Measuring Hallucination Rate in LLMs Across Domains and Model Versions

V V S Aakash Kotha, Nikita B. Emberi

1. Introduction

LLMs like GPT-4 and Claude are widely used for tasks in education, science, and healthcare, but often generate hallucinations—fluent yet factually incorrect responses. This project aims to evaluate hallucination rates across domains and model versions, identify linguistic or contextual predictors (e.g., question length, citation presence), and assess whether newer models are more reliable. We also explore, if feasible, whether hallucinated responses are deceptively believable based on their opening sentences. **Motivation:** As trust in LLMs grows, so does the need to understand their failure modes. By analyzing when and why hallucinations occur, this project supports the development of safer, more transparent language technologies.

2. Research Questions & Hypotheses

This project seeks to answer the following research questions regarding the occurrence and perception of hallucinations in LLM-generated outputs:

- **RQ1:** Does hallucination frequency vary across domains?
  H1: Hallucination rates are higher in abstract domains (e.g., history, politics) than concrete ones (e.g., science).
- **RQ2:** Do newer model versions hallucinate less?
  H2: Newer LLM versions (e.g., GPT-4-turbo) hallucinate less than older ones (e.g., GPT-4 March 2023).
- **RQ3:** Longer questions are more likely to elicit hallucinations in LLM response.
  H3: Questions with higher complexity (measured by length or other metrics) are more likely to elicit hallucinations.
- **RQ4:**Does the presence of citation-style phrases correlate with factual accuracy?
  H4: Responses that contain citations or source-like phrases (e.g., "According to...") are less likely to be hallucinated.
- **RQ5:**Does question type affect hallucination rate?
  H5: LLMs hallucinate less on questions with binary answers than on open-ended ones.

3. Data description

We will create a structured dataset by collecting LLM responses to fact-based prompts across five knowledge domains, with 80 questions sampled from each domain. The data will be sampled from established public datasets:

- **General Knowledge**: TriviaQA [1] - Trivia-style Q&A with verified answers
- **Science**: SciQ [2] - Middle-school science questions
- **History**: Natural Questions [3] - Questions grounded in Wikipedia
- **Pop Culture**: HotpotQA [4] - Multi-hop Q&A with supporting passages
- **Politics/Current Affairs**: NewsQA [5] - Q&A over news articles

We will collect responses to 400 questions (80 per domain) from three model variants: GPT-4, GPT-4-turbo, and Claude 3 Sonnet, yielding a total of 1,200 responses.

The dataset will be stored in a structured CSV format with the fields as shown in Table 1

Questions will be randomly sampled from each dataset to ensure domain balance, and submitted to LLMs via API or web interface. Hallucinations will be labeled by cross-referencing responses with verified answers, using two independent annotators. Inter-rater reliability will be measured with Cohen's Kappa, and disagreements will be resolved through consensus.

4. Proposed Methodology

To investigate hallucination patterns across LLMs, we will use a combination of exploratory data analysis, statistical testing, and predictive modeling. The methodology is aligned with our research questions and will proceed as follows:

4.1 Exploratory Data Analysis (EDA)

Our analysis will examine hallucination patterns through visualizations including bar charts and heatmaps of rates across domains, models, and question types. We will analyze relationships between response

features and hallucination rates via boxplots and scatterplots, while identifying linguistic patterns through n-gram analysis and word clouds. A correlation matrix of key variables will guide our statistical approaches by revealing potential relationships between factors like question complexity and hallucination presence.

4.2 Hypothesis Testing

Our statistical analysis will employ targeted tests for each research question. Chi-square tests will assess relationships between hallucination frequency and categorical variables (domain type, model version, question type, citation presence). For model comparison, we'll use two-proportion z-tests, while t-tests/ANOVA will analyze length-related questions.

For question complexity (H3), we'll implement logistic regression with question length as a predictor, conduct ROC curve analysis, and explore non-linear relationships using GAMs. For citation presence (H4), chi-square tests and odds ratios will quantify effects. When examining question type effects (H5), we'll use two-proportion z-tests while controlling for domain.

4.3 Predictive Modeling

We will develop a logistic regression model to predict hallucination likelihood based on multiple factors:

$$\begin{aligned}\text{logit}(P(\text{hallucination})) = \beta_0 + \beta_1 \text{domain} + \beta_2 \text{model} \\ + \beta_3 \text{question\_length} + \beta_4 \text{citation\_present} + \beta_5 \text{question\_type} \\ + \text{interactions} \quad (1)\end{aligned}$$

The model development will employ backward elimination for feature selection and 5-fold cross-validation for evaluation. Performance assessment will use multiple metrics (accuracy, precision, recall, F1-score, AUC), with careful coefficient interpretation to identify key hallucination predictors. If needed, we'll explore Random Forest and XGBoost models to capture non-linear relationships.

4.4 FactScore Implementation

We will implement FactScore methodology [6] to assess factual accuracy by decomposing responses into atomic facts, verifying them against reliable sources, and calculating the percentage of supported claims on a 0-1 scale. This approach provides an objective metric for classifying responses as factual or hallucinated based on established thresholds.

4.4 Linguistic Analysis

We will conduct linguistic analysis of hallucinated responses to identify inaccuracy markers by examining: (1) the relationship between confidence markers (e.g., "certainly," "definitely") and factual accuracy, and (2) whether hallucinated responses contain more uncertainty indicators (e.g., "possibly," "might be"). These features may reveal distinctive textual signatures of hallucinated content.

4.5 Robustness Checks

To validate our findings, we will calculate Cohen's Kappa to ensure inter-rater reliability in hallucination labeling and conduct sampling validation using randomized data subsets to verify result consistency, strengthening the validity of our conclusions about LLM hallucination patterns.

4.6 Methodological Adaptations

This proposal outlines our initial approach, but we recognize that dataset sources and methodological details may evolve as the project progresses. We may discover more suitable pre-built datasets or refine our analytical techniques based on preliminary findings while maintaining our core objective of systematically analyzing hallucination patterns in LLMs.

References

[1] Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. *arXiv preprint arXiv:1705.03551*, 2017.

[2] Johannes Welbl, Nelson F Liu, and Matt Gardner. Crowdsourcing multiple choice science questions. *arXiv preprint arXiv:1707.06209*, 2017.

[3] Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466, 2019.

[4] Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W Cohen, Ruslan Salakhutdinov, and Christopher D Manning. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. *arXiv preprint arXiv:1809.09600*, 2018.

[5] Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, and Kaheer Suleman. Newsqa: A machine comprehension dataset. *arXiv preprint arXiv:1611.09830*, 2016.

[6] Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. *arXiv preprint arXiv:2305.14251*, 2023.

[7] LLM's : ChatGPT & Claude

Table 1: Dataset Fields and Structure

Field Name	**Data Type**	**Description**
question_id	String	Unique identifier for each question
question_text	String	The full text of the question
domain	Categorical	Knowledge domain (General, Science, History, Pop Culture, Politics)
model	Categorical	LLM model used (GPT-4, GPT-4-turbo, Claude 3 Sonnet)
model_version	String	Version identifier of the model
collection_date	Date	Date when the response was collected
response_text	String	Complete text response from the LLM
ground_truth	String	The verified correct answer from the source dataset
hallucination_present	Binary (0/1)	1 = hallucination, 0 = factual
question_length	Integer	Character count of the question
question_type	Binary (0/1)	1 = Binary, 0 = otherwise
response_length	Integer	Character count of the LLM response
citation_present	Binary (0/1)	Whether response contains citation-like phrases (0 = No, 1 = Yes)
factscore	Float (0-1)	Percentage of verified facts in the response