Human Annotation and Evaluation of Confabulations in Large Language Models

Česnik, M., Kapelj, M. B., Vidmar, M.

University of Ljubljana
mv71814@student.uni-lj.si

1 Introduction

Large language models (LLMs) enable fluent and coherent text generation but this fluency does not guarantee factual reliability. One major limitation of LLMs is their tendency to produce hallucinations or plausible but factually unsupported content. Because such outputs can appear highly convincing, they are difficult to detect and raise concerns for real-world use. Huang et al. (2024).

This project examines the prevalence of confabulations and the reliability of human annotation. Confabulations are defined as outputs that present unsupported, fabricated, or misleading information as factual. The main research questions are: (1) how frequently do confabulations occur and (2) how consistent are human judgments? The study is based on the concepts of factuality and faithfulness hallucinations.

2 Methods

Responses were generated using Llama 3.1 8B Instruct with prompts designed to elicit confabulations. A dataset of 608 responses was annotated by 3 cognitive science students, who labeled confabulation presence and type with short justifications. The prompts were presented in two conditions: history (with conversational context) and no-history. Confabulatory responses were classified into eight categories: knowledge gap, demand for precision, explicit constraint violation, false presupposition, false memory induction, authority appeal, grounding, and neutral confabulation. Inter-rater agreement was measured between annotators using intraclass correlation (ICC).

3 Results

Inter-rater agreement was moderate to high, with ICC(1,k) = 0.821 and ICC(1,1) = 0.605. Overall confabulation statistics are shown in Table 1.

Mean response length was 563.7 tokens, mean prompt length was 28.7 words, and 20.9% of responses exceeded 900 tokens.

Tab. 1: Confabulation statistics across all responses.

Metric	Value
Total responses	608
Confabulatory responses	359
Overall confabulation rate	59.05%

Figure 1 shows the distribution of confabulation types by history condition. The no-history condition contained a higher proportion of confabulatory responses, especially false presupposition and knowledge gap, whereas the with-history condition contained more non-confabulatory responses.

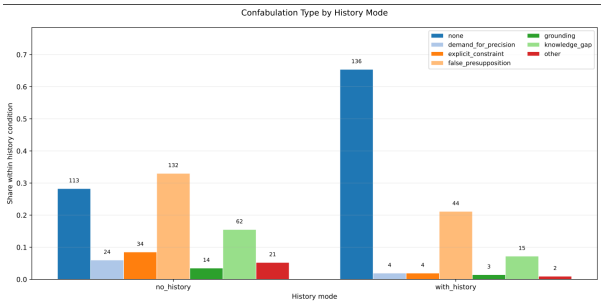

Fig. 1: Distribution of confabulation types by history condition.

4 Discussion

Results suggest human annotation is useful approach for evaluating LLM confabulations. Distinction between confabulation types enabled a more fine-grained analysis than binary detection alone, although some categories remained difficult to separate in borderline cases. No-history and with-history conditions suggests that conversational context may reduce some forms of confabulation.

As this study included only one model, the findings should be interpreted with caution. Future work should include more models, larger datasets, and automated methods for detecting confabulations.

Disclaimer. This research was conducted as part of the ARIS research project J7-70254, *Investigating Generative AI Confabulations by Comparison With Human Cognition: Physiological, Neurocognitive, and Philosophical Perspectives*.

References

Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B. and Liu, T. (2024). A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Office Information Systems*, 43(2).