Do Language Models Know When They're Hallucinating References?

Anonymous Author(s)
Affiliation
Address
email

Abstract

State-of-the-art language models (LMs) are famous for "hallucinating" references. These fabricated article and book titles lead to harms, obstacles to their use, and public backlash. While other types of LM hallucinations are also important, we propose hallucinated references as the "drosophila" of research on hallucination in large language models (LLMs), as they are particularly easy to study. We show that simple search engine queries reliably identify such hallucinations, which facilitates evaluation. To begin to dissect the nature of hallucinated LM references, we attempt to classify them using black-box queries to the same LM, without consulting any external resources. Consistency checks done with *direct* queries about whether the generated reference title is real (inspired by Kadavath et al. (2022); Lin et al. (2022); Manakul et al. (2023)) are compared to consistency checks with *indirect* queries which ask for ancillary details such as the authors of the work. These consistency checks are found to be partially reliable indicators of whether or not the reference is a hallucination. In particular, we find that LMs often hallucinate *differing* authors of hallucinated references when queried in independent sessions, while *consistently* identify authors of real references. This suggests that the hallucination may be more a generation issue than inherent to current training techniques or representation.[1]

1 Introduction

Language models (LMs) famously hallucinate[2], meaning that they fabricate strings of plausible but unfounded text. As LMs become more accurate, their fabrications become more believable and therefore more problematic. A primary example is "hallucinated references" to non-existent articles with titles readily fabricated by the LM. For instance, a real *New York Times* article entitled "When A.I. Chatbots Hallucinate" leads with a ChatGPT-fabricated *New York Times* article titled "Machines Will Be Capable of Learning, Solving Problems, Scientists Predict" (Weise and Metz, 2023).

In this work, we study the problem of hallucinated computer science references. We suggest the AI community study this type of hallucination as it presents a tractable model problem—much like the fruit fly, Drosophila melanogaster, has within biology. Hallucinated references exhibit key properties that make their study feasible. First, they can be automatically classified more easily than other types of hallucination. We provide a pipeline for classifying hallucinations using a search engine API and show that, on a sample of 100 potential articles, it agreed with experienced human annotators on at least 98% of the references. References generally have consistent titles and are widely advertised so as to be likely to be present in any training set that aims to be comprehensive. Other types

[1] All our code and results are available at LINK.

[2] Though it is an anthropomorphism, we use the term *hallucinate* due to its widespread adoption, following the use-theory of meaning (Wittgenstein, 1953). We use the terms *hallucinate* and *fabricate* interchangeably throughout the paper.

Submitted to 37th Conference on Neural Information Processing Systems (NeurIPS 2023). Do not distribute.

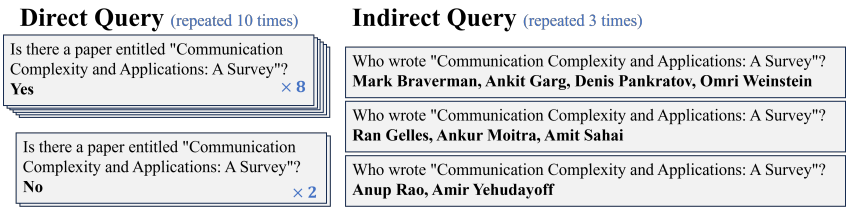


Figure 1: Direct vs. indirect queries for predicting whether a given paper title is hallucinated. LM generations are **boldface**. Prompts in this figure have been shortened for illustrative purposes.

types of hallucinations, such as factoids, are more complex to classify due to ambiguities in their wordings and difficulty of assessing their presence in the training data. Second, many researchers studying hallucination possess expertise that bears directly on the understanding of hallucinations this domain. Studying this modest manifestation of hallucination provides a blueprint for detecting and mitigating more complex types. Focusing on this tractable niche lays the groundwork for countering hallucinations in high-impact AI applications. Just as the genetics of fruit flies have yielded biological insights, targeted inquiry into reference hallucination can yield insights into LMs.

We provide an initial investigation into the questions *why do LMs hallucinate references, and what can be done about it?* Is it a problem of LM *representation*, a problem of *training* (maximizing next-word likelihood), or a problem due to the way they are used for *generation*? Specifically, we investigate whether an LM itself can be used to detect whether or not an output it has produced is a hallucination, without any external resources. While this does not provide a complete answer to the questions of why and what to do, it does inform the discussion. In particular, to the extent that LMs can be used to detect their own hallucinations, this suggests that the hallucination problem is not inherently one of training or representation but is rather one of generation because the models contain enough information to at least reduce the hallucination rate.

In this work, hallucinations refer to open-domain fabricated text with little or no grounding in the training data, as opposed to closed-domain hallucinations (see, e.g., Ji et al., 2023). Groundedness, the opposite of fabrication, is based on the training corpus, while correctness is evaluated with respect to absolute truth as discussed by Evans et al. (2021). For instance, the statement *the earth is flat* is incorrect but appears on many web pages and is likely to be grounded in the training data. In the case of references, however, groundedness and correctness are often closely related. To evaluate groundedness, we use exact-match Web search as a heuristic, as it is a reasonable approximation for identifying texts grounded in the training data, considering that article titles are specific and designed for discoverability on the web.

Direct queries (DQs). Our work builds upon and is inspired by two recent works that show how to use black-box generative LMs to assess confidence in generations, without consulting external references or inspecting weights. In particular, Kadavath et al. (2022) introduce multiple direct black-box strategies for using an LM to extract confidence estimates by querying the LMs on question-answer problems. Manakul et al. (2023) apply a similar direct self-consistency check to identify relative hallucinations in a summarization context.

Indirect queries (IQs). We suggest a new approach that uses *indirect queries*. A DQ may ask, *Is the following paper real?* while an IQ may ask, *Who are the authors of this paper?*, as illustrated in Figure 1. Answers are then generated to the IQ in $i > 1$ independent sessions, and tested for consistency. The motivation for IQs comes from investigative interviews, where detectives are advised to interview individuals separately and ask open-ended questions (Vredeveldt et al., 2014).

Contributions. There are several contributions of this work. First, we propose the problem of hallucinated computer science references as a model instance worth studying, like Drosophila. Second, we demonstrate that they can be *reliably* and *automatically* classified. Third, we perform a systematic LM study of hallucinated references, enabling us to compare hallucination rates across LMs. Fourth, we introduce IQs for evaluating hallucinations. Finally, we compare these to DQs across GPT and Llama based LMs. A conclusion of our work for reducing hallucination is the recognition that changing the generation pipeline can certainly help, while it is less clear if training or representation changes are necessary.

List 5 existing references related to "Artificial intelligence: Planning and scheduling". Just output the titles. Output format should be <num.> <title>
1. Artificial Intelligence: A Modern Approach
2. Automated Planning: Theory and Practice
3. Principles of Artificial Intelligence: Planning
4. AI Planning, Scheduling, and Constraint Satisfaction: From theory to practice
5. Intelligent Scheduling Systems

Figure 2: The prompt used to generate $k = 5$ reference titles. This method generates both grounded and hallucinated references. Topics are chosen from the ACM Computing Classification System.

2 Methodology

We now give an overview of our methodology followed by further details on our DQs and IQs. Note that this full pipeline is run separately for each of our LMs, so there is no mixing across LMs. We first describe how we generate lists of candidate reference titles.

Generating references. The input to our evaluation is a set of topics from which we generate k references each using the LM by prompting it with temperature 1 as illustrated in Figure 2. The procedure is re-run if the LM fails to generate a list of k candidate titles. We then run our classification procedures, described below, on each of the candidate titles.

Hallucination estimation procedures. Each of our procedures takes three inputs: (1) A candidate reference title. Given that there is generally less ambiguity in the title of a reference than in the spelling or abbreviation of its authors names, for each reference we chose to use only its title as input. (2) A dialogue-based LM such as ChatGPT or Llama2chat. (3) A number of queries made to the LM per title. For DQs, $j \geq 1$ specifies how many judgments to make. For IQs, $i \geq 1$ specifies how many indirect responses to request.

In our experiments, the candidate title will have been generated using the LM, though this is not a requirement. The procedure detects (possibly) hallucinated references by querying the LM to check the existence of the reference. It does so by making black-box completion queries to the same LM. Finally, the procedure outputs a real-valued prediction in $[0, 1]$ of the probability the title is grounded (G) or a hallucination (H). We consider $j > 1$ to implement a version of the procedure that outputs probabilities rather than just G/H judgments. Since we do not have access to the probability distribution of the completions of some of the SOTA LMs such as GPT-4, the above procedure effectively simulates probabilities using sampling at temperature 1.

Labeling. For labeling, we use exact match in a search engine as a heuristic for labeling G/H. The reference title surrounded by quotes is searched in the web using Web search (e.g., "LMs are few-shot learners"). If no results are retrieved, we label the reference title as hallucinated and vice versa. Final receiver operating characteristic (ROC) curves and false discovery rates (FDR) are determined by comparing the ground truth labels to the classifications. Note that we also experimented with academic reference APIs such as Semantic Scholar. While these gave thorough details about each paper in its index, many grounded references (even for real papers) did not appear in their indexes, and we found search engine results to be significantly more complete.

To test the efficacy of Bing search as an automatic labelling heuristic, we performed a human annotation of 100 references generated by GPT-4. Four computer scientists (with IRB approval) who are experienced in searching for academic references in this domain, independently labeled each reference as grounded or a hallucination, without examining the labels of the Bing search procedure. On $98/100$ references, there was complete agreement among the annotators and the Bing search labeling strategy. We provide further details in the Section A.3. A manual inspection of 120 random examples is presented in Section A.7.

2.1 Direct query details

The direct query (DQ) procedures simply query whether or not the given title exists following the format shown in Figure 3. We created three query templates (DQ1, DQ2, and DQ3) based on the multiple direct query approaches advocated by Kadavath et al. (2022); Manakul et al. (2023). The first query asks whether the reference exists directly. However, as discussed in prior work, some LMs can be strongly biased in answering the question when phrased this way, e.g., it may be presumed

real without any context about where the reference came from. DQ2 and DQ3 establish the context indicating that the reference was generated by an assistant or LM. DQ3 goes further by giving additional comparisons, as advocated for in prior work. For DQ3, all k queries from our generation step (using the same LM) are shown.

For each query, we generate $j \geq 1$ completions to approximate the probability distribution of the model. These strings are converted to binary judgements as follows: We calculate how many completions contained the word *yes* and divide it by the total number of completions to get the estimates of groundedness. This means that empty or otherwise invalid answers were assigned *no*. We do not assume that this score is calibrated as our analysis considers arbitrary probability thresholds.

We sample j completions for each direct prompt. Temperature 1 is used when $j > 1$ and temperature 0 is used when $j = 1$ to approximate the most likely LM completion.

Direct Query 1 (DQ1)

U: Does the reference "Principles of Artificial Intelligence: Planning" exist? Output just yes/no.
A: **YES**

Direct Query 2 (DQ2)

U: Give a famous reference for reading.
A: Principles of Artificial Intelligence: Planning
U: Does the above reference exist? Output just yes/no.
A: **NO**

Direct Query 3 (DQ3)

U: A language model generated references related to a research topic with the following titles:
*A:*1. Artificial Intelligence: A Modern Approach
2. Automated Planning: Theory and Practice
3. Principles of Artificial Intelligence: Planning
4. AI Planning and Scheduling: A Survey
5. Intelligent Scheduling Systems
U: Does the reference with title #3 exist? Output just yes/no.
A: **YES**

Figure 3: Examples of the three direct prompts used for the DQs.

2.2 Indirect query details

The IQs proceed in two steps.

Step 1: Interrogation. Separately for each reference, an IQ is made of the LM $i > 1$ times at temperature 1, as shown in Figure 6 (top).

Step 2: Overlap estimation.. The LM is used to evaluate overlap between the i responses. For each pair of answers, an estimate is computed by calling the overlap query, as shown in Figure 6 (bottom). The leading number is extracted, or, if no number is given, then a 0 is used. (We divide by 100 and clip the answer to the interval $[0, 1]$ to convert the percentages to fractions.)

The rationale for this approach is that we expect consistent responses to indirect questions to indicate the existence of a grounded reference title, while inconsistent responses may be taken as an warning sign for hallucination. Our method does not rely on external resources and uses the same LM for hallucination detection end-to-end. Of course, parsing and string-matching could be used in place of a LM for the overlap step, though this would require name matching which is known to be a thorny problem and one which is well suited for pretrained LMs.

3 Results and Discussion

The code and data generated in our experiments will be made available upon publication.

3.1 Experiment details

Models. We utilize the OpenAI LMs, including GPT-3 (*text-davinci-003*), ChatGPT (*gpt-35-turbo*), and GPT-4 (*gpt-4*). Furthermore, we employ open-source models from the Llama 2 Chat Touvron et al. (2023) *llama-2-*-chat* series referred to as L2-7B, L2-13B, and L2-70B. To access the OpenAI LMs, we make use of the Azure OpenAI API.

Topics. We use the ACM Computing Classification System (CCS; (Rous, 2012) for topics. CCS contains 12 high level categories, 84 second level concepts, and 543 subconcepts at the third level of granularity. For generating the dataset, we sample 200 of the 543 subconcepts uniformly at random, describing each by a topic string of the form *concept: subconcept* (e.g., *Information retrieval: Retrieval models and ranking*). For each topic, we generate $k = 5$ references. In this manner, we generate $200 \times 5 = 1000$ candidate paper titles using each LM.

Parameters. We selected $i = 3$ IQ results and averaged the overlapping evaluations to compute the final score for each IQ experiment. For DQs experiments, we sampled $j = 10$ judgments at temperature 1.0 and reported the fraction of *yes* responses as a final score.

Search engine labels. The Bing search engine API is used for searching for the candidate title string on the Web. Note that even with exact string match, some flexibility beyond capitalization and punctuation is allowed.

3.2 Quantitative metrics

First, Table 1 shows the rates of hallucination for the six models studied. As expected, references produced by the newer models (which achieve higher scores on other benchmarks (Srivastava et al., 2022)) also exhibit a higher grounding rate or, equivalently, a lower hallucination rate.

Table 1: Hallucination rates out of 1,000 generated titles, measured by our automatic labeler.

LM	GPT-4	ChatGPT	GPT-3	L2-70B	L2-13B	L2-7B
Hallucination rate	46.8%	59.6%	73.6%	66.2%	76.7%	68.3%

Since each of our querying strategies outputs a real-valued score, one can trade-off accuracy on G (i.e., how often truly grounded references are labeled G) and H (how often truly hallucinated references are labeled H) by thresholding the score to form a G or H classification. The standard ROC curves based on these thresholded scores are shown for each approach and model in Figure 4. Due to the space limitation, we show the results for GPT-4, ChatGPT and L2-70B and refer Section A.4 for the results of additional models. These figures enable one to explore different points on this trade off for each classifier. For the L2-70B and ChatGPT models, the IQ procedure performs best as quantified via the area under the ROC curve (AUC). For GPT-4 (Figure 4c), both the IQ and DQ approaches work well for classifying hallucination and groundedness with the IQ (AUC: 0.878) and DQ1 (AUC: 0.887) performing the best. The performance of each procedure generally improves as the model size increases. For smaller models, where the procedures perform worst, others have found that users are less likely to believe the generated text due to its inaccuracy (OpenAI, 2023).

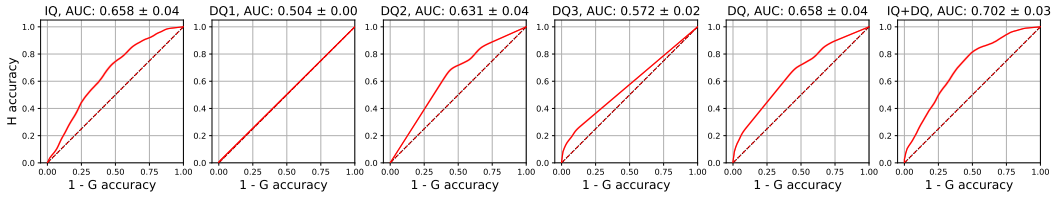
Each groundedness classifier can also be used as a filter to generate a list of likely grounded references for a literature review based on the raw generations of an LM. Aside from relevance, which we do not study in this work, two primary quantities of interest to a user of this filter would be the fraction of references preserved (more references provide a more comprehensive review) and the fraction of preserved references which are actually hallucinations. Figure 5 shows how these two quantities can be traded off. As one varies the threshold of G/H classification and returns only those references classified as grounded, the FDR captures the fraction of references produced which are hallucinations. Users may have a certain rate of tolerance for hallucinations, and one would like to maximize the number of generated references subject to that constraint. For L2-70B and ChatGPT, the IQ method achieves significantly lower FDR and a provides a substantially better FDR-preservation rate trade-off than the other approaches. For GPT-4, both IQ and DQ methods offer low FDR with comparable trade-offs.

Overall, our hypothesis that IQs would be more reliable than DQs appears to hold for ChatGPT and L2-70B; for GPT-4 the DQs were similarly effective.

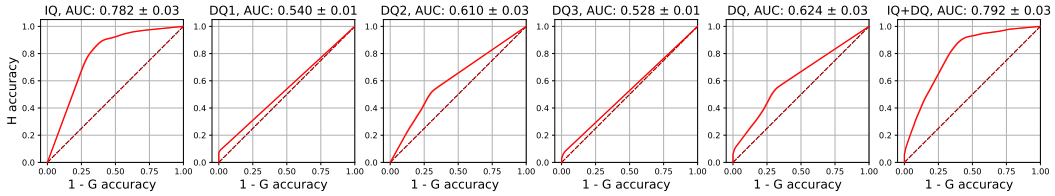
We find that classification performance increases when we take ensemble of different approaches, as illustrated by ROC curves in Figure 4. The ensemble is simply the mean of the scores and use them as thresholds. The ensemble of IQ and DQ (computed using the 50-50 mean of IQ and the DQ mean), referred to as *IQ+DQ* performs the best for every model. A qualitative analysis of the types of hallucinations and errors encountered are in Appendix A.1. A manual examination of 120 examples is given in Section A.7. The compute costs, which involve ≈6.6 million tokens and $412, are discussed in Section A.6.

4 Limitations and Conclusions

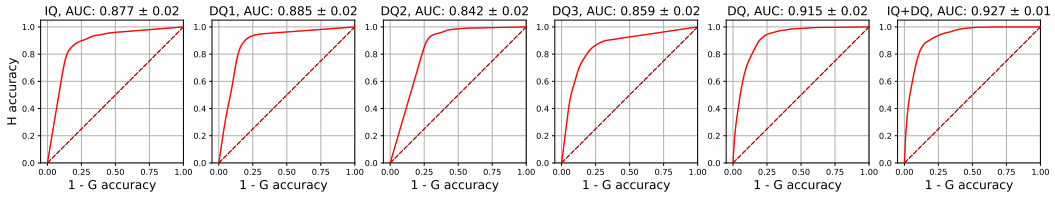
There are several limitations of this work: 1) We consider web as a contending proxy for the models' training data. However, we cannot conclude what is truly grounded versus hallucination since we don't have access to the training data. 2) The notion of hallucination is not entirely black and white



(a) Llama-2-70b-chat



(b) ChatGPT



(c) GPT-4

Figure 4: ROC Curves for IQ and DQ approaches (1-3, left to right) along with the ensemble of DQ, and IQ combined with DQ approaches (4-5, left to right). 95% confidence intervals for ROC curves and AUC are also shown.

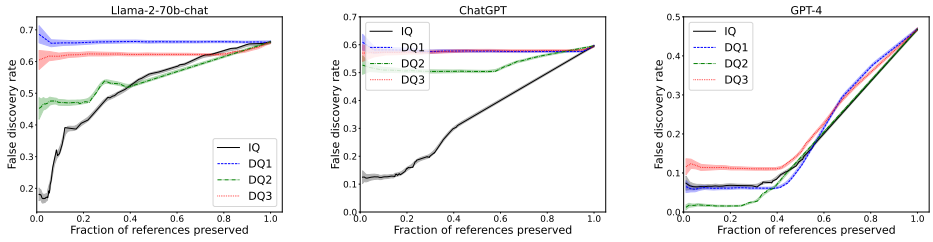


Figure 5: False discovery rate (FDR) vs. fraction of references preserved for each groundedness filter (IQ, DQ1, DQ2, DQ3) and LM. 95% confidence intervals are also shown.

as considered in this work and in prior works. 3) LMs are notoriously sensitive to prompt wording (Lu et al., 2022; Jiang et al., 2020). Thus, some of our findings comparing DQs and IQs may be sensitive to the specific wording in the prompt. 4) Since we use ACM CCS for our topics, the results are biased towards computer science references, though it would be straightforward to re-run the procedure on any given list of topics. 5) LMs have been shown to exhibit gender and racial biases (Swinger et al., 2019) which may be reflected in our procedure–in particular: our procedure may not recognize certain names as likely authors, or it may perform worse at matching names of people in certain racial groups where there is less variability in names.

Open-domain hallucination is an important but slippery concept that is difficult to measure. By studying it in the context of references using search engine results, we can quantitatively compare hallucinations across LMs and we can also quantitatively compare different black-box detection methods. We hope that our study of black-box self-detection of hallucinated references may shed light on the nature of hallucination more broadly, where classifying hallucinations is more challenging. It suggests that hallucination is not entirely a problem of training but rather one that can be addressed using only the same internal model representation with different generation procedures.

References

Sai Anirudh Athaluri, Sandeep Varma Manthena, V S R Krishna Manoj Kesapragada, Vineel Yarlagadda, Tirth Dave, and Rama Tulasi Siri Duddumpudi. 2023. Exploring the Boundaries of Reality: Investigating the Phenomenon of Artificial Intelligence Hallucination in Scientific Writing Through ChatGPT References. *Cureus* (April 2023). https://doi.org/10.7759/cureus.37432

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. 2023. Sparks of Artificial General Intelligence: Early experiments with GPT-4. https://doi.org/10.48550/arXiv.2303.12712 arXiv:2303.12712 [cs].

Owain Evans, Owen Cotton-Barratt, Lukas Finnveden, Adam Bales, Avital Balwit, Peter Wills, Luca Righetti, and William Saunders. 2021. Truthful AI: Developing and governing AI that does not lie. https://doi.org/10.48550/arXiv.2110.06674 arXiv:2110.06674 [cs].

Kelvin Guu, Albert Webson, Ellie Pavlick, Lucas Dixon, Ian Tenney, and Tolga Bolukbasi. 2023. Simfluence: Modeling the Influence of Individual Training Examples by Simulating Training Runs. https://doi.org/10.48550/arXiv.2303.08114 arXiv:2303.08114 [cs].

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of Hallucination in Natural Language Generation. *Comput. Surveys* 55, 12 (Dec. 2023), 1–38. https://doi.org/10.1145/3571730

Zhengbao Jiang, Frank F Xu, Jun Araki, and Graham Neubig. 2020. How can we know what language models know? *Transactions of the Association for Computational Linguistics* 8 (2020), 423–438.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, Deep Ganguli, Danny Hernandez, Josh Jacobson, Jackson Kernion, Shauna Kravec, Liane Lovitt, Kamal Ndousse, Catherine Olsson, Sam Ringer, Dario Amodei, Tom Brown, Jack Clark, Nicholas Joseph, Ben Mann, Sam McCandlish, Chris Olah, and Jared Kaplan. 2022. Language Models (Mostly) Know What They Know. https://doi.org/10.48550/arXiv.2207.05221 arXiv:2207.05221 [cs].

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Teaching Models to Express Their Uncertainty in Words. https://doi.org/10.48550/arXiv.2205.14334 arXiv:2205.14334 [cs].

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2022. Fantastically Ordered Prompts and Where to Find Them: Overcoming Few-Shot Prompt Order Sensitivity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 8086–8098.

Potsawee Manakul, Adian Liusie, and Mark J. F. Gales. 2023. SelfCheckGPT: Zero-Resource Black-Box Hallucination Detection for Generative Large Language Models. http://arxiv.org/abs/2303.08896 arXiv:2303.08896 [cs].

OpenAI. 2023. GPT-4 Technical Report. https://doi.org/10.48550/arXiv.2303.08774 arXiv:2303.08774 [cs].

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. https://doi.org/10.48550/arXiv.2203.02155 arXiv:2203.02155 [cs].

Bernard Rous. 2012. Major update to ACM's Computing Classification System. *Commun. ACM* 55, 11 (Nov. 2012), 12. https://doi.org/10.1145/2366316.2366320

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W. Kocurek, ...(421-others), and Ziyi Wu. 2022. Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models. https://doi.org/10.48550/ARXIV.2206.04615

Nathaniel Swinger, Maria De-Arteaga, Neil Thomas Heffernan IV, Mark DM Leiserson, and Adam Tauman Kalai. 2019. What are the biases in my word embedding?. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. 305–311.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288* (2023).

Annelies Vredeveldt, Peter J. van Koppen, and Pär Anders Granhag. 2014. The Inconsistent Suspect: A Systematic Review of Different Types of Consistency in Truth Tellers and Liars. In *Investigative Interviewing*, Ray Bull (Ed.). Springer, New York, NY, 183–207. https://doi.org/10.1007/978-1-4614-9642-7_10

Karen Weise and Cade Metz. 2023. When A.I. Chatbots Hallucinate. *The New York Times* (May 2023). https://www.nytimes.com/2023/05/01/business/ai-chatbots-hallucination.html

Ludwig Wittgenstein. 1953. *Philosophical Investigations*. Wiley-Blackwell, New York, NY, USA.

A Appendix

A.1 Qualitative findings

A qualitative examination of the titles generated by the LMs and their classifications according to the Bing search API revealed several interesting observations: 1) *Title mashups:* Many hallucinated titles were combinations of multiple existing titles. For example, a hallucinated title "Privacy-Preserving Attribute-Based Access Control in Cloud Computing" could be "fabricated" from (of the many possibilities) existing titles "Privacy-Preserving Attribute-Based Access Control for Grid Computing" and "Access Control in Cloud Computing". 2). *Bing's search flexibility:* The Bing quoted search heuristic is more lenient than exact match, ignoring more than just capitalization and punctuation. However, presumably since Bing quoted search is designed to facilitate title searches, it works well. 3) *Deceptive plausibility:* Some hallucinations were "plausible sounding" such as *A survey on X* for topic *X*, even when such a survey did not exist. 4) *DQ's false positives:* Direct methods may fail to identify hallucinations on "plausible sounding" titles such as surveys or book chapters. The indirect method also sometimes failed to identify a hallucination because the LM would consistently produce a "likely author" based on the title, for a given non-existent paper. For example, GPT-4 hallucinated the title *Introduction to Operations Research and Decision Making*, but there is a real book called *Introduction to Operations Research*. In all three IQs, it hallucinated the authors of the existing book, *Hillier Frederick S., Lieberman Gerald J.*. Similarly, for the hallucinated title *Exploratory Data Analysis and the Role of Visualization*, 2 of 3 IQs produced *John W. Tukey*, the author of the classic, *Exploratory Data Analysis*. 5) *IQ's false negatives:* The indirect method may sometimes fail to identify a grounded paper title which it can recognize/generate, as it may simply not be able to generate authors not encoded in its weights. Since, in many applications, identifying potential hallucinations is more important than recognizing all grounded citations, errors due to falsely marking an H as a G are arguably more problematic than classifying a G as an H. A manual examination of 120 examples is given in Section A.7.

Indirect Query (IQ)

U: Who were the authors of the reference, "Communication Complexity and Applications: A Survey"? Please, list only the author names, formatted as - AUTHORS: <firstname> <lastname>, separated by commas. Do not mention the reference in the answer.
A: AUTHORS: **Mark Braverman, Ankit Garg, Denis Pankratov, Omri Weinstein**

Overlap Query

U: Below are what should be two lists of authors. On a scale of 0-100%, how much overlap is there in the author names (ignore minor variations such as middle initials or accents)? Answer with a number between 0 and 100. Also, provide a justification. Note: if either of them is not a list of authors, output 0. Output format should be ANS: <ans> JUSTIFICATION: <justification>.
1. Mark Braverman, Ankit Garg, Denis Pankratov, Omri Weinstein
2. Ran Gelles, Ankur Moitra, Amit Sahai
A: ANS: **0 JUSTIFICATION: There is no overlap in the author names between the two lists.**

Figure 6: Top: Example of the IQ prompt templates instantiated with a candidate title. Bottom: An example of how we estimate overlap between a pair of answers using the LM.

A.2 Related Works

Open-domain hallucination were discussed in the context of GPT-4 (OpenAI, 2023; Bubeck et al., 2023), due to their prevalence and potential danger, Bubeck et al. (2023, page 82) write: "*Open domain hallucinations pose more difficult challenges, per requiring more extensive research, including searches and information gathering outside of the session.*" We show that open domain hallucinations can in fact be addressed, at least in part, without consulting external resources.

As mentioned, there are multiple definitions of hallucination. In this work, we use the term hallucinations to mean fabricated text that is not grounded in the training data. Factually incorrect generations can be decomposed into two types of errors (Evans et al., 2021): grounded errors which may be due to fallacies in the training data (e.g., that people use only 10% of their brains) and ungrounded errors. These two types of errors may need different techniques for remedy.

The grounded errors may be reduced by curating a training set with fewer errors or other techniques such as RLHF (Ouyang et al., 2022). However, the ungrounded errors which we study[3] are a fascinating curiosity which still challenge the AI community and one which is not clearly addressable by improving training data.

There is comparatively little prior work studying *open-domain groundedness* like ours. Some work (e.g., Guu et al., 2023) in attribution aims to understand which training examples are most influential in a given output. In recent independent work in the health space, Athaluri et al. (2023) did an empirical evaluation of hallucinated references within the medical domain. Similar to our approach, they used a Google search for exact string match as a heuristic for evaluating hallucinations. Our study of hallucinated references enables us to estimate the hallucination rates of different models, and, as discussed in prior work, the hallucination problem interestingly becomes more pressing as models become more accurate because users trust them more (OpenAI, 2023).

Related recent works include black-box techniques for measuring confidence in LM generations. Although these works are targeted at factual confidence, the approaches are highly related to our work. While Kadavath et al. (2022) use probability estimates drawn from LMs, it is straightforward to extend their procedures to generation-only LMs like ChatGPT using sampling. Lin et al. (2022) show that LMs can be used to articulate estimates by generating numbers or words as we do. Finally, Manakul et al. (2023) perform self-checks in the context of summarizing a document. All of these works use direct queries which influenced the design of our direct queries.

Due to space limitations, we do not discuss the work studying closed-domain hallucination (e.g., in translation or summarization) but instead refer the reader to recent survey of Ji et al. (2023).

A.3 Bing Search Reliability

The authors used Google search and other tools such as Google scholar in the course of their inquiry. Specifically, they adopted the following labeling protocol for labels: *"Grounded" if the search results yield a reference with an exact match for the title, or which is close enough to be naturally attributed to human error. Otherwise, it is "Hallucinated"*. For consistency, the human labelers also agreed on the following labels for four exemplars shown in Figure 7.

We show inter-rater reliability agreement computed using cohen's κ score among the labelers and the Bing in Table 2. This study shows that the labelling done using Bing search exact match is indeed reliable and could be used for identifying hallucinated references.

Generation: Theory of Computation: Design and Practise
Closest match: Theory of Computation
Label: Hallucinated

Generation: Cryptography through quantum lenses
Closest match: Crypography through quantum lenses: an insightful parody
Label: Hallucinated

Generation: Cryptography through quantum lenses: insightful parody
Closest match: Cryptography through quantum lenses: an insightful parody
Label: Grounded

Generation: Effective Classification using Negative Mining (ECNM)
Closest match: ECNM: Effective Classification with Negative Mining
Label: Grounded

Figure 7: Exemplar labels on which annotators agreed upon

A.4 Additional Experiments

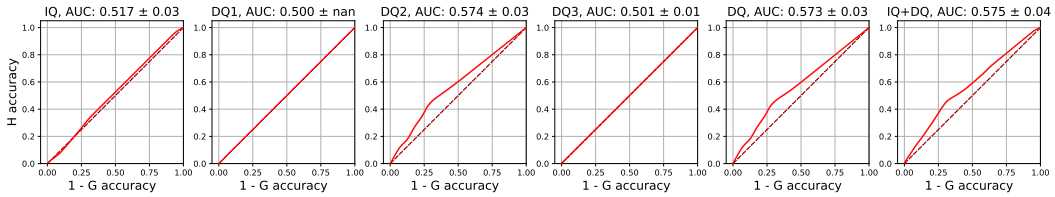
We show ROC and FDR metrics for L2-13B, L2-7B and GPT-3 models in Figure 9. We find that the procedures are not effective in detecting hallucinations, performing the worst for the L2-7B. Though

[3]One can also imagine ungrounded correct generations, such as a generated paper title that exists but is not in the training data, but we find these to be quite rare.

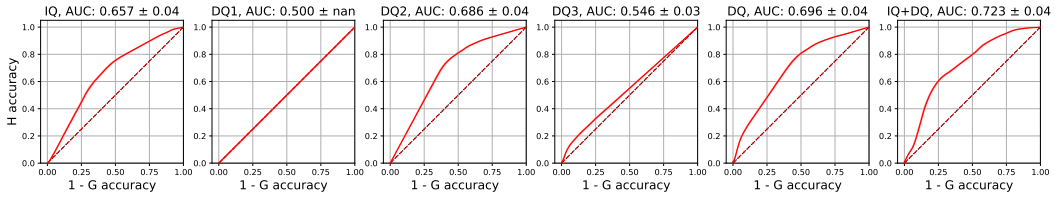
Table 2: Comparison of Cohen's Kappa

	Cohen's Kappa (κ)
person A and person B	0.96
person A and person C	0.98
person B and person C	0.98
person D and person A	0.96
person D and person B	1.0
person D and person C	0.98
person A and Bing	0.98
person B and Bing	0.98
person C and Bing	1.0
person D and Bing	1.0

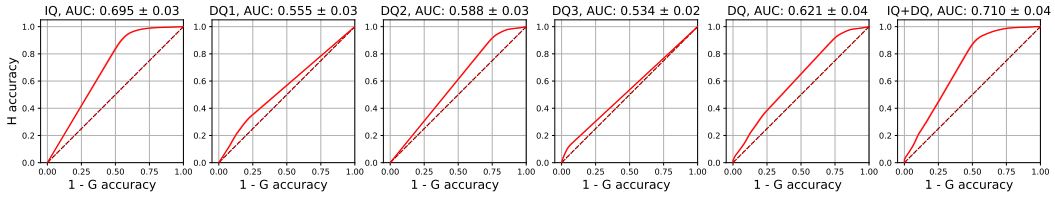
IQ helps the most for GPT-3, DQ2 approach helps the most for L2-13B and L2-7B. Consistent with our findings of other models, IQ+DQ ensemble approach performs the best.



(a) Llama-2-7b-chat



(b) Llama-2-13b-chat



(c) GPT-3

Figure 8: ROC Curves for the IQ and DQ approaches along with the ensemble approaches

A.5 Licenses and Terms of Use

According to the OpenAI terms of use Sharing and Publication policy,[4] they "welcome research publications related to the OpenAI API." Following the Bing Search API Legal Information[5], we do not store the results of the search queries but rather only whether or not there were any results. According to the ACM,[6] "The full CCS classification tree is freely available for educational and research purposes." (This section will be included with any published version of our paper.)

[4] `https://openai.com/policies/sharing-publication-policy`

[5] `https://www.microsoft.com/en-us/bing/apis/legal`

[6] `https://www.acm.org/publications/class-2012`

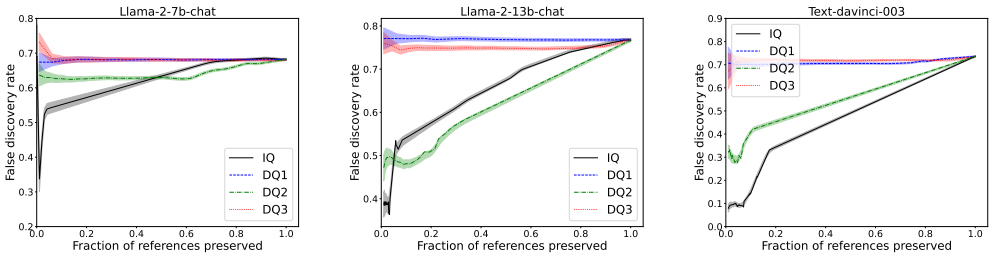


Figure 9: False discovery rate (FDR) vs. fraction of references preserved for each groundedness filter (IQ, DQ1, DQ2, DQ3) and LM. The preservation rate indicates the fraction of references preserved when a groundedness filter is applied to the raw generations of a LM. The FDR represents the fraction of preserved references that are actually hallucinations. For unachievable values of the fraction of references preserved (below the minimal fraction achievable by thresholding), we extrapolate each curve by uniformly subsampling references with maximal scores.

A.6 Computation and cost

We use OpenAI API for running the experiments on GPT-4, ChatGPT and GPT-3. We show the average tokens consumed for prompt and completion for each of the approaches and data generation per candidate query in Tables 3, 4 and 5. We estimate the cost based on the pricing details available as of May 2023.[7] For GPT-4, around 2.2M tokens were used amounting to roughly $74 to evaluate all approaches. For ChatGPT, around 2.3M tokens were used amounting to roughly $5. For GPT-3, around 2.1M tokens were used amounting to roughly $258. For Bing Search, we use an S1 instance of the Bing Search API [8]. We made 3,000 queries in all to this endpoint amounting to $75. Summing these costs gives a total of $412. The compute requirements of combining these results were negligible. While the exact model sizes and floating point operations are not publicly available for these models, the total cost gives a rough idea on the order of magnitude of computation required in comparison to the hourly cost of, say, a GPU on the Azure platform.

For running the experiments on Llama-2-chat series, we used a node with 8 V100 GPUs.

Table 3: GPT-4: Average number of tokens consumed

	DS	IQ	DQ1	DQ2	DQ3
Prompt	40.1	443.4	221.2	299.6	946.1
Completion	64.8	140.1	67.2	12.2	30.3

Table 4: ChatGPT: Average number of tokens consumed

	DS	IQ	DQ1	DQ2	DQ3
Prompt	40.1	437.3	224.1	302.2	1009.6
Completion	71.8	144.9	28.8	45.5	75.8

Table 5: GPT-3: Average number of tokens consumed

	DS	IQ	DQ1	DQ2	DQ3
Prompt	39.7	399.53	232.36	332.4	995.1
Completion	68.4	90.6	30.3	21.8	30.4

[7] `https://openai.com/pricing`

[8] `https://www.microsoft.com/en-us/bing/apis/pricing`

A.7 Examples of hallucinations and references

Tables 6, 7, 8, and 9 each display a careful inspection of 30 random candidate paper titles classified as H and G as determined by whether the Bing Search API returned any results. A manual search for each suggested title indicated that the vast majority of Hs are in fact hallucinations and the vast majority of Gs are in fact real references. We show the titles classified as H by Bing search along with closest manually discovered match for ChatGPT (Table 6) and GPT-4 (Table 8). We show the titles classified as G by Bing search along with the web links to the matched titles for ChatGPT (Table 7) and GPT-4 (Table 9). We also list the score assigned by the IQ method for all the sampled candidate titles. Interestingly, for both models there was a case in which the IQ method assigned the score of 1 to an H title. These H titles were *Design and Implementation of Digital Libraries: Technological Challenges and Solutions* for ChatGPT (Table 6) and *Enterprise Modeling: Tackling Business Challenges with the 4EM Approach* for GPT-4 (Table 8). In both of these cases, the titles were very similar to the closest manually discovered matched titles - *Design and Implementation of Digital Libraries* and *Enterprise Modeling with 4EM: Perspectives and Method*, respectively.

Table 6: Reference titles classified as H (hallucination) by Bing generated from ChatGPT. 30 randomly sampled titles are shown.

Reference title generated (Closest Match, if found)	**IQ Prob**
Quantum sensing for healthcare (NA)	0
Challenges and Solutions in Managing Electronic Records in Storage Systems (Electronic Records Management Challenges)	0
Hardware Verification Using Physical Design Techniques (NA)	0
A Framework for Verifying Recursive Programs with Pointers using Automata over Infinite Trees (Verification of recursive methods on tree-like data structures)	0
Robust Control for Nonlinear Time-Delay Systems with Faults (Robust Control for Nonlinear Time-Delay Systems)	0
Intelligent Scheduling for Autonomous UAVs using Discrete Artificial Intelligence Planning Techniques (NA)	0
An Overview of Database Management System Engines for Distributed Computing (NA)	0
The Aesthetics of Digital Arts and Media (VOICE: Vocal Aesthetics in Digital Arts and Media)	0
Improving Human-Robot Team Performance through Integrated Task Planning and Scheduling in a Complex Environment (Improved human–robot team performance through cross-training, an approach inspired by human team training practices)	0
Web Application Security: From Concept to Practice (Web Application Security)	0
A 28 nm high-density and low-power standard cell library with half-VDD power-gating cells (NA)	0
An Acoustic Interface for Touchless Human-Computer Interaction (NA)	0
Advances in Solid State Lasers Development and Applications: Proceedings of the 42nd Polish Conference on Laser Technology and Applications (Advances in Solid State Lasers Development and Applications)	0
Designing mobile information systems for healthcare (Design and Implementation of Mobile-Based Technology in Strengthening Health Information System)	0
Fault-tolerance and Reliability Techniques for Dependable Distributed Systems (Reliability and Replication Techniques for Improved Fault Tolerance in Distributed Systems)	0
Cyber-physical systems: A Survey and Future Research Directions on Sensor and Actuator Integration (Cyber-physical systems: A survey)	0
Performance evaluation of wireless sensor networks using network simulator-3 (NA)	0
Communication-Based Design for VLSI Circuits and Systems (NA)	0
Digital Media: The Intersection of Art and Technology (NA)	0
Toward a tool-supported software evolution methodology (NA)	0
Performance evaluation of temperature-aware routing protocols in wireless sensor networks (Performance Evaluation of Routing Protocols in Wireless Sensor Networks)	0
Computer-managed instruction and student learning outcomes: a meta-analysis (Effects of Computer-Assisted Instruction on Cognitive Outcomes: A Meta-Analysis)	0
An Empirical Analysis of Enterprise Resource Planning (ERP) Systems Implementation in Service Organizations in Jordan (Contributions of ERP Systems in Jordan)	0
Optimization of production planning in consumer products industry (Optimizing production planning at a consumer goods company)	0.01
Efficient Text Document Retrieval Using an Inverted Index with Cache Enhancement (NA)	0.11
Service OAM in Carrier Ethernet Networks	0.13
Introduction to Logic: Abstraction in Contemporary Logic (Introduction to Logic)	0.17
Query Processing and Optimization for Information Retrieval Systems (Query Optimization in Information Retrieval)	0.33
Cross-Platform Verification of Web Applications (Cross-platform feature matching for web applications)	0.33
Design and Implementation of Digital Libraries: Technological Challenges and Solutions (Design and Implementation of Digital Libraries)	1

Table 7: Reference titles classified as G (grounded) by Bing, generated from ChatGPT. 30 randomly sampled titles are shown.

Reference title generated (Matched title)	**IQ Prob**
JavaScript: The Good Parts (exact match)	1
Essentials of Management Information Systems (exact match)	1
Visualization Analysis and Design (exact match)	1
Forecasting: Methods and Applications (exact match)	1
Python for Data Analysis (exact match)	1
Introduction to Parallel Algorithms and Architectures: Arrays Trees Hypercubes (exact match)	1
Linear logic and its applications (Temporal Linear Logic and Its Applications)	1
Coding and Information Theory (exact match)	1
Introduction to Electric Circuits (exact match)	1
Concurrent Programming in Java: Design Principles and Patterns (exact match)	1
Cross-Platform GUI Programming with wxWidgets (exact match)	1
Embedded Computing and Mechatronics with the PIC32 Microcontroller (exact match)	0.87
Quantum entanglement for secure communication (Quantum entanglement breakthrough could boost encryption, secure communications)	0.78
An Introduction to Topology and its Applications (An introduction to topology and its applications: A new approach)	0.67
SQL Server Query Performance Tuning (exact match)	0.67
WCAG 2.1: Web Content Accessibility Guidelines (exact match)	0.61
Session Announcement Protocol (SAP) (exact match)	0.5
Introduction to Atmospheric Chemistry (exact match)	0.33
Data modeling and database design: Using access to build a database (exact match)	0.33
Introductory Digital Electronics: From Truth Tables to Microprocessors (exact match)	0.33
Trust Management: First International Conference, iTrust 2003, Heraklion, Crete, Greece (exact match)	0.25
Random geometric graphs (exact match)	0.08
Statistical Inference: An Integrated Approach (exact match)	0
Network Service Assurance (exact match)	0
Higher Order Equational Logic Programming (exact match)	0
Network Mobility Route Optimization Requirements (Network Mobility Route Optimization Requirements for Operational Use in Aeronautics and Space Exploration Mobile Networks)	0
Thermal management of electric vehicle battery systems (exact match)	0
Handbook of Imaging Materials (exact match)	0
The Secure Online Business Handbook: E-commerce, IT Functionality and Business Continuity (exact match)	0
Advanced Logic Synthesis (exact match)	0

Table 8: Reference titles classified as H (hallucination) by Bing generated from GPT-4. 30 randomly sampled titles are shown.

Reference title generated (Closest Match, if found)	**IQ Prob**
Privacy-Preserving Attribute-Based Access Control in Cloud Computing (Accountable privacy preserving attribute-based access control for cloud services enforced using blockchain)	0
Policy Measures for Combating Online Privacy Issues (NA)	0
Storage Security: Protecting Sanitized Data Attestation (NA)	0
Design of Scalable Parallel Algorithms for Graph Problems (NA)	0
Very Large Scale Integration (VLSI) Design with Standard Cells: Layout Design and Performance Analysis (NA)	0
Object-Oriented Modeling and Simulation of Complex Systems (Modelling and simulation of complex systems)	0
Overview of Electronic Design Automation (EDA) Tools & Methodologies (The Electronic Design Automation Handbook)	0
Printers and Modern Storage Solutions: The Role of the Cloud and Mobile Devices (NA)	0
Algebraic Algorithms and Symbolic Analysis Techniques in Computer Algebra Systems (Computer algebra systems and algorithms for algebraic computation)	0
Measuring Software Performance in Cross-platform Mobile Applications (NA)	0
A Comparative Study of OAM Protocols in Ethernet Networks (Carrier Ethernet OAM: an overview and comparison to IP OAM)	0
Best Practices in Board- and System-level Hardware Test Development (NA)	0
Algorithms for Symbolic and Algebraic Computations in Science and Engineering (NA)	0
Cryptography and Secure E-Commerce Transactions: Methods, Frameworks, and Best Practices (NA)	0
Quantum Computing: A Primer for Understanding and Implementation (A primer on quantum computing)	0
Understanding Network Management: Concepts, Standards, and Models (Network management: principles and practice)	0
Assessing network reliability: An analytical approach based on graph entropy (NA)	0
Language Models and their Applications to Information Retrieval (Language models for information retrieval)	0
Automated Support for Legacy Software Maintenance and Evolution (NA)	0
In-Network Traffic Processing: Advancements and Perspectives (NA)	0
Intellectual Property Law and Policy in the Digital Economy (Intellectual Property Law and Policy in the Digital Economy)	0
The Art and Science of Survey Research: A Guide to Best Practices (The Art and Science of Reviewing (and Writing) Survey Research)	0
Review of Network Mobility Protocols: Solutions and Challenges (A Review of Network Mobility Protocols for Fully Electrical Vehicles Services)	0
Program Semantics, Higher-Order Types, and Step Counting (NA)	0
Network Services: Management Strategies and Techniques (NA)	0
Machine Learning-Based Power Estimation and Management in Energy Harvesting Systems (NA)	0
The Evolution of Distance Education: Historical and Theoretical Perspectives (Distance Education: Historical Perspective)	0.17
The Economics of VLSI Manufacturing: A Cost Analysis Approach (NA)	0.5
Digital Decisions: The Intersection of e-Government and American Federalism (NA)	0.78
Enterprise Modeling: Tackling Business Challenges with the 4EM Approach (Enterprise Modeling with 4EM: Perspectives and Method)	1

Table 9: Reference titles classified as G (grounded) by Bing generated from GPT-4. 30 randomly sampled titles are shown.

Reference title generated (Matched title)	**IQ Prob**
Art and Electronic Media (exact match)	1
Network+ Guide to Networks (exact match)	1
Handbook of Automated Reasoning (exact match)	1
System Dynamics: Modeling, Simulation, and Control of Mechatronic Systems (exact match)	1
Information Visualization: Perception for Design (exact match)	1
The Emperor's New Mind: Concerning Computers, Minds and the Laws of Physics (exact match)	1
Computer Networks: A Systems Approach (exact match)	1
DNS and BIND: Help for System Administrators (exact match)	1
Introduction to Modern Cryptography (exact match)	1
Beyond Software Architecture: Creating and Sustaining Winning Solutions (exact match)	1
Practical Byzantine Fault Tolerance and Proactive Recovery (exact match)	1
Real-Time Systems: Scheduling, Analysis, and Verification (exact match)	1
Computational Complexity: A Modern Approach (exact match)	1
The Foundations of Cryptography: Volume 1, Basic Techniques (exact match)	1
Digital Library Use: Social Practice in Design and Evaluation (exact match)	1
Transactional Information Systems: Theory, Algorithms, and the Practice of Concurrency Control and Recovery (exact match)	1
Database System Concepts (exact match)	1
Pattern Recognition and Machine Learning (exact match)	1
File System Forensic Analysis (exact match)	1
The Archaeology of Science: Studying the Creation of Useful Knowledge (exact match)	0.78
Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data (exact match)	0.67
Electronic Design Automation for Integrated Circuits Handbook (exact match)	0.47
Modern VLSI Design: IP-Based Design (exact match)	0.39
Computational Complexity and Statistical Physics (exact match)	0.33
Probabilistic Methods for Algorithmic Discrete Mathematics (exact match)	0.33
Digital Rights Management: Protecting and Monetizing Content (exact match)	0.08
Deep Learning for Computer Vision: A Brief Review (exact match)	0.08
Random Geometric Graphs and Applications (exact match)	0.07
Concurrent Separation Logic for Pipelined Parallelization (exact match)	0
High-Level Synthesis for Real-time Digital Signal Processing (exact match)	0