

HALLUCINATIONS IN BIBLIOGRAPHIC RECOMMENDATION: CITATION FREQUENCY AS A PROXY FOR TRAINING DATA REDUNDANCY

A PREPRINT

 **Junichiro Niimi***^{1,2}
¹Meijo University
²RIKEN AIP

ABSTRACT

Large language models (LLMs) have been increasingly applied to a wide range of tasks, from natural language understanding to code generation. While they have also been used to assist in bibliographic recommendation, the hallucination of non-existent papers remains a major issue. Building on prior studies, this study hypothesizes that an LLM’s ability to correctly produce bibliographic information depends on whether the underlying knowledge is generated or memorized, with highly cited papers (i.e., more frequently appear in the training corpus) showing lower hallucination rates. We therefore assume citation count as a proxy for training data redundancy (i.e., the frequency with which a given bibliographic record is repeatedly represented in the pretraining corpus) and investigate how citation frequency affects hallucinated references in LLM outputs. Using GPT-4.1, we generated and manually verified 100 bibliographic records across twenty computer-science domains, and measured factual consistency via cosine similarity between generated and authentic metadata. The results revealed that (i) hallucination rates vary across research domains, (ii) citation count is strongly correlated with factual accuracy, and (iii) bibliographic information becomes almost verbatimly memorized beyond approximately 1,000 citations. These findings suggest that highly cited papers are nearly verbatimly retained in the model, indicating a threshold where generalization shifts into memorization.

Keywords Large language model · Natural language processing · Hallucination · Information retrieval · Recommendation system

1 Introduction

Large language models (LLMs) have achieved remarkable fluency across a wide range of domains [1]. However, they are also known to generate hallucinations that are nonsensical or unfaithful to the provided source content [2, 3]. In particular, the generation of non-existent academic references or legal precedents has been widely recognized as a critical issue [4]. For example, in the field of marketing, where Recency–Frequency–Monetary (RFM) analysis [5, 6, 7] is commonly employed as a customer relationship management (CRM) [8], when prompted to “Please suggest recent academic papers on RFM analysis with Author (Year) Title, Journal, Vol, No, pp style,” the model (GPT-4.1) produced the following response:

Chitturi, P., Raghunathan, B., Scian-
dra, R., & Sikora, J. (2010). “RFM
and CLV: Using Customer Data for Im-
proved Decision Making.” Journal of

Direct, Data, and Digital Marketing
Practice, 12(1), 1–10.

Although the output follows the correct bibliographic format, the paper itself does not exist. Each component imitates genuine studies, such as the author names (e.g., Chitturi and Raghunathan [9]), journal name (Journal of Direct, Data, and Digital Marketing Practice [10]), and the paper title (e.g., “RFM and CLV” [11]), but the numerical details are fictitious, suggesting that multiple authentic entries were probabilistically merged into a coherent yet fabricated citation.

These fabricated yet plausible references suggest that hallucinations in bibliographic recommendation may not occur arbitrarily, but rather reflect how knowledge is represented within the model. For example, the probability of reproducing training data has been shown to correlate with its frequency of appearance [12]. This study therefore focuses on bibliographic recommendation using LLMs and empirically examines how factual correctness

varies with domain popularity (number of papers in the field) and citation prominence (citation count of the generated reference). Our findings suggest that hallucinations arise not randomly but systematically from imbalanced knowledge distributions within the representation space.

2 Related Study

Hallucination in LLMs has been examined from diverse perspectives [2, 4, 13, 14, 15]. OpenAI’s analysis [15] argued that reinforcement learning with human feedback (RLHF) [16, 17] may inherently encourage hallucination, as current LLMs are penalized for responding “I don’t know” (IDK) and instead rewarded for producing statistically plausible continuations. This alignment objective can thus promote confident but unreliable statements.

Conversely, security-oriented studies have highlighted the opposite tendency: information repeated multiple times during pretraining is more likely to be memorized and reproduced verbatim [18, 19, 20, 12]. This view aligns with recent theoretical accounts positioning LLMs as probabilistic pattern recognizers that approximate data distributions rather than explicitly “understanding” knowledge [21, 22]. From this perspective, hallucination and *exposure* [23] (i.e., training data leakage) represent opposite outcomes of the same probabilistic learning dynamics, where the frequency of exposure governs whether information is faithfully recalled or spuriously synthesized.

In the context of citation recommendation [24, 25, 26], this implies that frequently cited papers which appear across numerous publications and other web sources are more likely to be verbatimly recalled by LLMs, whereas sparsely represented works tend to be fabricated. This study hypothesizes that hallucination in bibliographic recommendation is systematically related to the *training data redundancy* (i.e., the frequency with which a given bibliographic record is repeatedly represented in the pretraining corpus). Highly cited papers are expected to be more robustly represented, leading to lower hallucination rates, while limited-redundancy papers are more prone to plausible but non-existent references.

3 Experiments

3.1 Methodology

Bibliographic records were generated using GPT-4.1 accessed via API (knowledge cutoff: June 2024). To minimize domain-specific citation bias, twenty topics were selected within the field of computer science (e.g., transformer [27], diffusion model [28], retrieval-augmented generation [29]). For each topic, the model was prompted to recommend five academic papers in JSON format (Fig. 3; see Appendix A), yielding a total of 100 samples.

Each output record was manually validated using Google Scholar. Existence of the paper was confirmed primarily by its title; minor coincidences in author or journal names

were not considered sufficient. Each record was scored as completely correct ($score = 2$), partially hallucinated ($score = 1$; when some metadata such as author names, journal, or year were inaccurate), or completely hallucinated ($score = 0$). Citation counts were retrieved from Google Scholar (as of October 2025). Semantic similarity between generated and authentic bibliographic metadata was computed using Sentence-BERT [30] (all-MiniLM-L6-v2) embeddings with cosine similarity, and the relationship between citation frequency and similarity was analyzed.

3.2 Results

Experiment 1. We first compared low- and high-citation groups divided at the median citation count ($M_{dn} = 818$). A one-tailed t -test revealed that the high-citation group achieved significantly higher factual scores than the low-citation group ($t(98) = -5.12, p < .001$; $M_{high} = 1.245, M_{low} = 0.725$).

Experiment 2. Figure 1 shows the average factual scores across research domains. The scores differ markedly among domains: Domains related to image processing, such as Vision Transformer (ViT) [31] and Diffusion Model [28], achieved notably higher accuracy, whereas recent LLM-oriented techniques, such as RAG [29] and LoRA [32], exhibited substantially lower scores. This discrepancy likely reflects the popularity of the research domain, which is related with the data redundancy. Consequently, hallucinations occurred more frequently in underrepresented domains such as LoRA and Graph Transformer [33].

Experiment 3. Using only valid records ($score > 0$; $n = 81$), we analyzed the relationship between log-transformed citation counts and cosine similarity (Fig. 2). A strong positive correlation ($r = 0.75, p < .001$) indicates a clear log-linear relationship, which reflects the memorization scaling law [12]. In this context, we may interpret our finding as a form of a redundancy scaling law, where the probability of factual recall significantly increases as $\log(citation)$ rises.

Experiment 4. Notably, Fig. 2 also indicates the saturation near 1.0 when $\log(citation)$ exceeded 7 (approximately 1,000 citations), suggesting that highly cited papers are almost verbatimly retained within the model. In that regime, frequently cited papers are not merely represented through probabilistic token associations but are instead recalled almost verbatim. To quantify this non-linear pattern, we conducted logistic regression using min-max-normalized cosine similarity as the dependent variable. The model revealed a significant positive relationship between $\log(citation)$ and normalized similarity ($\beta_1 = 0.523, p = 0.003$) with a low intercept ($\beta_0 = -2.360, p = 0.020$). The estimated inflection point ($-\beta_0/\beta_1 \approx 5$) corresponds to roughly 100 citations,

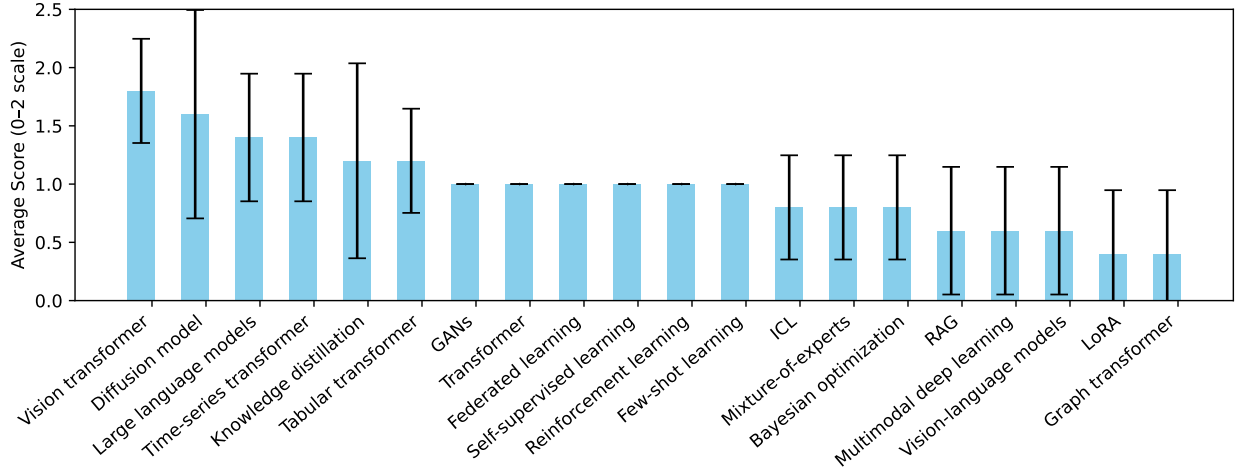


Figure 1: Average factual score by domain. Error bars indicate the 95% confidence interval. Accuracy varies by domain familiarity.

marking the transition from generalization to memorization.

These four experiments collectively support our initial hypothesis that citation count acts as a proxy for the training data redundancy. The positive and non-linear relationship between $\log(\text{citation})$ and cosine similarity indicates that hallucination is not random but structurally linked to uneven knowledge distributions within the model’s representation space.

4 Conclusion

This study empirically examined how citation frequency functions as a proxy for hallucination in bibliographic recommendation by LLMs. The model was instructed to output JSON-formatted results without explanations, effectively disabling IDK responses. In line with previous study [15], such output constraints encourage the model to produce plausible yet non-existent entries.

The key findings are as follows: (i) hallucination rates vary across research domains, (ii) citation count is strongly correlated with factual accuracy, and (iii) bibliographic information becomes almost verbatimly memorized beyond approximately 1,000 citations. In other words, while LLMs can faithfully reproduce information about highly cited papers, they struggle in domains with shorter publication histories or limited redundancy in the training corpus. Contrary to the discoverability phenomenon [12] where memorization emerges only when sufficient context is given, our results suggest the opposite direction: highly redundant knowledge can be recalled even with minimal prompting. Although this study focused on GPT-4.1 within the computer science domain, future research should extend the analysis to other models, disciplines, and multilingual contexts to assess the generality of this threshold behavior.

References

- [1] Jan Ole Krugmann and Jochen Hartmann. Sentiment analysis in the age of generative ai. *Customer Needs and Solutions*, 11(1):3, 2024.
- [2] Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630, 2024.
- [3] Ziwei Xu, Sanjay Jain, and Mohan Kankanhalli. Hallucination is inevitable: An innate limitation of large language models. *arXiv preprint arXiv:2401.11817*, 2024.
- [4] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55, 2025.
- [5] Connie L Bauer. A direct mail customer purchase model. *Journal of Direct Marketing*, 2(3):16–24, 1988.
- [6] Jan Roelf Bult and Tom Wansbeek. Optimal selection for direct mail. *Marketing Science*, 14(4):378–394, 1995.
- [7] Sunil Gupta and Donald R Lehmann. Customer life-time value and firm valuation. *Journal of Relationship Marketing*, 5(2-3):87–110, 2006.
- [8] Jacob Jacoby and Robert W Chestnut. *Brand loyalty: Measurement and management*. John Wiley & Sons Incorporated, 1978.
- [9] Ravindra Chitturi, Rajagopal Raghunathan, and Vijay Mahajan. Form versus function: How the intensities of specific emotions evoked in functional

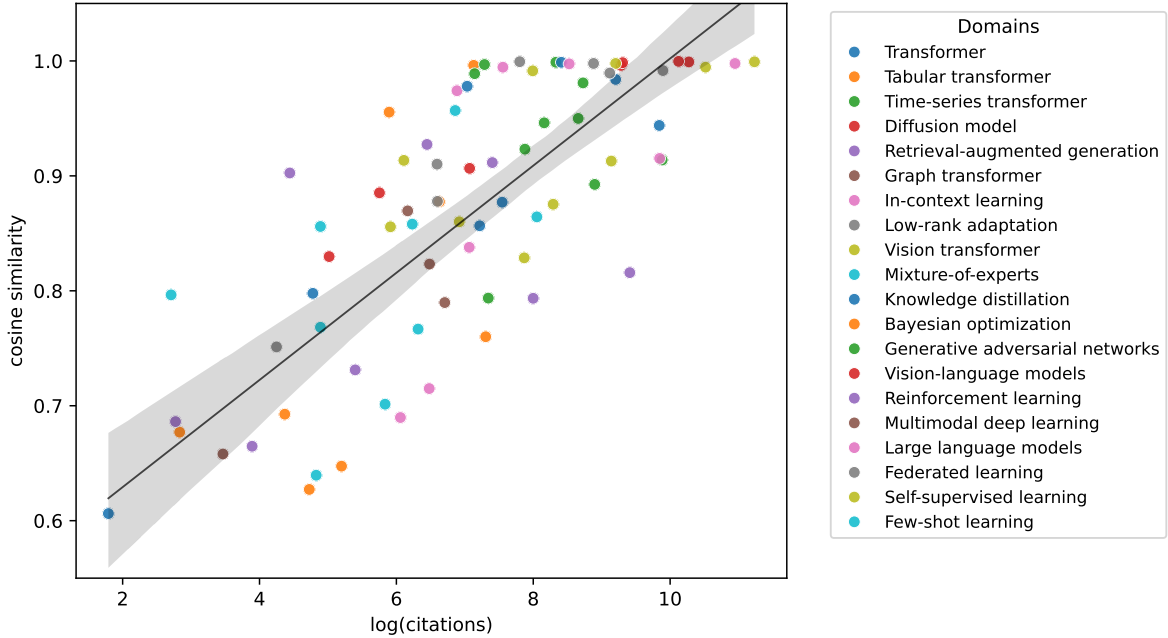


Figure 2: Relationship between citation frequency and generation fidelity. Each dot represents a factual citation generated by the model, colored by research domain. The dashed line indicates the fitted linear regression (95% CI in gray). The correlation ($r = 0.75$, $p < .001$) demonstrates a strong log-linear relationship between citations and factual accuracy, with saturation near $\log(\text{citation}) \approx 7$.

- versus hedonic trade-offs mediate product preferences. *Journal of marketing research*, 44(4):702–714, 2007.
- [10] Efthymios Constantinides and Stefan J Fountain. Web 2.0: Conceptual foundations and marketing issues. *Journal of direct, data and digital marketing practice*, 9(3):231–244, 2008.
- [11] Peter S Fader, Bruce GS Hardie, and Ka Lok Lee. Rfm and clv: Using iso-value curves for customer base analysis. *Journal of marketing research*, 42(4):415–430, 2005.
- [12] Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramèr, and Chiyuan Zhang. Quantifying memorization across neural language models. In *The Eleventh International Conference on Learning Representations*, 2022.
- [13] Hoang Anh Dang, Vu Tran, and Le-Minh Nguyen. Survey and analysis of hallucinations in large language models: attribution to prompting strategies or model behavior. *Frontiers in Artificial Intelligence*, 8:1622292, 2025.
- [14] Joseph Spracklen, Raveen Wijewickrama, AHM Nazmus Sakib, Anindya Maiti, and Bimal Viswanath. We have a package for you! a comprehensive analysis of package hallucinations by code generating {LLMs}. In *34th USENIX Security Symposium (USENIX Security 25)*, pages 3687–3706, 2025.
- [15] Adam Tauman Kalai, Ofir Nachum, Santosh S Vempala, and Edwin Zhang. Why language models hallucinate. *arXiv preprint arXiv:2509.04664*, 2025.
- [16] Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017.
- [17] Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*, 2019.
- [18] Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. Deduplicating training data makes language models better. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8424–8445, 2022.
- [19] Nikhil Kandpal, Eric Wallace, and Colin Raffel. Deduplicating training data mitigates privacy risks in language models. In *International Conference on Machine Learning*, pages 10697–10707. PMLR, 2022.
- [20] Nicholas Carlini, Florian Tramèr, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar

- Erlingsson, et al. Extracting training data from large language models. In *30th USENIX security symposium (USENIX Security 21)*, pages 2633–2650, 2021.
- [21] Suvir Mirchandani, Fei Xia, Pete Florence, Brian Ichter, Danny Driess, Montserrat Gonzalez Arenas, Kanishka Rao, Dorsa Sadigh, and Andy Zeng. Large language models as general pattern machines. In *Conference on Robot Learning*, pages 2498–2518. PMLR, 2023.
- [22] Pablo Contreras Kallens and Morten H Christiansen. Distributional semantics: Meaning through culture and interaction. *Topics in cognitive science*, 2024.
- [23] Nicholas Carlini, Chang Liu, Úlfar Erlingsson, Jernej Kos, and Dawn Song. The secret sharer: Evaluating and testing unintended memorization in neural networks. In *28th USENIX security symposium (USENIX security 19)*, pages 267–284, 2019.
- [24] Chanwoo Jeong, Sion Jang, Eunjeong Park, and Sungchul Choi. A context-aware citation recommendation model with bert and graph convolutional networks. *Scientometrics*, 124(3):1907–1922, 2020.
- [25] Zitong Zhang, Braja Gopal Patra, Ashraf Yaseen, Jie Zhu, Rachit Sabharwal, Kirk Roberts, Tru Cao, and Hulin Wu. Scholarly recommendation systems: a literature survey. *Knowledge and Information Systems*, 65(11):4433–4478, 2023.
- [26] Jie Zhu, Braja G Patra, and Ashraf Yaseen. Recommender system of scholarly papers using public datasets. *AMIA summits on translational science proceedings*, 2021:672, 2021.
- [27] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.
- [28] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.
- [29] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474, 2020.
- [30] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, page 3982. Association for Computational Linguistics, 2019.
- [31] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words. *International Conference on Learning Representations (ICLR 2021)*, 2021.
- [32] Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.
- [33] Chengxuan Ying, Tianle Cai, Shengjie Luo, Shuxin Zheng, Guolin Ke, Di He, Yanming Shen, and Tie-Yan Liu. Do transformers really perform badly for graph representation? *Advances in neural information processing systems*, 34:28877–28888, 2021.
- [34] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021.
- [35] Xuezhe Ma, Chunting Zhou, Xiang Kong, Junxian He, Liangke Gui, Graham Neubig, Jonathan May, and Luke Zettlemoyer. Mega: Moving average equipped gated attention. *arXiv preprint arXiv:2209.10655*, 2022.

A Model Settings

Actual prompt to generate bibliographic information is shown in Fig. 3. To strictly generate structured records, we include JSON schema and "no additional explanation" descriptions.

B Generated Records

Some records generated with GPT-4.1 are shown in Table 1. For the partially hallucinated records, the tendency for the first author, year and paper title to be strictly reproduced, while other information is reproduced in a plausible manner.

```

### Instruction:
You are an academic assistant that outputs structured bibliographic data in JSON format.
Please suggest 5 recent academic papers related to "{domain}".

Each paper should be represented as a JSON object following this schema:
{
  "author": "Author name(s) in APA style, e.g., 'Smith, J. & Tanaka, K.'",
  "year": 2023,
  "title": "Title of the paper",
  "journal": "Name of the academic journal",
  "volume": "12",
  "number": "3",
  "pages": "123--145"
}

Output must be a single valid JSON array of objects and contain no additional explanation.
If you are unsure about any field, please leave it as an empty string ("").

```

Figure 3: Prompt to generate bibliographic information

Table 1: Examples of generated bibliographic records. Hallucinated parts are shown in bold.

Class	Domain	Cos	Results
Completely correct	Vision transformer	0.995	Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., & Guo, B. (2021). Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. Proceedings of the IEEE/CVF International Conference on Computer Vision, 10012–10022.[34]
Partially hallucinated	Transformer	0.798	Ma, X., Xu, J., Sun, Y., He, Y., & Lin, J. (2022) Mega: Moving average equipped gated attention. International Conference on Machine Learning, 162, 15369–15384. [35]
Completely hallucinated	Tabular transformer	n.a.	Kossen, J., Probst, P., Schirrmeister, R. T. & Bischl, B. (2023) Self-Attention for Raw Numerical Tabular Data. IEEE Transactions on Neural Networks and Learning Systems.