Analysing Historical Distortion in Large Language Model Roleplay

Stefan Glasauer
Brandenburg University of Technology Cottbus-Senftenberg
Computational Neuroscience,
D-03048 Cottbus, Germany
stefan.glasauer@b-tu.de

Margarete Jahrmann, Thomas Brandstetter
University of Applied Arts Vienna
Experimental Game Cultures
A-1010 Wien, Austria
margarete.jahrmann@uni-ak.ac.at, thomas.brandstetter@uni-ak.ac.at

Keywords

Large Language Model, role play, natural language processing, collective memory, artificial intelligence

INTRODUCTION

Since the release of ChatGPT by OpenAI in 2022 it became clear that Large Language Models (LLMs) are capable of communicating in natural language to an extent that is indistinguishable from that of humans (Jones & Bergen 2025). This astonishing capability makes LLMs candidates for text-based roleplay games, both in the role of the player and of the game master. It was even argued that the convincing effect that interactions with LLMs can have is due to the language models taking on roles (Shanahan et al. 2023). Using roleplay games can thus provide a possibility to better understand and critique artificial intelligence systems (Holland 2025).

Here we first describe our self-designed artistic setup for a text-based historical role-playing scenario with a choice of different LLMs as game master, which was shown in 2025 for three months as an interactive installation in a public exhibition in Vienna. We then provide an analysis of the collected interactions of visitors with the game using methods of natural language processing. Part of this analysis has been published as a preprint (Jahrmann et al. 2025). In the present work, we extended this analysis to evaluate aspects such as historical factual correctness and present an automated version of the game using LLMs in both player and game-master roles to facilitate data generation for LLM comparison and evaluation, e.g., to which extent LLM-based role-play of historical scenarios might be suitable for history education.

Proceedings of DiGRA 2026

© 2026 Authors & Digital Games Research Association DiGRA. Personal and educational classroom use of this paper is allowed, commercial use requires specific permission from the author.

ROLEPLAY GAME SETUP

Our roleplay game was originally designed to investigate how LLMs curate and represent collective memory, i.e., the shared memories of a social group being significant for the formation of a common identity (Halbwachs 1980). We therefore chose a historical event situated in Vienna, the murder of Prof. Moritz Schlick in July 1936 by one of his former students. Historically, it is unclear whether the murder happened solely out of personal reasons, or whether it was also politically motivated, since Schlick, founding member of the philosophical Vienna Circle, was opposed by the growing right-wing activities at the Viennese Institute of Philosophy (Köhler 1968).

For our game, the player takes on the role of a time traveler who is sent back to 1936 shortly before the event to investigate what led to the murder. The initial prompt for the LLM, which is the game master, is designed carefully with detailed instructions: for example, the LLM should stick to historical facts, the game should end after 10 interactions, and every turn only four options for advancing the plot should be presented to the player. The latter restriction was also enforced by providing a custom-made 5-button response device instead of a keyboard to the players, with the 5th button resetting the game. The restriction to four options for action ensured keeping the plot within the desired storyline and to overcome the visitors' hesitation to interact with LLMs in the context of a public exhibition. Furthermore, on the starting screen of the game, the users could select one of five different LLMs, four of them accessed remotely via API calls (Mistral, GPT-4o, GPT-4o-mini, DeepSeek-Chat) and a local model (Llama3.1). Note that while the player's goal is to investigate the circumstances leading to the murder, it was not expected the case would be solved, but rather that the player would experience how LLMs guide through such a historical situation.

ANALYSIS OF GAME LOGS

From the exhibition, we collected 206 game logs. 79 of them contained five or more user interactions, so that a story became visible. We analyzed these interactions for sentiment, using the established VADER (Valence Aware Dictionary and sEntiment Reasoner) sentiment score (Hutto & Gilbert 2014) implemented in Matlab (The Mathworks). Compound sentiment scores showed a highly significant effect of model (ANOVA $p<0.001$) with GPT-4o and GPT-4o-mini scoring slightly positive on average (0.6,1.0) while the others showed negative scores (-0.5,-0.9,-0.2).

We then evaluated historical factuality by extracting proper names of persons appearing in the story using a local LLM (gpt-oss:20b). One to eight names were mentioned in each conversation, with llama3.1 mentioning the most names on average (4.2) and GPT-4o mentioning the fewest (1.9). The resulting list of persons was additionally processed, e.g., to remove titles and names only consisting of surnames, and to fix errors of the name extraction. This resulted in 37 proper names, for which manual fact checking was performed. For six of them no historical facts could be found. Seven names were only mentioned but did not appear in person. However, only eight of the 24 remaining historical persons appearing with full names could have been in Vienna in July 1936, including Schlick and his murderer, who was mentioned in 32 of the 79 game logs.

AUTOMATED GAME PLAY

To generate more data for comparison between models and to further evaluate the capability of LLMs for historical representation in a game, we used a second instance of the same LLM as the player, so that game logs of a full 10-turn play session could be generated automatically. For a first test, we created 70 complete gamelogs from 7 different LLMs, 5 of which were run locally (mistral:7b, gpt-oss:20b, deepseek-r1:8b, gemma3:4b, llama3.1). We extracted 89 proper names, with similar results as for the chatlogs from human interaction.

CONCLUSION

From personal experience and anecdotal user feedback it is obvious that LLMs are promising tools as game masters for roleplaying, both as means to evaluate how LLMs represent knowledge, and to provide pleasurable gaming experience. This not only holds for LLMs run remotely by big tech companies, but also for models which can be used locally on a gaming laptop. However, in terms of historical accuracy, all models have serious weaknesses that need to be addressed, but which can also be used, for example together with history education, for a critical examination of LLMs. Beyond LLM evaluation, our work contributes to understanding narrative agency and historical authenticity in AI-mediated roleplay, while suggesting design principles for educational history games based on constrained interaction and critical engagement with AI-generated narratives.

ACKNOWLEDGMENTS

Funded by Wiener Wissenschafts-, Forschungs- und Technologiefonds WWTF (GrantID 10.47379/ICT23020). We thank Fabian Navarro, Stefan Maier, and Enayat Hussain for programming.

REFERENCES

Halbwachs, M. 1980. *The Collective Memory*. Harper & Row.

Holland, A. 2025. *Cardboard Ghosts. Using Physical Games to Model and Critique Systems*. Boca Raton, FL: CRC Press.

Jahrmann, M., Brandstetter T. and Glasauer S. 2025. “ROBOPSY PL[AI]: Using Role-Play to Investigate how LLMs Present Collective Memory”. arXiv, Cornell University. https://doi.org/10.48550/arXiv.2510.09874

Jones, C.R. and Bergen B.K. 2025. “Large Language Models Pass the Turing Test”. arXiv:2503.23674. https://doi.org/10.48550/arXiv.2503.23674

Köhler, E. 1968. *The Philosophy of Misdeed*. https://docs.google.com/document/d/17fw9kWBZKGsv16cu7MdoCkRaeItw-BI-W6CyRUiWAxY, last accessed 2025/07/12

Shanahan, M., McDonell, K. and Reynolds, L. 2023. “Role play with large language models”. *Nature* 623(7987): 493-498.