Balancing Bias & Creativity: Fine Tuning LLM Agents to Refine User Prompts Ethically

Author: Alex Zwakenberg

Affiliations: Hogeschool Utrecht, Monks

Generative AI has revolutionized creative industries but inherits biases that can influence user prompts and outputs. Current bias-mitigation methods focus on post-generation fixes, often restricting creativity. This research explores **fine-tuning LLM agents** to **detect** and **refine biased prompts** in real time while preserving creative intent. By developing and evaluating a prototype LLM agent, this study aims to support ethical, inclusive AI-driven creativity. Findings will inform AI development in fields like media and advertising, ensuring that AI fosters innovation without reinforcing harmful biases, contributing to more responsible and effective generative AI applications.

INTRODUCTION

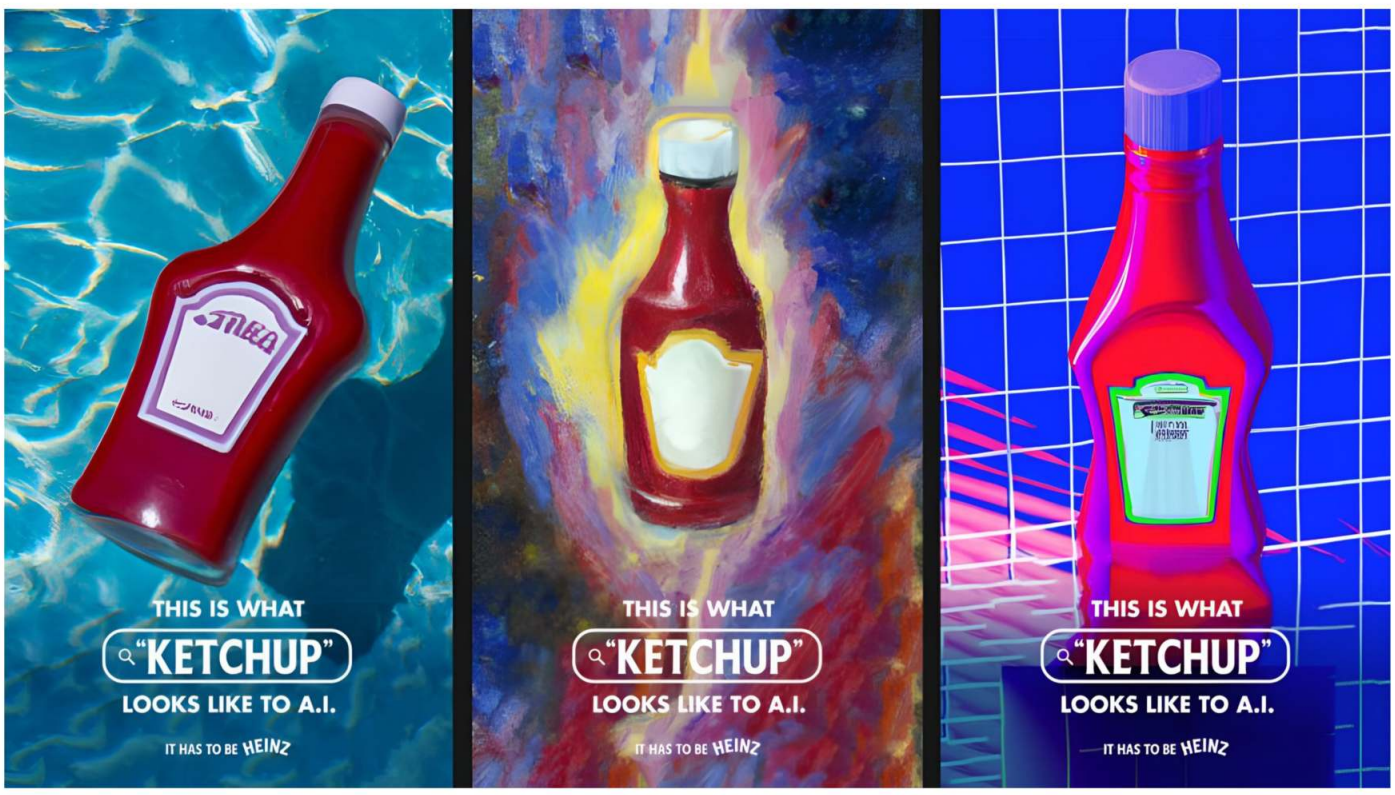
Generative AI (GenAI) models, especially large language models (LLMs), have transformed the creative industry by enhancing ideation and automating tasks. However, these models are **inherently prone to biases** inherited from their training data, which can manifest in outputs, potentially reinforcing stereotypes and presenting ethical challenges [1]. Existing methods often focus on post-generation filtering but **neglect choices made in prompts**, limiting user control and creativity. This research explores fine-tuning LLM agents to detect and mitigate bias in real-time while preserving artistic intent. By developing a **prototype that refines prompts**, we aim to help creative professionals generate diverse, ethically sound outputs.

Hypothesis:

LLM agents fine-tuned to improve user prompts in real-time can effectively reduce biases while maintaining creative intent, producing more ethically sound and diverse outputs compared to unrefined prompts or post-generation methods.

LITERATUREREVIEW

The integration of large language models (LLMs) in creative tasks raises challenges in balancing creativity and ethical responsibility. Models like PromptDoctor [2] aim to optimize prompts for improved outputs but **struggle with complex ethical issues** and the risk of undermining creative intent. Research stresses the need for systems that detect and mitigate bias, **ensuring that generated content adheres to ethical standards** without stifling creativity [3]. Tools like NBIAS have been developed to detect **textual biases**, yet most current solutions address



Birds, M. (2024, August 24). 5 Amazing AI-Powered Ads Shaping The Marketing World! Marketing Birds. https://themarketingbirds.com/5-amazing-ai-powered-ads-shaping-the-marketing-world/

bias in text not images, **leaving the root problem in image generation untouched** [4]. Additionally, while a prompt itself may not contain inherent bias, LLMs can still reinforce stereotypes present in the training data, reflecting the biases learned during model training [5].

Generated using Adobe Express

This research will explore **how LLM agents can refine prompts in real-time** to address biases, preserve creativity, and maintain ethical integrity in generative AI applications.

RESEARCH METHODOLOGY

The dataset will consist of diverse user-generated prompts collected from open-source platforms and previous generative AI research. These prompts will serve as the baseline to fine-tune and test the LLM agent. To evaluate the LLM agent multiple HBO ICT research methods will be applied including literature studies, interviews, benchmark tests and model evaluations improving **triangulation** of the results [6].

PRELIMINARY CONSIDERATIONS

This study explores balancing ethics and creativity in generative AI, helping creative professionals produce inclusive, diverse, and ethical content. Findings will guide AI tool development in industries like advertising, media, and art, where responsible AI is key. By **refining prompts early**, this research ensures ethically aligned yet creatively rich outputs. It may also inspire further work on real-time AI feedback, broadening AI's applications.

CONCLUSION

This research seeks to balance creativity and ethical responsibility in generative AI by developing LLM agents that refine user prompts in real-time. By **mitigating and flagging biases** in prompts while preserving creative intent, the study will help creative professionals produce more diverse and ethically sound content. The findings have the potential to revolutionize AI-driven ideation across industries, ensuring that AI outputs aligns with ethical standards while fostering innovation. This research will contribute to more inclusive and responsible AI applications in creative fields.

References

[1] Rzig, D. E., Paul, D. J., Pister, K., Henkel, J., & Hassan, F. (2025). An Empirically-grounded tool for Automatic Prompt Linting and Repair: A Case Study on Bias, Vulnerability, and Optimization in Developer Prompts (No. arXiv:2501.12521). arXiv. https://doi.org/10.48550/arXiv.2501.12521

[2] Raza, S., Garg, M., Reji, D. J., Bashir, S. R., & Ding, C. (2023). NBIAS: A Natural Language Processing Framework for Bias Identification in Text (No. arXiv:2308.01681). arXiv. https://doi.org/10.48550/arXiv.2308.01681

[3] Amankwah-Amoah, J., Abdalla, S., Mogaji, E., Elbanna, A., & Dwivedi, Y. K. (2024). The impending disruption of creative industries by generative AI: Opportunities, challenges, and research agenda. International Journal of Information Management, 79, 102759. https://doi.org/10.1016/j.ijinfomgt.2024.102759

[4] HBO-i. (n.d.). ICT Research Methods—Methods Pack for Research in ICT. ICT Research Methods. Retrieved January 29, 2025, from https://ictresearchmethods.nl/

[5] Chen, C., Gong, X., Liu, Z., Jiang, W., Goh, S. Q., & Lam, K-Y. (2025). Trustworthy, Responsible, and Safe AI: A Comprehensive Architectural Framework for AI Safety with Challenges and Mitigations (No. arXiv:2408.12935). arXiv. https://doi.org/10.48550/arXiv.2408.12935

[6] Zhou, M., Abhishek, V., Derdenger, T., Kim, J., & Srinivasan, K. (2024). Bias in Generative AI (No. arXiv:2403.02726). arXiv. https://doi.org/10.48550/arXiv.2403.02726

The HCAIM (the Human-Centred AI Master's Programme) Project is Co-Financed by the Connecting Europe Facility of the European Union Under Grant NoCEF-TC-2020-1 Digital Skills 2020-EU-IA-0068. This poster was created as part of the Blended Intensive Programme organised under the Erasmus + Programme of the European Union.