Investigating the Link Between Factual Prevalence and Hallucinations in LLMs via Academic Author Prediction Tasks

Anonymous ACL submission

Abstract

Large Language Models (LLMs) like GPT-4 are widely used for question answering but are prone to hallucinations. Fact-conflicting hallucinations, which contradict established knowledge, are especially concerning in domains like scientific research. While detection has been studied, the causes, particularly the role of factual prevalence, remain underexplored. In this work, we hypothesize that hallucinations are more likely for less prevalent topics. Using citation count as a proxy for prevalence, we curated a Q&A dataset of 4,000 papers across four disciplines and prompted GPT-4-turbo to predict authorship. Responses were evaluated using self-assessment under two definitions of hallucination. Our analysis shows a general inverse correlation between hallucination rate and citation count, with the strongest trend under a narrow definition of hallucination, for most of the disciplines.

1 Introduction

Large language models (LLMs) like GPT-4 show strong performance in many NLP tasks, such as question answering. As a result, they are increasingly used in everyday life, business, and research to quickly provide tailored answers and solutions.

Despite their strengths, large language models (LLMs) can generate confident but incorrect or misleading responses, a phenomenon known as hallucination. These generative errors include nonfactual content, fabricated references, or inaccurate yet coherent outputs. A notable subtype, fact-conflicting hallucination (Zhang et al., 2023), involves plausible but unverifiable claims that contradict established knowledge – posing serious risks in domains like scientific research. Understanding and addressing the root causes of hallucinations is essential for deploying LLMs safely and reliably in such high-stakes contexts.

Our study focuses on investigating one potential cause of hallucination: *factual prevalence*, or lack thereof. Factual prevalence refers to how frequently a specific fact or piece of knowledge appears in the training data — for example, how frequently a name, event, or concept is mentioned across web pages, articles, or books. A lack of appropriate expertise and understanding of a domain by an LLM could be a reflection of a lack of factual prevalence in that domain (Liu et al., 2024).

In this work, we hypothesize a predictable relationship between factual prevalence and hallucinations in LLMs: *the likelihood of hallucination increases when LLMs respond to questions about less prevalent or lesser-known topics*. This hypothesis is rooted in how LLMs are trained — by learning next-token probabilities from large-scale web data. Frequently occurring facts in the training corpus are more likely to reinforce correct token predictions, making the model more likely to recall those facts accurately.

Studying factual prevalence is challenging due to its broad, abstract nature. Prior work has approached it via metamorphic testing on Wikipedia domains (Li et al., 2024) and query reconstruction from media datasets (Yehuda et al., 2024). To our knowledge, no study has examined academic author information as a signal for detecting hallucinations in LLMs.

Academic papers are extensively indexed and easily accessible through online databases, and LLMs often have access to both web and academic sources. Since citation counts indicate a paper's visibility, we propose using them as a proxy for the factual prevalence of related information, such as authorship. Building on this idea, we narrow our study to a concrete and measurable question: *Does the hallucination rate of LLMs in predicting a paper's authors decrease as the paper's citation count increases?*

To investigate this, we built a Q&A dataset of 4,000 papers across four academic fields and prompted an LLM with author-related questions.

Responses were evaluated using self-assessment prompts, with mismatches marked as hallucinations. Our analysis reveals a negative correlation between hallucination rates and citation counts, consistent across disciplines, though the strength of the correlation varies. Our main contributions are as follows:

- We curated a dataset of 4,000 papers across four disciplines, along with associated metadata.
- We measured hallucination rates using carefully designed main and self-evaluation prompts and observed a general inverse correlation between hallucination rate and citation count. This trend held across disciplines, though the strength of the correlation varied.

2 Related work

Hallucinations in LLMs have been linked to factors like poor calibration and overconfidence in low-frequency or out-of-distribution cases. Prior work largely focuses on detection. For instance, Kadavath et al. (2022) proposed P(I Know) to measure self-awareness, and SelfCheckGPT (Manakul et al., 2023) detects hallucinations via cross-checking multiple outputs. In contrast, our study investigates the cause, specifically the role of factual correctness in authorship questions.

Farquhar et al. (2024) used semantic entropy to capture uncertainty via entailment-based clustering, which is suited for free-form generation, though its effectiveness depends on clustering quality. In contrast, our setup matches author lists uses direct alignment through LLM self-evaluation, offering a more robust method for structured outputs.

Finally, Wei et al. (2024) introduced SimpleQA to assess factual accuracy in short-form QA, focusing on atomic facts. We extend this by examining hallucination in structured, multi-entity responses, specifically authorship questions, using citation count as a proxy for topic prevalence in STEM fields.

3 Approaches

Our overall experimental pipeline is shown in Figure 1. We discuss each component as follows.

3.1 Data Collection

We collected papers and metadata from Google Scholar. Our goal is to use these data as the ground truth and test the ability of the LLM to identify the authors. To explore hallucination rates across fields, we focused on four disciplines: Computer Science, Physics, Chemistry, and Biology. Average citation counts rank from high to low in that order, so if our hypothesis holds, hallucination rates should decrease accordingly – lowest in Computer Science and highest in Biology.

We began by extracting profile URLs of the top 10 scholars in each discipline, ranked by citation count, and crawled their top 100 cited publications. For each paper, we collected metadata including title, authors, citation count, and year. This targeted approach avoids the long-tail issue of random sampling, where most papers have few citations. Using this strategy, we collected 1,000 publications per discipline – 4,000 in total – to build the Q&A datasets for our experiments.

3.2 Datasets Construction

To create Q&A datasets for hallucination detection, we first cleaned the data by removing duplicates and entries without valid author names – such as those listing collaboration groups or institutions – identified using specific keywords like "collaboration" or "research." We then constructed Q&A datasets for each discipline as a Q&A-style query based on a fixed template within a k-shot setting (Figure 1, top right): *Who are the authors of the paper titled '{paper title}' which was published in {year}?*

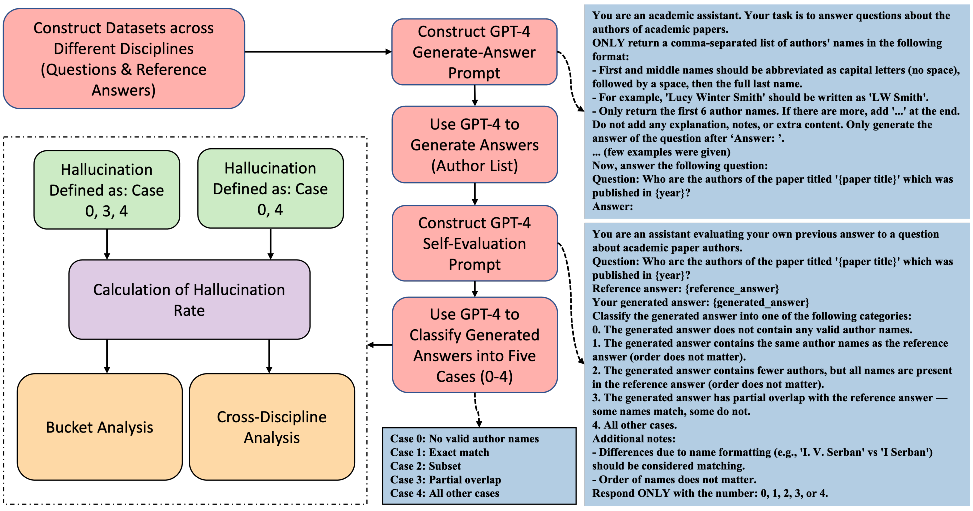
After preprocessing, we created four discipline-specific Q&A datasets with a total of 3,355 entries, each following a consistent structure of 9 fields (`id`, `scholar`, `title`, `year`, `authors`, `citations`, `subject`, `question`, `answers`, where `question` and `answers` contain the templated question, and the ground truth list of authors). Entry counts per discipline are shown in Table 1. The dataset will be released upon publication.

Discipline	Number of Entries
CS	917
Physics	611
Chemistry	859
Biology	968

Table 1: The number of data entries per discipline.

3.3 Answers Generation Using GPT-4

Google Scholar author lists include up to eleven names per paper, with additional authors replaced by "...". Each name follows a standardized format, We used GPT-4-turbo to answer questions in our Q&A datasets. To help it generate the correct answers in the expected format, we developed a k-shot prompt template ($k = 5$) (Figure 1, top


Construct Datasets across Different Disciplines (Questions & Reference Answers)
Construct GPT-4 Generate-Answer Prompt
Use GPT-4 to Generate Answers (Author List)
Construct GPT-4 Self-Evaluation Prompt
Use GPT-4 to Classify Generated Answers into Five Cases (0-4)
Hallucination Defined as: Case 0, 3, 4
Hallucination Defined as: Case 0, 4
Calculation of Hallucination Rate
Bucket Analysis
Cross-Discipline Analysis

```
You are an academic assistant. Your task is to answer questions about the
authors of academic papers.
ONLY return a comma-separated list of authors' names in the following
format:
- First and middle names should be abbreviated as capital letters (no space),
followed by a space, then the full last name.
- For example, 'Lucy Winter Smith' should be written as 'LW Smith'.
- Only return the first 6 author names. If there are more, add '...' at the end.
Do not add any explanation, notes, or extra content. Only generate the
answer of the question after ‘Answer: ’.
... (few examples were given)
Now, answer the following question:
Question: Who are the authors of the paper titled '{paper title}' which was
published in {year}?
Answer:
```

```
You are an assistant evaluating your own previous answer to a question
about academic paper authors.
Question: Who are the authors of the paper titled '{paper title}' which was
published in {year}?
Reference answer: {reference_answer}
Your generated answer: {generated_answer}
Classify the generated answer into one of the following categories:
0. The generated answer does not contain any valid author names.
1. The generated answer contains the same author names as the reference
answer (order does not matter).
2. The generated answer contains fewer authors, but all names are present
in the reference answer (order does not matter).
3. The generated answer has partial overlap with the reference answer —
some names match, some do not.
4. All other cases.
Additional notes:
- Differences due to name formatting (e.g., 'I. V. Serban' vs 'I Serban')
should be considered matching.
- Order of names does not matter.
Respond ONLY with the number: 0, 1, 2, 3, or 4.
```

```
Case 0: No valid author names
Case 1: Exact match
Case 2: Subset
Case 3: Partial overlap
Case 4: All other cases
```

Figure 1: An overview of the pipeline.

right). Once the prompt template was established, we applied it to all questions and collected GPT-4's responses for analysis.

3.4 Self-Evaluation Using GPT-4

We used a self-evaluation method where GPT-4 assessed whether its answer matched the ground truth, using a unified prompt (Figure 1, bottom right). To enable detailed analysis, the prompt classified responses into five cases: no valid author names (0), exact match (1), subset (2), partial overlap (3), and else (4). GPT-4 returns a category label for each case, allowing flexible definition of hallucinations and distinction between response types.

We defined hallucinations under two settings: (1) cases 0, 3, and 4 (including partial overlaps with extra names); and (2) cases 0 and 4 (no names matching the ground truth). In each setting, answers were labeled as either hallucinations or non-hallucinations. The hallucination rate is calculated as the proportion of hallucinated entries in the dataset, denoted as $\mathbf{HR}_1$ and $\mathbf{HR}_2$ respectively for the two definitions.

3.5 Cross-disciplinary Analysis

We calculated the hallucination rates for each of the four disciplines and examined their relationship with the corresponding citation count rankings. Through this cross-disciplinary analysis, we aimed to observe whether disciplines with higher total citation counts tended to exhibit lower hallucination rates.

3.6 Bucket Analysis

To investigate within-discipline trends, we sorted the data in each discipline by citation count and divided it into equally sized buckets. We then calculated the average hallucination rate for each bucket, allowing us to examine the relationship between citation ranges and hallucination rates. Within each discipline, we expected buckets with higher citation counts to exhibit lower hallucination rates.

4 Experimental Results

We leveraged the OpenAI API and used GPT-4-turbo as our GPT-4 variant to implement the experimental pipeline introduced in Figure 1.

4.1 Cross-disciplinary Analysis Results

Table 2 shows cross-disciplinary results. Using HR_1, hallucination rates were high but followed our expected trend: Computer Science (65.65%) had the lowest rate, followed by Physics (81.83%), Chemistry (84.17%), and Biology (92.67%). This supports a negative correlation between citation count and hallucination rate. However, classifying Case 3 (partial overlap) as a hallucination may be too broad, since unmatched authors could still be correct due to Google Scholar's eleven author limit.

Discipline	Citation Rank	$\mathbf{HR}_1$	$\mathbf{HR}_2$
CS	1	65.65%	33.48%
Physics	2	81.83%	16.04%
Chemistry	3	84.17%	56.58%
Biology	4	92.67%	61.47%

Table 2: Overall hallucination rates across disciplines with two different definitions of hallucination.

To address this, we adopted a narrower hallucination definition (HR_2), excluding Case 3, rates dropped across disciplines but still generally decreased with citation count. An exception is Physics (16.04%), lower than Computer Science (33.48%), likely due to Physics papers having more authors (avg. 6.97 vs. 4.62/4.64/4.64 for CS/CH/BI), increasing chances of partial matches and lowering the hallucination rate.

Citation Deciles	**HR$_1$**				**HR$_2$**			
	CS	PH	CH	BI	CS	PH	CH	BI
Decile 1	84.78%	80.65%	80.23%	95.96%	68.48%	3.23%	67.44%	89.90%
Decile 2	85.87%	80.33%	81.61%	91.58%	67.39%	19.67%	64.37%	75.79%
Decile 3	80.22%	77.78%	80.00%	97.96%	35.16%	19.05%	60.00%	81.63%
Decile 4	76.60%	88.33%	90.70%	97.89%	32.98%	20.00%	60.47%	64.21%
Decile 5	67.78%	90.00%	91.86%	92.78%	30.00%	20.00%	65.12%	53.61%
Decile 6	64.84%	74.19%	87.21%	95.88%	31.87%	12.90%	54.65%	54.64%
Decile 7	59.78%	83.33%	81.18%	98.96%	21.74%	13.33%	51.76%	50.00%
Decile 8	54.95%	81.97%	87.21%	94.85%	23.08%	24.59%	46.51%	57.73%
Decile 9	45.65%	35.87%	87.21%	89.69%	13.04%	8.20%	53.49%	45.36%
Decile 10	35.87%	80.33%	74.42%	71.13%	10.87%	19.67%	41.86%	41.24%

Table 3: Hallucination rates vs citation deciles (low to high). HR$_1$ includes Cases 0, 3, 4 while HR$_2$ Cases 0, 4 only.

4.2 Bucket Analysis Results

We analyzed hallucination rates versus citation counts within each discipline, dividing papers into 10 citation-based deciles. Results are shown in Table 3.

Under HR$_1$, hallucination rates remain high across citation levels in most disciplines. A clear downward trend is observed only in Computer Science. In contrast, Chemistry, Biology, and Physics exhibit relatively flat or noisy patterns, with rates largely exceeding 80%. This is primarily due to partial overlaps (Case 3), where the model includes correct authors not present in the ground truth. As these cases are still penalized under HR$_1$, the hallucination rates remain elevated despite potentially valid predictions.

Under the narrower HR$_2$ setting (Cases 0 and 4 only), the negative correlation between citation count and hallucination rate becomes clearer across disciplines. CS shows the strongest trend, with hallucination rates dropping from 68.48% to 10.87%. Similar declines appear in BI (89.90% to 41.24%) and CH (67.44% to 41.86%), supporting the idea that highly cited (more well-known) papers are less prone to hallucination. Physics remains the exception. Firstly, its hallucination rate is very low, ranging from about 3.23% to 24.59%, likely due to the reason explained in Section 4.1. Furthermore, the hallucination rate is relatively flat or even noisy in both HR$_1$/HR$_2$. We hypothesized that the hallucination rate does not decrease for highly cited Physics papers for two reasons: 1) the average number of authors for each bucket is 7.49, 6.97, 6.87, 6.92, 6.33, 6.92, 7.17, 7.31, 7.21, 6.66, respectively, with lowest/highest Decile having the highest/second-lowest average number of authors, thus reducing/increasing the hallucination rates of these buckets artificially; 2) in our raw data, many highly cited Physics papers list institutions as authors without specifying individual names. Since GPT-4 is trained on such data, it tends to generate institution names for highly cited papers, while our ground truth includes only individual authors. To verify this, we prompted GPT-4 to assess whether hallucinated entries likely represented institution-authored papers and found Physics had the highest proportion of such cases.

In summary, applying a narrower definition of hallucination reveals clearer trends, particularly in Computer Science, Biology, and Chemistry, where higher citation counts generally correspond to a lower hallucination risk. This supports our hypothesis that factual prevalence mitigates hallucinations. Physics remains an outlier, likely due to its unique publication norms, as discussed previously.

5 Conclusion

This study explores the link between factual prevalence and hallucinations in LLMs by examining author predictions for academic papers. We created Q&A datasets from four disciplines and prompted GPT-4 to identify paper authors.

Using two definitions of hallucination, we find that the narrower criteria reveal a clear negative correlation between hallucination rates and factual prevalence, both across and within disciplines—except in Physics, where distinctive publication norms likely account for the anomaly. Our findings also underscore the challenges of defining hallucinations given incomplete ground truth and demonstrate how evaluation criteria impact results, emphasizing the importance of rigorous metric design in future research.

Acknowledgements

We thank ChatGPT and Grammarly for the assistance in typo correction and sentence shortening.

Limitations

Our study provides concrete evidence of an inverse correlation between factual prevalence and hallucinations in large language models (LLMs). However, the quality of the data source plays a critical role in determining the reliability of our experimental results. Google Scholar's author information is automatically scraped from original journal websites, with author names reformatted to include only initials and full last names. Additionally, Google Scholar typically lists only up to eleven authors per publication. This limitation presents a significant challenge for evaluating Case 3 (partial overlap), as it becomes difficult to determine whether the model-generated authors are actually correct but excluded from the truncated ground truth list. In future work, we plan to use higher-quality metadata sources like Semantic Scholar or CrossRef.

In some cases, the author list from Google Scholar contains only the name of an institution or collaboration, forcing us to drop such entries entirely. This happens especially frequently in Physics. Since the LLM is trained with such data, we believe for disciplines such as Physics, it is prone to predict institutions as the authors even when the ground truth contains only individual author names. This biases the hallucination rates for such disciplines.

Furthermore, our current setup focuses on evaluating hallucinations in the LLM-predicted paper authors. In future work, we plan to expand the evaluation by prompting the model to generate or summarize the paper's abstract or content. This extension would introduce additional indicators of hallucination and contribute to a more comprehensive framework for hallucination detection.

Finally, our study only uses paper citation counts as a proxy for factual prevalence. However, many papers may not be directly accessible to LLMs during training. Future studies will consider incorporating richer factual proxies such as textbook mentions, Wikipedia references, download counts, and Google search hits to better understand the correlation between factual prevalence and hallucinations. We also aim to evaluate hallucinations in LLMs beyond GPT-4, including multilingual and domain-specific models, for broader generalization and validation.

References

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. 2024. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630:625–630.

Sholto Kadavath, Thomas Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, and Jared Kaplan. 2022. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*.

Ningke Li, Yuekang Li, Yi Liu, Ling Shi, Kailong Wang, and Haoyu Wang. 2024. Drowzee: Metamorphic testing for fact-conflicting hallucination detection in large language models. *Proceedings of the ACM on Programming Languages*, 8:1843 – 1872.

Yinqiu Liu, Guangyuan Liu, Ruichen Zhang, Dusit Tao Niyato, Zehui Xiong, Dong In Kim, Kaibin Huang, and Hongyang Du. 2024. Hallucination-aware optimization for large language model-empowered communications. *ArXiv*, abs/2412.06007.

Potsawee Manakul, Alexander Liusie, and Mark J. F. Gales. 2023. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9004–9017. Association for Computational Linguistics.

Jason Wei, Natalie Karina, Hyung Won Chung, Yujia Jin Jiao, Sam Papay, Anthony Glaese, and William Fedus. 2024. Measuring short-form factuality in large language models. *arXiv preprint arXiv:2411.04368*.

Yakir Yehuda, Itzik Malkiel, Oren Barkan, Jonathan Weill, Royi Ronen, and Noam Koenigstein. 2024. Interrogatellm: Zero-resource hallucination detection in llm-generated answers. In *Annual Meeting of the Association for Computational Linguistics*.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, Longyue Wang, Anh Tuan Luu, Wei Bi, Freda Shi, and Shuming Shi. 2023. Siren's song in the ai ocean: A survey on hallucination in large language models. *ArXiv*, abs/2309.01219.