *Internet of Things and Cyber-Physical Systems, 3*, 2023, 121-154.

[9] S. Baack, "A Critical Analysis of the Largest Source for Generative AI Training Data: Common Crawl", *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, 2024, 2199-2208.

[10] L. Li, D. Bamman, "Gender and Representation Bias in GPT-3 Generated Stories", *Proceedings of the 3rd Workshop on Narrative Understanding*, 2021. 48-55.

[11] P. Cardon, C. Fleischmann, J. Aritz, M. Logemann, J. Heidewald, “The Challenges and Opportunities of AI-assisted Writing: Developing AI Literacy for the AI Age”, *Business and Professional Communication Quarterly, 86(3)*, 2023, 257-295.
[12] J. Barroso Da Silveira, E. Alves Lima, “Racial Biases in AIs and Gemini’s Inability to Write Narratives About Black People”, *Emerging Media 2(2)*, 2024, 277–287.
[13] T. Le Scao, A. Fan, C. Akiki, E. Pavlick, S. Ilić, D. Hesslow, et al., “BLOOM: A 176B-parameter Open-access Multilingual Language Model”, *CoRR,* 2023, URL: https://arxiv.org/abs/2211.05100.
[14] M. Müller, F. Laurent, Cedille: A Large Autoregressive French Language Model, 2022. URL: https://arxiv.org/pdf/2202.03371.
[15] S. Harmon, S. Rutman, “Prompt Engineering for Narrative Choice Generation. In International Conference on Interactive Digital Storytelling”, *Cham: Springer Nature Switzerland*, 2023. 208-225.
[16] E. M. Bender, T. Gebru, A. McMillan-Major, S. Shmitchell, “On the dangers of stochastic parrots: Can language models be too big?” *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 2021, 610-623.
[17] N. Montfort, R. Pérez y Pérez, "Computational Models for Understanding Narrative”, *Revista de Comunicação e Linguagens, 58*, 2023, 97-117.
[18] P. Guilherme, “Collocation and semantic prosody–a gauge for bias in political discourse?”, *Confluência. Revista Hispánica de Cultura y Literatura, 60*, 2021, 273-289.
[19] A. M. Jacobs, B. Herrmann, G. Lauer, J. Lüdtke, J., S. Schroeder, “Sentiment analysis of children and youth literature: is there a pollyanna effect?” Frontiers in Psychology, 11, 2020, 574746.
[20] P. Rao, M. Taboada, “Gender bias in the news: A Scalable Topic Modelling and Visualization Framework”, *Frontiers in Artificial Intelligence, 4*, 2021, 664737.
[21] T. Bolukbasi, K. W. Chang, J. Y., Zou, V. Saligrama, A. T. Kalai, “Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings”, *Advances in Neural Information Processing Systems, 29*, 2016.
[22] A. S. Luccioni, S. Viguier, A.-L. Ligozat, A.-L. "Estimating the Carbon Footprint of BLOOM, a 176B-Parameter Language Model”, *Journal of Machine Learning Research, 24(1)*, 2024.
[23] L. Escouflaire, A. Descampe, C. Fairon, "Automated Text Classification of Opinion vs. News French Press Articles. A Comparison of Transformer and Feature-based Approaches”, *Language & Communication, 99*, 2024, 129-140.
[24] Q. Lyu, M. Apidianaki, C. Callison-Burch, “Towards Faithful Model Explanation in NLP: A Survey”, *Computational Linguistics, 70*, 2024.
[25] M. T. Ribeiro, S. Singh, C. Guestrin, *Model-Agnostic Interpretability of Machine Learning,* 2016. URL: https://arxiv.org/abs/1606.05386
[26] S. Lundberg, G. G. Erion, S. I. Lee, *Consistent Individualized Feature Attribution for Tree Ensembles*, 2018. URL: https://arxiv.org/abs/1802.03888.
[27] Chefer, H., Gur, S., & Wolf, L. (2021). Transformer interpretability beyond attention visualization. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 782-791.
[28] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. & Polosukhin, I. (2017). Attention is all you Need. Advances in Neural Information Processing Systems, 30.
[29] Bogaert, J., de Marneffe, M. C., Descampe, A., Escouflaire, L., Fairon, C., & Standaert, F. X. (2023). Sensibilité des explications à l’aléa des grands modèles de langage: le cas de la classification de textes journalistiques. Traitement Automatique Des Langues, 64(3).
[30] Rettberg, J. W. (2024). How Generative AI Endangers Cultural Narratives. Issues in Science and Technology 40 (2), 77-79.