LLM CITATION HALLUCINATION EVALUATION: A MULTI-METRIC ANALYSIS USING THE HAGRID DATASET

By
ZHENG Ying

Submitted to Central European University - Private University
Department of Economics and Business
Business Analytics Program
Expected Graduation: June 20, 2025

In partial fulfilment of the requirements for the degree of Master of Science in Business Analytics

Supervisor: Eduardo Arino de la Rubia
Project Sponsor: Berta Eszter Böjte, Hiflylabs Zrt.

Vienna, Austria

2025

COPYRIGHT NOTICE

[1]

[1] Icon by Font Awesome.

AUTHOR'S DECLARATION

I, ZHENG Ying, the undersigned candidate for the MSc degree in Business Analytics declare herewith that the present thesis titled "LLM Citation Hallucination Evaluation: A Multi-Metric Analysis Using the HAGRID Dataset" is exclusively my own work, based on my research and only such external information as properly credited in notes and bibliography. I declare that no unidentified and illegitimate use was made of the work of others, and no part of the thesis infringes on any person's or institution's copyright.

I also declare that no part of the thesis has been submitted in this form to any other institution of higher education for an academic degree.

Vienna, 09 June 2025

ZHENG Ying

TABLE OF CONTENTS

PROJECT SUMMARY

Large Language Models (LLMs), such as Gemini and ChatGPT, have achieved state-of-the-art performance in natural language generation tasks, including question answering (QA), summarization, and dialogue. However, despite their fluency and coherence, LLMs often suffer from hallucination—generating text that appears factually correct but lacks grounding in retrieved or known information. Among all hallucination types, citation hallucination—where LLMs generate incorrect, unsupported, or fabricated references—is especially problematic in knowledge-intensive applications such as academic writing, research assistance, or legal and medical QA systems.

This project focuses on evaluating and analyzing citation hallucination in Retrieval-Augmented Generation (RAG) pipelines, where an external retrieval module is used to fetch relevant documents for the LLM to condition its response. Using the HAGRID dataset, which contains 1,922 QA samples with associated ground-truth citations, the study systematically measures hallucination using four complementary evaluation metrics:

1. Retrieval Recall (Recall@k) – how many gold documents were retrieved;
2. Answer–Citation Alignment – whether cited documents in the answer match the gold citations;
3. TF-IDF Keyword Coverage – lexical overlap between answer and source;
4. Semantic Similarity – embedding-based similarity between answer and reference.

To avoid relying on a single perspective, hallucination is defined multi-dimensionally: a response is marked as hallucinated if it fails multiple metrics simultaneously. According to this operational definition, 65.0% of answers in the evaluated RAG pipeline were found to be hallucinated, with the majority of failure attributable to retrieval errors rather than issues in the language model's decoder. The findings highlight the critical role of retrieval quality in ensuring factual grounding.

Further analysis includes taxonomy of hallucination types (retrieval-only, citation-only, combined), and visualization of overlapping metric failures to identify systemic patterns. The study also reviews the impact of alternative reranking strategies and embedding models on hallucination rate, offering practical insights for future system design.

In conclusion, this project underscores the limitations of current RAG systems in citation-sensitive tasks and provides a rigorous, metric-rich framework for hallucination evaluation. It suggests that meaningful improvements in citation faithfulness may require not just better language models, but stronger retrieval and ranking pipelines, along with multi-perspective evaluation rather than single-metric assessments.