Exploring Gender Bias in LLM-Generated Hero and Heroine Narratives

Irene C. E. van Blerck[1,2(✉)] and Edirlei Soares de Lima[1]

[1] Academy for AI, Games and Media, Breda University of Applied Sciences, Breda, The Netherlands
blerck.i@buas.nl, soaresdelima.e@buas.nl

[2] Department of Computer Science, University of Antwerp, Antwerp, Belgium

Abstract. Narrative structures such as the Hero's Journey and Heroine's Journey have long influenced how characters, themes, and roles are portrayed in storytelling. When used to guide narrative generation in systems powered by Large Language Models (LLMs), these structures may interact with model-internal biases, reinforcing traditional gender norms. This workshop examines how protagonist gender and narrative structure shape storytelling outcomes in LLM-based storytelling systems. Through hands-on experiments and guided analysis, participants will explore gender representation in LLM-generated stories, perform counterfactual modifications, and evaluate how narrative interpretations shift when character gender is altered. The workshop aims to foster interdisciplinary collaborations, inspire novel methodologies, and advance research on fair and inclusive AI-driven storytelling in games and interactive media.

Keywords: Storytelling · Large Language Models · Gender Bias · Counterfactuals

1 Introduction

The increasing adoption of Large Language Models (LLMs) in storytelling raises important concerns about the persistence and reinforcement of gender biases. Recent studies have demonstrated that LLMs frequently reflect and, in some cases, amplify gender stereotypes present in the training data [6]. In narrative generation, prior research has shown that LLMs tend to associate female characters with themes of family and appearance while portraying them with significantly less agency than male characters [16]. Furthermore, the influence of literary genres and narrative conventions on LLM-generated texts complicates efforts to disentangle biases inherent to Artificial Intelligence (AI) models from those embedded in the storytelling traditions they replicate [5]. While some studies suggest that LLMs may introduce progressive gender representations in certain contexts [1], others reveal persistent biases in how AI systems associate gender with professions and social interactions [17]. Beyond gender, research has

Published by Springer Nature Switzerland AG 2025
R. Yamanishi et al. (Eds.): ICEC 2025 Workshops, LNCS 15935, pp. 11–16, 2025.
https://doi.org/10.1007/978-3-032-02534-0_3

also identified structural biases in narrative outcomes, such as a strong preference for positive endings in game-related stories [21,22].

Despite these findings, existing work has largely overlooked the role of narrative structures in shaping gender biases in AI-generated storytelling. Most studies focus on word-level associations, thematic biases, or profession-based stereotypes but do not systematically examine how protagonist gender interacts with structured storytelling frameworks, such as the Hero's Journey [2] and Heroine's Journey [18]. Prior work has demonstrated the importance of narrative structures in managing character arcs and plot progression, particularly in interactive environments [9–11], yet little attention has been given to how these structures may influence gender stereotypes when operationalized in LLM-based generation systems. Furthermore, prior research has not applied counterfactual estimation to assess causal effects, leaving open the question of whether LLM biases primarily stem from this interaction or if they are significantly influenced by other, potentially unobserved factors.

Building on our experience developing LLM-based narrative generation systems [7,8,12–15], this workshop aims to address key gaps in current research by providing a hands-on exploration of gender bias in LLM-generated storytelling. It focuses specifically on the interaction between protagonist gender and narrative structure – an area that remains underexamined despite its influence on the generated narratives. Participants will engage in practical experiments, including direct bias analysis and counterfactual generation using state-of-the-art LLMs. Additionally, the workshop will foster discussion on bias mitigation strategies, encouraging participants to discuss different approaches for addressing gender bias in AI-generated storytelling.

The workshop's key objectives are to:

1. Examine gender bias in LLM-generated narratives by analyzing protagonist representation in different narrative structures.
2. Evaluate the impact of counterfactual narrative generation as a methodology for identifying and quantifying gender bias in storytelling systems.
3. Explore different bias mitigation strategies, providing participants the opportunity to discuss and assess approaches for promoting fairer and more inclusive LLM-based storytelling systems.

By integrating insights from ethics, computational creativity, and interactive storytelling, this workshop aims to advance the discourse on responsible AI in narrative generation and contribute to shaping future storytelling technologies that transcend traditional gender biases.

2 Exploring Gender Bias in LLM-Generated Narratives

This section presents the practical and analytical activities through which gender bias in LLM-generated narratives is investigated. Building on recent advances in interactive storytelling and LLM-based narrative generation, we outline a structured exploration that combines hands-on experimentation, comparative

analysis, and group discussion. The aim is to foster a deeper understanding of how narrative structures, protagonist gender, and model behavior interact to shape storytelling outputs.

2.1 Phase 1: Narrative Generation and Gender Bias Analysis

The exploration begins with the generation of fictional stories using PatternTeller [12,15], an AI-powered storytelling system guided by predefined narrative structures. Participants use the system to generate narratives based on neutral prompts and well-established narrative structures: the Hero's Journey [2] and the Heroine's Journey [18]. These generated narratives serve as the primary data for analysis.

Participants then evaluate the gender representation in their generated stories, focusing on protagonist identity, thematic associations, and role positioning. To complement qualitative interpretation, we apply the GenBit Score [20] – a metric for quantifying gender associations in text – using an interactive tool developed for this purpose. Through this process, participants identify patterns in how LLMs assign gender roles within given narratives and how these roles relate to societal stereotypes or literary conventions.

2.2 Phase 2: Counterfactual Narrative Generation and Structural Classification

In the second phase, we examine whether protagonist gender influences how LLMs interpret and classify stories in relation to the same narrative structures that guided their generation. To this end, participants apply a counterfactual approach using an LLM-assisted workflow, which modifies the gender of all characters in the stories they generated in Phase 1 while preserving all other narrative elements. These counterfactual versions serve as comparative counterparts to the factual narratives and form the basis for analyzing how shifts in gender representation may affect narrative classification outcomes.

Both the original and counterfactual narratives are then classified through an LLM-assisted workflow, using the same model that generated the stories. The classification prompt asks the model to determine whether each narrative aligns more closely with the Hero's Journey or the Heroine's Journey by comparing the story to key structural stages. Participants review the resulting classifications and analyze cases where a shift in protagonist gender leads to a different structural interpretation. This process offers insight into how gendered assumptions may influence model behavior, with a particular focus on recurring misclassification patterns and instances where stereotypical associations appear to override the narrative's actual content.

2.3 Phase 3: Bias Mitigation Strategies and Collective Reflection

The final phase involves a reflective group discussion on observed biases and potential strategies to mitigate them in LLM-driven storytelling systems. Drawing from examples generated during earlier phases, participants discuss diverse

mitigation approaches. These include fine-tuning or Reinforcement Learning with Human Feedback (RLHF) [23], where models are aligned with human preferences through curated reward signals to discourage biased outputs; debiasing post-processing [3], which adjusts generated content to reduce stereotype exposure or gender imbalance; output filtering or re-ranking [19], where multiple generations are evaluated and filtered to select those that demonstrate more inclusive or balanced representations; and system-level interventions [4], which integrate external control layers or workflows that monitor and guide LLM behavior without altering the model itself. Ethical considerations – such as the trade-offs between creative freedom, fairness, and interpretability – are also critically discussed. This dialogue aims to seed future collaborations and encourage the development of new methodologies for fair and inclusive AI storytelling.

3 Expected Outcomes

This workshop aims to stimulate critical discussions on gender bias in LLM-generated narratives while equipping participants with practical skills to analyze and mitigate such biases. By engaging with hands-on experiments and structured discussions, attendees will gain insights into how narrative structures and protagonist gender influence the outputs of LLM-based storytelling systems. A key outcome is to raise awareness of the challenges in ensuring fair and unbiased LLM-generated narratives, particularly in entertainment applications such as games and interactive storytelling.

Beyond the immediate learning experience, this workshop is intended to facilitate new research collaborations and interdisciplinary projects. Insights gained from discussions may contribute to the development of new research initiatives, including collaborative publications and project proposals exploring strategies for mitigating bias in AI storytelling. We aim to facilitate networking opportunities that could lead to long-term collaborations between researchers, developers, and practitioners working in AI-driven creative fields.

As a tangible post-workshop outcome, we will invite interested participants to co-author a position or research paper summarizing workshop discussions, emergent insights, and open challenges. We also plan to form a working group to explore the possibility of forming a working group to develop project proposals focusing on bias mitigation in AI storytelling, with potential applications in game development, interactive narratives, and ethical AI research.

4 Conclusion

As LLM-based storytelling systems are increasingly used for narrative generation, examining how LLMs internalize and reproduce gender norms becomes a critical area of research. This workshop responds to that need by offering a collaborative environment where participants engage directly with LLM-generated narratives, observe how bias emerges in relation to narrative structure, and reflect on the broader implications for storytelling systems.

Rather than providing definitive solutions, this workshop emphasizes the open questions around fairness, creativity, and accountability in LLM-driven storytelling. By encouraging critical discussion and hands-on experimentation, it aims to lay the groundwork for sustained interdisciplinary research at the intersection of narrative theory, machine learning, and ethics. The insights developed through this workshop are intended not only to inform technical design but also to challenge and refine our collective understanding of what inclusive and responsible storytelling should look like in the age of generative AI.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this workshop.

References

1. Begus, N.: Experimental narratives: a comparison of human crowdsourced storytelling and AI storytelling (2023). https://arxiv.org/abs/2310.12902
2. Campbell, J.: The Hero With a Thousand Faces. New World Library (2008)
3. Ghanbarzadeh, S., Huang, Y., Palangi, H., Moreno, R.C., Khanpour, H.: Gender-tuning: empowering fine-tuning for debiasing pre-trained language models. In: Rogers, A., Boyd-Graber, J., Okazaki, N. (eds.) Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, pp. 5448–5458. Association for Computational Linguistics (2023). https://doi.org/10.18653/v1/2023.findings-acl.336
4. Huang, D., Zhang, J.M., Bu, Q., Xie, X., Chen, J., Cui, H.: Bias testing and mitigation in LLM-based code generation. ACM Trans. Softw. Eng. Methodol. (2025). https://doi.org/10.1145/3724117
5. Jackson, D., Courneya, M.: Unreliable narrator: reparative approaches to harmful biases in AI storytelling for the he classroom and future creative industries. Braz. Creat. Ind. J. **3**(2), 59–75 (2023). https://doi.org/10.25112/bcij.v3i2.3540
6. Kotek, H., Dockum, R., Sun, D.: Gender bias and stereotypes in large language models. In: Proceedings of The ACM Collective Intelligence Conference, CI 2023, pp. 12–24. Association for Computing Machinery, New York (2023). https://doi.org/10.1145/3582269.3615599
7. de Lima, E.S., Casanova, M.A., Feijó, B., Furtado, A.L.: Semiotic structuring in movie narrative generation. In: Ciancarini, P., Di Iorio, A., Hlavacs, H., Poggi, F. (eds.) Entertainment Computing – ICEC 2023, pp. 161–175. Springer, Singapore (2023). https://doi.org/10.1007/978-981-99-8248-6_13
8. de Lima, E.S., Feijó, B., Casanova, M.A., Furtado, A.L.: ChatGeppetto - an AI-powered storyteller. In: Proceedings of the 22nd Brazilian Symposium on Games and Digital Entertainment, pp. 28–37. ACM (2024). https://doi.org/10.1145/3631085.3631302
9. de Lima, E.S., Feijó, B., Furtado, A.L.: Adaptive branching quests based on automated planning and story arcs. In: 2021 20th Brazilian Symposium on Computer Games and Digital Entertainment (SBGames), pp. 9–18 (2021). https://doi.org/10.1109/SBGames54170.2021.00012
10. de Lima, E.S., Feijó, B., Furtado, A.L.: Procedural generation of branching quests for games. Entertain. Comput. **43**, 100491 (2022). https://doi.org/10.1016/j.entcom.2022.100491

11. de Lima, E.S., Feijó, B., Furtado, A.L.: Managing the plot structure of character-based interactive narratives in games. Entertain. Comput. **47**, 100590 (2023). https://doi.org/10.1016/j.entcom.2023.100590
12. de Lima, E.S., Neggers, M.M.E., Casanova, M.A., Feijó, B., Furtado, A.L.: A pattern-oriented AI-powered approach to story composition. In: Figueroa, P., Di Iorio, A., Guzman del Rio, D., Gonzalez Clua, E.W., Cuevas Rodriguez, L. (eds.) Entertainment Computing – ICEC 2024, pp. 1–16. Springer, Cham (2024). https://doi.org/10.1007/978-3-031-74353-5_10
13. de Lima, E.S., Neggers, M.M.E., Furtado, A.L.: Multigenre AI-powered story composition (2025). https://arxiv.org/abs/2405.06685
14. de Lima, E.S., Neggers, M.M., Casanova, M.A., Furtado, A.L.: From images to stories: exploring player-driven narratives in games. In: Marto, A., Prada, R., Gouveia, P., Contreras-Espinosa, R., Gonçalves, A., Abrantes, E., Ribeiro, R. (eds.) Videogame Sciences and Arts, pp. 228–242. Springer, Cham (2025). https://doi.org/10.1007/978-3-031-81713-7_16
15. de Lima, E.S., Neggers, M.M., Feijó, B., Casanova, M.A., Furtado, A.L.: An AI-powered approach to the semiotic reconstruction of narratives. Entertain. Comput. **52**, 100810 (2025). https://doi.org/10.1016/j.entcom.2024.100810
16. Lucy, L., Bamman, D.: Gender and representation bias in GPT-3 generated stories. In: Proceedings of the Third Workshop on Narrative Understanding, pp. 48–55. Association for Computational Linguistics, Virtual (2021). https://doi.org/10.18653/v1/2021.nuse-1.5
17. Marin, A., Eger, M.: Towards evaluating profession-based gender bias in chatGPT and its impact on narrative generation. In: Farrokhimaleki, M., Rahmati, P., Saadat, K., Zhao, R. (eds.) Proceedings of the AIIDE Workshop on Intelligent Narrative Technologies co-located with the 20th AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment (AIIDE 2024), Lexington, Kentucky, USA (2024)
18. Murdock, M.: The Heroine's Journey: Woman's Quest for Wholeness. Shambhala (1990)
19. Nouriinanloo, B., Lamothe, M.: Re-ranking step by step: investigating pre-filtering for re-ranking with large language models (2024). https://arxiv.org/abs/2406.18740
20. Sengupta, K., Maher, R., Groves, D., Olieman, C.: GenBiT: measure and mitigate gender bias in language datasets. Microsoft J. Appl. Pharm. Res. **16**, 63–71 (2021)
21. Taveekitworachai, P., et al.: What is waiting for us at the end? inherent biases of game story endings in large language models. In: Holloway-Attaway, L., Murray, J.T. (eds.) Interactive Storytelling, pp. 274–284. Springer, Cham (2023)
22. Taveekitworachai, P., Plupattanakit, K., Thawonmas, R.: Assessing inherent biases following prompt compression of large language models for game story generation. In: 2024 IEEE Conference on Games (CoG), pp. 1–4 (2024). https://doi.org/10.1109/CoG60054.2024.10645609
23. Ziegler, D.M., et al.: Fine-tuning language models from human preferences (2020). https://arxiv.org/abs/1909.08593