Do Language Models Know When They're Hallucinating References?

Anonymous ACL submission

Abstract

State-of-the-art language models (LMs) are notoriously susceptible to generating hallucinated information. Such inaccurate outputs not only undermine the reliability of these models but also limit their use and raise serious concerns about misinformation and propaganda. In this work, we focus on hallucinated book and article references and present them as the *Drosophila* of research on hallucinations in language models due to their frequent and easy-to-discern nature. We posit that if a language model cites a particular reference in its output, then it should ideally possess sufficient information about its authors and content, among other relevant details. Using this basic insight, we illustrate that one can identify and sometimes even correct hallucinated references without ever consulting any external resources, by asking a set of *direct* or *indirect* queries to the language model about the references. These queries can be considered as "consistency checks." Our findings highlight that while LMs, including GPT-4, often produce inconsistent author lists for hallucinated references, they also often accurately recall the authors of real references. In this sense, the LM can be said to "know" when it is hallucinating references. Furthermore, these findings show how hallucinated references can be dissected, like Drosophila, to shed light on their nature. [1]

1 Introduction

Despite their unparalleled capabilities, recent large language models (LLMs) still exhibit a tendency to generate seemingly credible yet incorrect or unfounded information. This phenomenon is often referred to as the "hallucination" problem in the field of natural-language processing (NLP).[2] As one might imagine, the ramifications of these hallucination generations can be profoundly detrimental when these outputs find their way to critical domains such as healthcare, finance, law, or academic publications, where factuality is essential and non-negotiable. In fact, a recent example underlining the gravity of this issue involved a U.S. judge imposing sanctions on two New York lawyers for submitting a legal brief that included several fictitious case citations that were generated by ChatGPT.[3]

The are two primary challenges ahead for both researchers and practitioners within the NLP community. The first requires developing a deeper understanding of why these language models resort to fabricating information, while the second demands creating mechanisms that can not only promptly detect but also mitigate, if not completely prevent, inaccurate information in model outputs. To that effect, in this work, we study the problem of hallucinated book and article references related to the field of computer science and present a simple yet effective method to detect hallucinated references without relying on external tools.

Drawing inspiration from the role of the fruit fly, Drosophila melanogaster, as a benchmark in biological research, we suggest that the NLP community focus on the study of hallucinated references to better understand and mitigate wider hallucination challenges. These hallucinated references present distinct characteristics that render them suitable for study. First, their automatic classification is more straightforward than other hallucination varieties.[4] As an illustration, our method that leverages a search engine API closely matches each of four human expert evaluations, in at least 99 out of a sample of 100 references. Moreover, the static nature of academic reference titles, combined

[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). Additionally, we use the terms *hallucinate* and *fabricate* interchangeably throughout the paper.

[3] The original newspaper article detailing this incident can be found at this link. (Merken, 2023)

[4] In contrast, hallucinations like factoids pose classification challenges due to their nuanced phrasing and the uncertainty regarding their presence in training datasets.

with their broad online availability (on platforms like Google Scholar, Semantic Search, and arXiv), suggests they frequently appear in large, popular language modeling corpora. Additionally, many within the research domain already possess the skills and knowledge pertinent to studying these hallucinations. We therefore believe that just as fruit fly studies have enriched our understanding of biology, focusing on these specific reference hallucinations can pave the way for insights and solutions for more complex and challenging hallucination forms.

We outline the rest of this work as follows. We begin our investigations by addressing the questions of *when and why language models produce hallucinated references* and *what can be done to prevent them*. We then explore whether LLMs such as GPT-4 can recognize their own hallucinated outputs without relying on any external tools. While this approach does not fully unravel the reasons behind and solutions to hallucinations, it adds valuable perspective. Specifically, if LLMs can identify their own hallucinations, it implies the root of the issue may not lie in training or representation, but rather in the generation (i.e., decoding) process, given that models inherently possess enough data to potentially lower the rate of hallucinations. Our experiments compared different questioning strategies to use the LM to detect its own hallucinations across GPT and Llama based LM's.

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 *indirect queries* for evaluating hallucinations. Finally, we compare these to *direct queries* 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.

2 Preliminaries and Background

Following Ji et al. (2023), we define "hallucination" as fabricated text that has little or no grounding in the training data. It is worth noting that this is sometimes referred to as *open-domain hallucination* to distinguish it from *closed-domain hallucination* (see: Ji et al., 2023).[5] Our usage of the term *hallucination* aligns with the open-domain variant.

Distinguishing Groundedness from Correctness. The measure of *correctness* (or factuality) relies upon a comparison with ground-truth answers. Previous work on hallucination has blurred the line between groundedness and factuality. (Sometimes this distinction is also referred to as *honesty* versus *truthfulness* (Evans et al., 2021)). For example, the common misconception that "people use 10% of their brains" might be considered grounded if it is mentioned in the training data and assumed to be a true statement; however, this does not mean that it is factual, as it is not a scientifically correct statement.

Evaluating groundedness. Perfectly evaluating hallucinations would require access to the LM's training data. An advantage of the hallucinated reference problem is ease of (approximate) evaluation in that exact-match Web search is a reasonable heuristic for groundedness. This is because the vast majority of article titles present in the training data are included in Web search results—articles are meant to be published and shared, and publishers aim to make their work discoverable by search. Furthermore, references generally have titles that are specific enough not to spuriously occur on the Web. Regarding other types of hallucinations, besides article names, which cannot be as easily evaluated, we still hope that our methodology and findings would apply, even if evaluating those types of hallucinations would require access to the training data.

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 language models on question-answer problems. Manakul et al. (2023) apply a similar direct self-consistency check called SelfCheckGPT to identify relative hallucinations in a summarization context. These queries are direct true/false correctness queries. We test

[5] Closed-domain hallucination is typically studied in areas like abstractive summarization and machine translation, where the outputs are compared relative to a specific source document to be summarized or translated as opposed to the entirety of the training data.

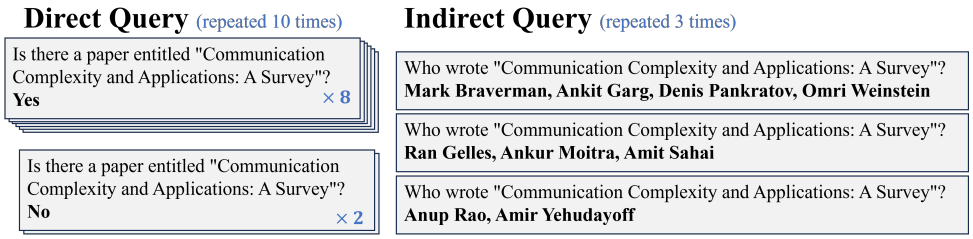


Figure 1: Example direct vs. indirect LM queries for predicting whether a given paper title is hallucinated or grounded. Direct queries are binary, repeated multiple times to estimate a probability. Indirect queries are open-ended, and their answers are compared to one another, using the LM, to output an agreement fraction. Language model generations are indicated in **boldface**. Prompts in this figure have been shortened for illustrative purposes.

similar approaches in the context of hallucinated references. Black-box generative approaches stand in contrast to the work that either introspects the weights on LMs (Azaria and Mitchell, 2023) or that consults existing databases (Guo et al., 2022).

Indirect queries (IQs). In addition, we suggest a new approach using what we call *indirect queries*. A direct query may ask, *Is the following paper real?* while an indirect query may ask, *Who are the authors of this paper?*, as illustrated in Figure 1. Answers are then generated to the indirect query in $i > 1$ independent sessions, and tested for consistency. The motivation for indirect queries comes from investigative interviews, where detectives are advised to interview individuals separately and ask open-ended questions. For instance, consistency may be better evaluated by asking multiple witnesses to *"Describe in detail what the suspect was holding"* rather than asking, *"Was the suspect holding a gun in their right hand?"* (Vredeveldt et al., 2014). In the context of reference hallucination, our hypothesis is that the likelihood of multiple generations agreeing on the same authors for a hallucinated reference would be smaller than the likelihood of multiple responses to a direct query indicating that the reference exists.

3 Related Work

Open-domain hallucinations, in the context of GPT-4 discussions (OpenAI, 2023; Bubeck et al., 2023), have garnered attention given their prevalence and associated hazards. Bubeck et al. (2023, pg. 82) comment: "*Open domain hallucinations pose more difficult challenges, per requiring more extensive research, including searches and information gathering outside of the session.*" Yet, our work provides evidence that addressing these hallucinations can be achieved without turning to 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[6] 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

[6]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.

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).

4 Methodology: Consistency Checks

We now provide an overview of our simple yet effective consistency check methodology, explaining how we perform a series of *direct* and *indirect* queries to detect hallucinated references.[7]

4.1 *Direct* Queries

The direct query (DQ) method examines if a particular title exists using a format illustrated in Figure 2. We use three simple DQ templates (DQ1, DQ2, and DQ3), drawing insights from Kadavath et al. (2022); Manakul et al. (2023). In each case, an LM to expected to answer "yes" if it believes that the reference *actually* exists and "no" otherwise.

DQ1 asks outright if the reference does indeed exist. While being simple, this approach can sometimes be problematic as some chat-bot-based LMs have strong biases in answering questions when phrased in a particular way (without any proper context) (Lu et al., 2022). DQ2 and DQ3, on the other hand, incorporate context by stating that the reference was generated by an LM or an assistant. Moreover, DQ3 takes it a step further by providing additional references for comparison, an approach advocated in Kadavath et al. (2022).

For each query, we generate $j \geq 1$ completions to approximate the probability distribution of the model about the existence of the generated reference.[8] We measure the *groundedness* rate (see Section 2) by dividing the number of completions containing the word "yes" by the total number of completions.[9] We also consider an *ensemble direct query*, denoted by DQ, that simply averages the scores of DQ1, DQ2, and DQ3.

4.2 *Indirect* Queries

The indirect query (IQ) method involves two main steps: *interrogation* and *overlap estimation*.

Step 1: Interrogation. For each reference, we first pose j indirect queries to the LM, asking about the authors of the generated reference, for instance, as shown in Figure 3 (top).

Step 2: Overlap estimation.. Next, we asses the degree of similarity (overlap) between the model responses from the previous step by using a separate query template, as shown in Figure 3 (bottom). We initially tested string-matching techniques which we found to be inaccurate and required hyperparameters. Name matching is known to be a thorny problem and one which we found could be performed accurately when using pretrained LMs to compare in pairs.[10]

The intuition behind our approach is simple: If a language model provides similar (that is, consistent) responses to multiple indirect queries, it can then be assumed that the model is most likely familiar with the reference and that it has seen the reference during its training; such a reference could therefore be deemed *grounded*. On the other hand, varied responses might signal that the model does not intrinsically possess knowledge about the author(s) and content of the reference; hence, it can be speculated that the model has presumably not seen the reference during its training and that the reference is mostly likely fabricated.

We also consider an ensemble IQ+DQ check that averages the scores of IQ and the DQ ensemble.

Finally, we highlight that our consistency checking methods do not rely on external resources such as Google Scholar or Semantic Search. It instead uses the same language model throughout the hallucination detection process.

[7] Note that this pipeline is run separately for each of our LMs, so there is no mixing across LMs.

[8] For both direct and indirect queries, we employ a temperature rate of 1 when $j > 1$ (i.e., generating multiple completions) and 0 when $j = 1$ (i.e., generating a single completion). The choice of 0 is intended to capture the model's top pick if a single output is generated.

[9] This means that empty or otherwise invalid answers are assigned "no." We do not assume that this score is calibrated as our analysis considers arbitrary probability thresholds.

[10] It is worth noting that LMs sometimes return responses that do not consist of a list of authors (e.g., a long response beginning with *"I could not find a specific reference titled..."*. In such cases, we simply set the overlap rate to 0. We also note that traditional parsing and string-matching techniques could be leveraged as an alternative to LMs in this overlap estimation phase.

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 2: Examples of the three direct prompt templates used for the direct queries, instantiated with candidate reference titles.

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

5 Experimental Details

Here, we describe the steps taken to build a corpus of article and book references pertaining to computer science topics for each language model, as well as the automatic labeling heuristic used to annotate these generated references.

5.1 Dataset Construction Using ACM CCS

To ensure that our corpus of references is representative of a broad spectrum of the topics in computer science, we used the ACM Computing Classification System (CCS; Rous, 2012) as our main source. The CCS provides a structured taxonomy for computer science, ranging from 12 high-level subjects down to 543 specific topics.

From the 543 topics, we selected a uniformly random subset of 200 topics, each denoted as *area: topic* (e.g., *Information retrieval: Retrieval models and ranking*). For each chosen topic, we then prompted each LM to generate five related reference titles, amounting to 1,000 total titles per LM as shown in Figure 4.

5.2 Automatic Labeling and Verification

Next, we employed the Bing search engine API[11] as an automatic labeling heuristic, labeling each of the 1,000 reference titles generated in the previous step as either *grounded* (G) or *hallucinated* (H) based on exact matches. The reference title surrounded by quotes is searched in the web (e.g., "LMs are few-shot learners"). We label the reference as hallucinated if no results are retrieved and as grounded otherwise.

To assess the efficacy of this automated pipeline, we asked four expert annotators (all computer scientists familiar with academic writing and publication) to manually label 10% of the GPT-4-generated references. One of the annotators agreed with Bing on 100% of the labels, and the other three each had 99% agreement with Bing, indicating strong support for the reliability of the automatic labeling pipeline. See Appendix A for more details.

[11] https://www.microsoft.com/en-us/bing/apis/bing-web-search-api

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

5.3 Models and Parameters

We evaluate the OpenAI LMs GPT-3 (*text-davinci-003*), ChatGPT (*gpt-35-turbo*), and GPT-4 (*gpt-4*) using the Azure OpenAI API and the open-source Llama 2 Chat *llama-2-*-chat* series LMs abbreviated as L2-7B, L2-13B, and L2-70B (Touvron et al., 2023).

We select $i = 3$ indirect query results and take the average of the overlapping evaluations to compute the final score for each indirect query experiment. For direct query experiments, we sample $j = 10$ judgments at temperature 1.0 and report the fraction of *yes* responses as a final score.

5.4 Metrics

Receiver Operating Characteristic (ROC) Curves. 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. We visualize this trade-off using a standard receiver operating characteristic (ROC) curve (Fawcett, 2006) and summarize overall detection performance using the area under the ROC curve (AUC).

False Discovery Rate (FDR) Curves. 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. We show how these two quantities can be traded off using false discovery rate (FDR) curves. 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.

6 Results and Discussion

In this section, we discuss the performance of the indirect and direct methods using quantitative metrics, and present interesting qualitative findings.

6.1 Quantitative Analysis

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.

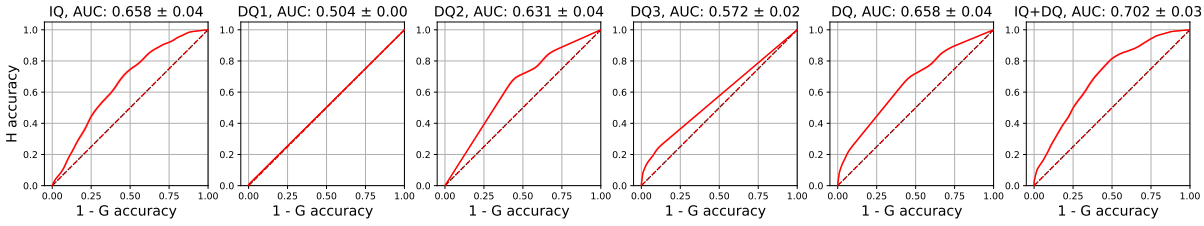
LLM	GPT-4	ChatGPT	GPT-3	L2-70B	L2-13B	L2-7B
H%	46.8%	59.6%	73.6%	66.2%	76.7%	68.3%

Table 1: The hallucination rate (out of 1000 generated titles), as determined by ground-truth labels assigned using the Bing search API.

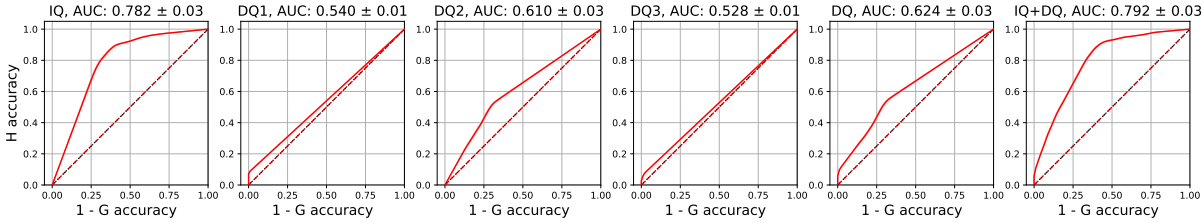
Due to space limitations, we show the ROC and FDR curves for GPT-4, ChatGPT, and L2-70B and defer additional LM results to Appendix B.

The ROC curves are shown for each approach and model in Figure 5. 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 overall as quantified via AUC. For GPT-4 (Figure 5c), 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.

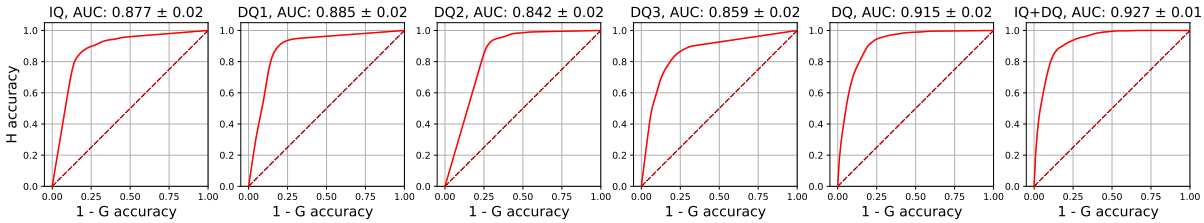
Figure 6 shows FDR curves for the three models. For L2-70B and ChatGPT, the IQ method achieves significantly lower FDR and a provides a substantially better FDR-preservation rate trade-off than the other approaches. For GPT-4, both IQ



(a) Llama-2-70b-chat



(b) ChatGPT



(c) GPT-4

Figure 5: For each individual (IQ, DQ1-3) and ensemble (DQ, IQ+DQ) consistency check, we display the trade-off between accuracy on grounded and hallucinated references with 95% confidence bands based on 100 bootstrap replicates and a 95% confidence interval for the AUC using the DeLong et al. (1988) estimate of standard error.

and DQ methods offer low FDR with comparable trade-offs.

Overall, IQ appears to be more accurate than DQ1-3 for ChatGPT and L2-70B, while for GPT-4 DQ1-3 and IQ were similarly effective. For each LM, ensembling further boosts classification performance with the IQ+DQ ensemble obtaining the best AUC and lower FDR curves for each LM.

The compute costs, which involve ≈6.6 million tokens and $412, are discussed in Section D.

6.2 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 indirect queries, 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 indirect queries produced *John W. Tukey*, the author of the classic, *Exploratory Data Analysis*. 5) *IQ’s false negatives:* The indirect method may some-

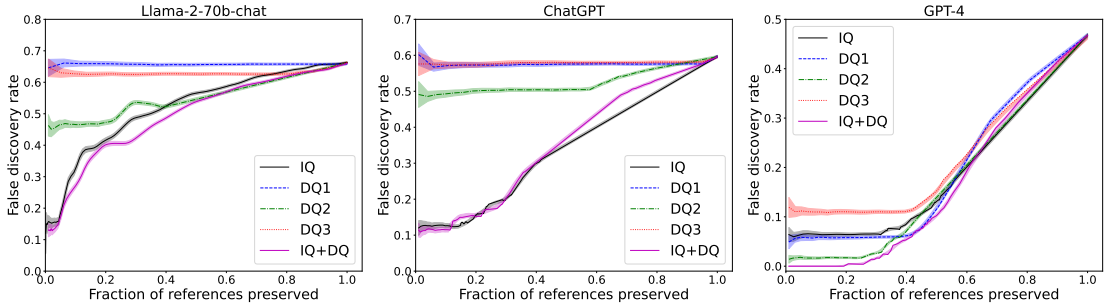


Figure 6: False discovery rate (FDR) vs. fraction of references preserved for each groundedness filter and LM. We compute 95% confidence intervals from a 100-replicate bootstrap mean ± 1.96 times the bootstrap standard error.

times 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 Appendix E.

7 Conclusions

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. Of course, for the sole purpose of detection, one could achieve higher accuracy by directly consulting curated publication indexes. However, we hope that our study of black-box self-detection of hallucinated references sheds light on the nature of open-domain hallucination more broadly, where detecting 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. While our direct and indirect query methods are only partially reliable and impractically expensive, we hope they may pave the way towards more efficient methods that generate text with fewer hallucinations and thereby reduce potential harms of language models.

There are several directions for future work. 1) *Improved decoding techniques:* An important consequence of our work is the recognition that reducing hallucination may be a problem at generation time. Thus, inventing improved (non-black-box) generation procedures is thus a crucial direction for future work. 2) *Additional indirect questions:* One may improve accuracy by adding more indirect questions such as year or venue. These pose additional challenges as a paper with the same title and authors may often appear in multiple venues (e.g., arXiv, a workshop, a conference, and a journal) in different years. 3) *Generalisability:* It would be very interesting to see if the methods we employ could be used to identify other types of open-domain hallucinations beyond references. Even though hallucinated references are often given as a blatant example of hallucination, perhaps due to the ease with which they can be debunked, these other types of hallucination are also important. Following the investigative interviewing analogy, one way to aim to discover general hallucinations would be to query the LM for "notable, distinguishing details" about the item in question. One could then use an LM to estimate the consistency between multiple answers. However, as mentioned for other domains besides references, it may be impossible to determine whether or not a generation is a hallucination without access to the training set (and unclear even with such access).

8 Limitations

There are several limitations of this work: 1) *Inaccessible training data:* We consider web as a contending proxy for the models' training data. However, we cannot conclude what is truly grounded versus hallucination since we do not have access to the training data. 2) *Hallucination spectrum:* The notion of hallucination is not entirely black and white as considered in this work and in prior works. For example, a generated reference that is a substring or superstring of an existing title is hard to classify with the binary scheme. 3) *Prompt sensitivity:* LMs are notoriously sensitive to prompt wording (Lu et al., 2022; Jiang et al., 2020; Shin

et al., 2020; Gao et al., 2021). Thus, some of our findings comparing direct and indirect queries may be sensitive to the specific wording in the prompt. 4) *Domain-specific reference bias:* Since we use ACM Computing Classification System 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) *Gender and racial biases:* 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. Since our work compares LMs and hallucination estimation procedures, the risk is lower compared to a system that might be deployed using our procedures to reduce hallucination. Before deploying any such system, one should perform a more thorough examination of potential biases against sensitive groups and accuracy across different research areas.

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*.

Amos Azaria and Tom Mitchell. 2023. The Internal State of an LLM Knows When its Lying. ArXiv:2304.13734 [cs].

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. ArXiv:2303.12712 [cs].

Elizabeth R DeLong, David M DeLong, and Daniel L Clarke-Pearson. 1988. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*, pages 837–845.

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. ArXiv:2110.06674 [cs].

Tom Fawcett. 2006. An introduction to roc analysis. *Pattern recognition letters*, 27(8):861–874.

Tianyu Gao, Adam Fisch, and Danqi Chen. 2021. Making pre-trained language models better few-shot learners. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3816–3830.

Zhijiang Guo, Michael Schlichtkrull, and Andreas Vlachos. 2022. A Survey on Automated Fact-Checking. *Transactions of the Association for Computational Linguistics*, 10:178–206.

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. 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. *ACM Computing Surveys*, 55(12):1–38.

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*, 8:423–438.

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. ArXiv:2207.05221 [cs].

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Teaching Models to Express Their Uncertainty in Words. 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)*, pages 8086–8098.

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

Mary L McHugh. 2012. Interrater reliability: the kappa statistic. *Biochemia medica*, 22(3):276–282.

Sara Merken. 2023. New york lawyers sanctioned for using fake chatgpt cases in legal brief. *Reuters*.

OpenAI. 2023. GPT-4 Technical Report. 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. ArXiv:2203.02155 [cs].

Bernard Rous. 2012. Major update to ACM's Computing Classification System. *Communications of the ACM*, 55(11):12.

Taylor Shin, Yasaman Razeghi, Robert L Logan IV, Eric Wallace, and Sameer Singh. 2020. Autoprompt: Eliciting knowledge from language models with automatically generated prompts. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4222–4235.

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.

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*, pages 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*.

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 Ray Bull, editor, *Investigative Interviewing*, pages 183–207. Springer, New York, NY.

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

A Bing Search Reliability

Before assigning manual grounded or hallucination labels to each reference title, each expert annotator was given the instructions shown in Figure 7. Along with a given reference title, the annotators were provided with a corresponding Google search link as shown in Table 2. For consistency, the human labelers also agreed on the labels for the four exemplars shown in Figure 8.

We show inter-rater reliability agreement computed using Cohen's κ score (McHugh, 2012) between the labelers and the automated Bing labels in Table 3. The results demonstrate that the automated labeling generated via Bing search exact match reliably matches the judgments of human experts.

B Supplementary Experimental Details

We show ROC and FDR metrics for L2-13B, L2-7B and GPT-3 models in Figure 9 and Figure 10 respectively. We find that the procedures are not effective in detecting hallucinations, performing the worst for the L2-7B. Though 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.

C Licenses and Terms of Use

According to the OpenAI terms of use Sharing and Publication policy,[12] they "welcome research publications related to the OpenAI API." Following the Bing Search API Legal Information[13], we do not store the results of the search queries but rather only whether or not there were any results. According to the ACM,[14] "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.)

D 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 4 to 6. We estimate the cost based on the pricing details available as of May 2023.[15] 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

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

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

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

[15] https://openai.com/pricing

```
You are provided with 100 reference titles.

Your task is to label these reference titles as "Grounded" (G) or
"Hallucinated" (H).

You are provided with the search_url against each title, please
go over that to observe the search results. Additionally, you may
also use  other tools such as Google scholar while assigning the
ground truth labels to the reference titles.

Label a title as "G" 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, label it as "H".
```

Figure 7: Labeling instructions shown to the expert human annotators.

Table 2: Sample of 2 titles out of 100 titles given to the expert human annotators for labeling.

Reference Title	Search Url	(H/G)
Introduction to Autonomous Robots: Mechanisms, Sensors, Actuators, and Algorithms	link	?
Timing Aware Placement and Routing in FPGAs	link	?

```
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 8: Exemplar labels upon which all expert human annotators agreed prior to assigning manual labels.

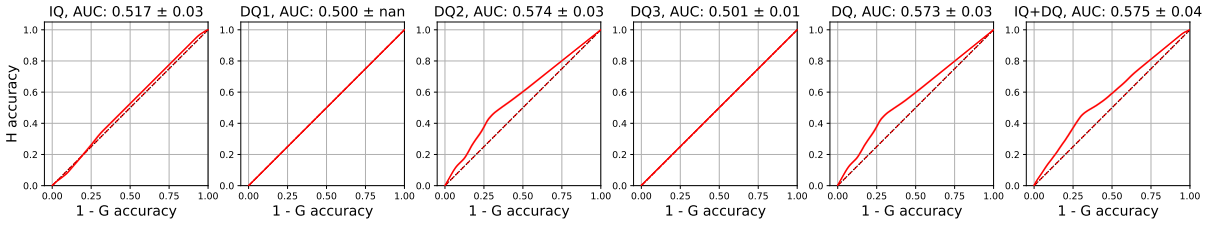
2.1M tokens were used amounting to roughly $258. For Bing Search, we use an S1 instance of the Bing Search API [16]. We made 3,000 queries in all to this endpoint amounting to $75. Summing these costs gives a total of $412. The compute requirements of combining these results were negligible. While the 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.

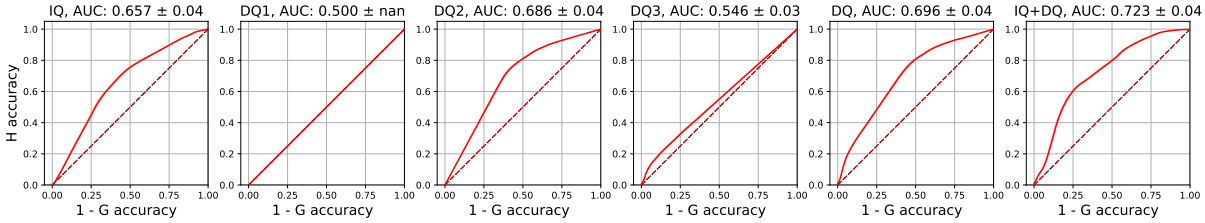
[16] https://www.microsoft.com/en-us/bing/apis/pricing

E Examples of Hallucinations and References

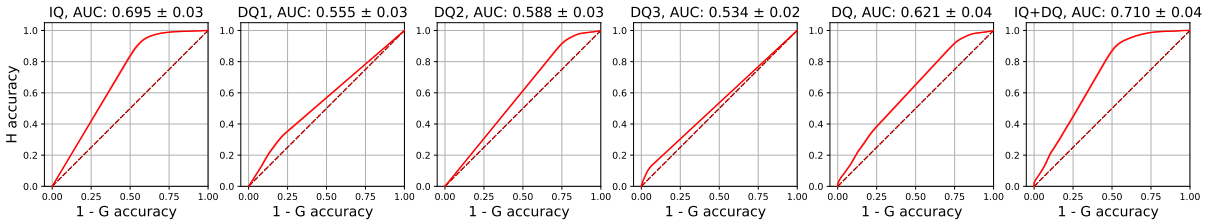
Tables 7 to 10 each display a careful inspection of 30 random candidate paper titles classified as H and G as determined by whether the Bing Search API returned any results. A manual search for each suggested title indicated that the vast majority of Hs are in fact hallucinations and the vast majority 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 7) and GPT-4 (Table 9). We show the titles classified as G by Bing search along with the web links to the matched titles for ChatGPT (Table 8) and GPT-4 (Table 10). We also list the score assigned



(a) Llama-2-7b-chat



(b) Llama-2-13b-chat



(c) GPT-3

Figure 9: ROC Curves for the IQ and DQ approaches along with the ensemble approaches

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 7) and *Enterprise Modeling: Tackling Business Challenges with the 4EM Approach* for GPT-4 (Table 9). 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 3: Cohen's κ measure of inter-rater reliability between each pair of expert human evaluators and between each expert and the automated Bing labeling described in Section 5.2. The range of Cohen's κ is $[-1, 1]$ with a value of 1 indicating perfect agreement. A value above 0.9 is considered "almost perfect" agreement (McHugh, 2012).

	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	0.98

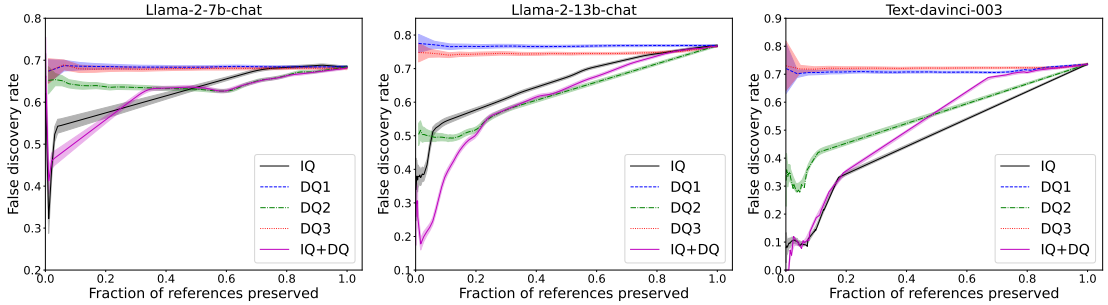


Figure 10: False discovery rate (FDR) vs. fraction of references preserved for each groundedness filter and LM. The preservation rate indicates the fraction of references preserved when a groundedness filter is applied to the raw generations of a LM. The FDR represents the fraction of preserved references 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. We compute 95% confidence intervals from a 100-replicate bootstrap mean ± 1.96 times the bootstrap standard error.

Table 4: 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 5: 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 6: 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

Table 7: 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 8: 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 9: 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 10: 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