REASONS: A benchmark for REtrieval and Automated citationS Of scieNtific Sentences using Public and Proprietary LLMs

Anonymous ACL submission

Abstract

Automatic citation generation for sentences in a document or report is paramount for intelligence analysts, cybersecurity, news agencies, and education personnel. In this research, we investigate whether large language models (LLMs) are capable of generating references based on two forms of sentence queries: (a) *Direct Queries*, LLMs are asked to provide author names of the given research article, and (b) *Indirect Queries*, LLMs are asked to provide the *title* of a mentioned article when given a sentence from a different article. To demonstrate where LLM stands in this task, we introduce a large dataset called **REASONS** comprising abstracts of the 12 most popular domains of scientific research on arXiv. From ∼ 20K research articles, we make the following deductions on public and proprietary LLMs: (a) State-of-the-art, often called anthropomorphic GPT-4 and GPT 3.5, suffers from high pass percentage (PP) to minimize the hallucination rate (HR). When tested with Perplexity.ai (7B), they unexpectedly made more errors; (b) Augmenting relevant metadata lowered the PP and gave the lowest HR; (c) Advance retrieval-augmented generation (RAG) using Mistral demonstrates consistent and robust citation support on indirect queries, and matched performance to GPT-3.5 and GPT-4. The HR across all domains and models decreased by an average of 41.93%, and the PP reduced to 0% in most cases. In terms of generation quality, the average F1 Score and BLEU were 68.09% and 57.51%, respectively; (d) Testing with adversarial samples showed that LLMs, including the Advance RAG Mistral, struggle to understand context, but the extent of this issue was small in Mistral and GPT-4-Preview. Our study contributes valuable insights into the reliability of RAG for automated citation generation tasks.

1 Introduction

The development of LLMs marks a significant advancement in computational linguistics and artificial intelligence (AI) (Tamkin and Ganguli, 2021).

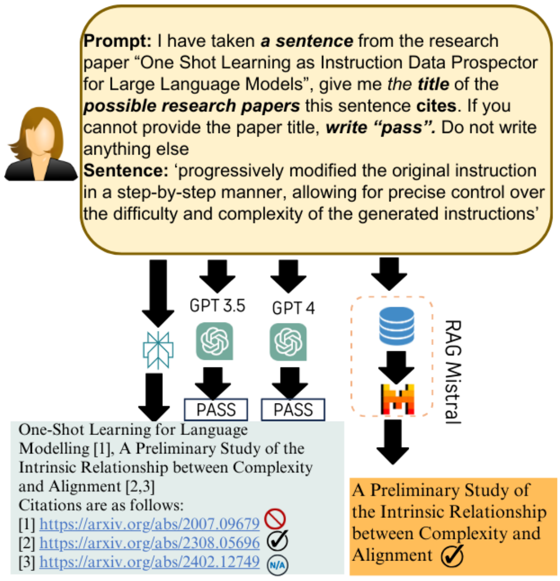

Figure 1: An illustration and motivating example for investigating LLMs for automatic citation generation task. Perplexity.ai, which is an LLM-based search engine, yields a citation that doesn't exist [1], an incorrect one [3], and a correct citation [2]. Advance RAG (defined in this research) improved context understanding and citation generation quality. Time: Feb. 05, 2024.

LLMs, such as OpenAI's GPT series, have shown remarkable capabilities in text generation (Zhao et al., 2023), and question-answering systems (Rasool et al., 2023; Elgedawy et al., 2024). However, their limitations become apparent as they become more integrated into various domains, including defense (Schwinn et al., 2023), news media (Fang et al., 2023), and education (Yan et al., 2024; Hung et al., 2023; Augenstein et al., 2023). The critical issue is their propensity to generate hallucinated sentences and propagate factually inaccurate pieces of information **without reference** (Ji et al., 2023; Rawte et al., 2023). These inaccuracies diminish the models' reliability and erode users' trust, a vital component in their widespread adoption.

Commercial LLM-based search systems, including Bing Search-powered GPT 4 (Mehdi, 2024) and Perplexity.ai (Roose, 2024), are still not capable enough of resolving the issue of citation generation to confirm the scientific feasibility of either a generated sentence(s) or given sentence(s) from the scientific literature. For instance, Figure 1 shows how proprietary LLMs respond to the zero-shot indirect query. It is evident from the figure that while general-purpose LLMs like GPT-3.5 and GPT-4 'pass' the query, task-specific LLM Perplexity does generate relevant citations but still shows hallucination. Consider the following three use cases motivating this research:

Citation Generation in Research Articles and News Reports: LLMs can generate highly persuasive and realistic content, especially in writing research articles or news reports, making it challenging for users to distinguish between genuine and fabricated information Nakano et al. (2021); Menick et al. (2022); Kumarage and Liu (2023).

Citation Generation in Reports for Organizational Cybersecurity: LLMs are trained on massive datasets and can inadvertently reveal sensitive information, which can put an organization at risk without proper citations (Yamin et al., 2024).

Citation Generation in Reports for Legal: In a significant event, an attorney tried employing ChatGPT for legal analysis during a trial (see subsection A.1)(Bohannon, 2023). While ChatGPT generated information, it failed to capture the nuanced complexities and critical legal precedents needed for the case. This underscores the importance of confirming and sourcing accurate legal citations and precedents relevant to the case. We **contribute** by addressing these challenges with the following: (**A**) Introduce **REASONS**, a dataset created by extracting related works from IEEE articles spanning 12 scientific domains from 2017 to 2023. (**B**) We employ a new RAG training regime to develop Advance RAG. Advance RAG and Naïve RAG examine the factual integrity of the information retrieved by dense retrievers and its presentation as citations by LLMs. (**C**) We evaluate both proprietary and public LLMs and their RAG counterparts (**10** models) to assess their contextual awareness using metrics like Pass Percentage (PP) and Hallucination rate (HR). Additionally, we have measured the quality of citation generation using F-1 and BLEU scores. (**D**) We conduct an adversarial examination to provide a clear assessment of context awareness regarding citation generation in LLMs.

Findings:(I) Perplexity, faces a major challenge when dealing with *indirect and direct query* on the **REASONS** dataset (Figure 2 - Figure 5, and in Appendix A Table 6 - Table 9).**(II)** Citation generation is enhanced uniformly across public and proprietary LLMs when metadata like abstract and title are considered with *indirect query* (Figure 3 and Figure 5, along with Table 7 and Table 9). **(III)** Advance RAG with Mistral LLM outperforms other competitive proprietary and public LLMs. This performance is realized by a reduction in the HR and increments in F-1 and BLEU scores (Figure 3 and Figure 5 (last two bars) and Table 7 and Table 9 (last two columns)). **(IV)** For domains such as Quantum Computing and Biomolecules that are heavy in mathematics and numerals, there was a substantial decline in citation generation quality and an increase in HR. Adversarial examination strengthens our understanding that despite being exorbitantly large, LLMs lack context awareness (Table 2). **(V)** Advance RAG did provide convincing evidence of context understanding (Table 2). Further improvements in RAG-based LLMs are desirable, and utilizing **REASONS** dataset can provide valuable insights into context understanding and provenance in tasks such as hypothesis generation.

2 Background

Early Techniques in Citation Recommendation: The practice of citing sources is a cornerstone of academic and professional writing, serving as the bedrock for reliability, and truthfulness in scholarly work (Cronin, 1981). The evolution of citation recommendation systems mirrors the broader advancements in computational linguistics and natural language processing (NLP) (Bai et al., 2019; Ali et al., 2021). Initial methods in citation recommendation focused on basic techniques such as text feature-based systems (Strohman et al., 2007), simple keyword matching, and basic statistical methods (Bethard and Jurafsky, 2010). Context-aware citation recommendation systems supplemented these methods (He et al., 2010; Ebesu and Fang, 2017; Jeong et al., 2020a; Huang et al., 2021). However, their inability to grasp deeper textual contexts limited their effectiveness.

Machine learning in Citation Recommendation The integration of machine learning marked a significant leap in citation recommendation systems (Agarwal et al., 2005; Küçüktunç et al., 2012). These systems began to exhibit an improved understanding of the text, although they still lacked a nuanced grasp of complex contexts (Tran et al., 2015). The application of neural networks revolutionized citation recommendation. NLP algorithms, capable of parsing complex sentence structures, started identifying relevant themes for contextually appropriate citation recommendations (Zarrinkalam and Kahani, 2013; Beel et al., 2016; Iqbal et al., 2020). Concurrently, graph-based models, visualizing literature as interconnected networks, enhanced citation recommendations by considering content similarity and citation patterns (Ali et al., 2020; Chakraborty et al., 2015). With deep learning, citation recommendation systems began incorporating semantic analysis, employing models like word embeddings and neural networks for a more nuanced understanding (Yang et al., 2018; Bhagavatula et al., 2018; Vajdecka et al., 2023). Adapted from commercial use, collaborative filtering also emerged, recommending citations based on similar citation behaviors (Wang et al., 2020).

Large Language Models in Citation Generation: The advent of LLMs like GPT-3 and its successors has further transformed NLP. Initial language model systems such as those based on BERT have significantly improved citation recommendation by converting unstructured text into meaningful vectors (Jeong et al., 2020b; Devlin et al., 2018; Bhowmick et al., 2021). Recent studies have focused on evaluating the fidelity of generated text to its sources (Ji et al., 2023). (Rashkin et al., 2023) introduced the "attributable to identified sources" (AIS) score, while (Bohnet et al., 2022) and others (Honovich et al., 2022; Yue et al., 2023) have focused on automating AIS. Concurrent work by (Liu et al., 2023) explored human evaluation of commercial generative search engines such as Bing. Chat, NeevaAI, perplexity.ai, and YouChat.

Despite these advancements, LLMs in citation recommendation still struggle with generating accurate information and providing references, as shown in studies by (Ji et al., 2023; Zheng et al., 2023). Even commercial systems like BingChat and Perplexity.ai, which boast advanced technologies, lack reliability, especially when generating analytical reports requiring proper citations.

This limitation necessitates an approach closely aligning with RAG. RAG compels LLMs to provide citations alongside the generated text. The concept of retrieval-augmented LLMs has gained traction in recent years following (Guu et al., 2020; Borgeaud et al., 2022; Izacard et al., 2022; Khandelwal et al., 2019; Schick et al., 2023; Jiang et al., 2023b; Yao et al., 2022; Gao et al., 2023). We evaluate public and proprietary LLMs and their RAG counterparts on citation generation using **REASONS**, a meticulously curated dataset from arXiv spanning key domains in computer science and related fields. This allows us to assess the LLM's ability to identify a given sentence's source accurately.

Domain	Paper Count	IEEE Papers	Citation Count
CV	5488	1028	3437
Robotics	3656	292	776
Graphics	1796	384	1417
IR	1741	564	1654
AI	1697	531	2021
NLP	1526	293	1092
Cryptography	1084	371	1106
NNC	892	111	326
HCI	761	112	229
Databases	723	115	182
QC	421	126	456
Biomolecules	119	17	27
Total	**19904**	**3944**	**12723**

Table 1: Our benchmark dataset, **REASONS**, includes papers and sentences from 12 domains. It primarily features ten domains in computer science and 2 in biology. Full forms of domain acronyms are provided in subsection A.5.

3 Problem Setup

Scope of REASONS: The dataset comprises sentences gathered from the *related work* sections of articles in computer science and biology available on arXiv (arX). Summary is provided in Table 1. Exclusions were made for mathematics, statistics, and physics due to the abundance of equations in the related work section, and the crawling method theoremKb[1] lacked the required versatility. The exclusive emphasis on related work in IEEE format papers stems from the notion that each sentence in the related work section encapsulates the author's thought process in citing related works: **(A)** Every sentence captures the author's interpretation and emphasis on original methodology, critique of prior work, corrections to previous research, or acknowledgment of pioneers. This encompasses summarizing these aspects briefly and concisely. **(B)** The cited work in the related work section is either incidental or important to current work (Valenzuela et al., 2015). **REASONS** is inspired by previously constructed **s2ORC and UnarXive datasets** containing academic papers (see Table 4 in Appendix A); however, we diverge on the following points: **(A)** We provide sentence-level annotation of citations on major computational domains

[1]https://github.com/PierreSenellart/theoremkb

on arXiv. **(B)** Each sentence is accompanied by its metadata, which includes the paper title, abstract, and author names of the paper it cites. It also contains the title of the paper from which it was taken. **(C)** The dataset structure allows for an easy examination of LLMs using indirect and direct queries.

Crawling Process: The web crawler employs the Oxylabs[2] SERP Scraper API as its methodology, enabling real-time data extraction from major search engines. This API offers a proxy chaining platform for efficient data extraction. The dataset is meticulously organized in JSON format with a detailed outline (see "JSON Structure"). A complete GitHub repository is provided, containing the dataset and the code for reproducibility (see details in subsection A.3). We plan to keep updating the repository with more articles and metadata. The associated costs are provided in (subsection A.2).

JSON Structure


```
{"Computer Vision": {
  "http://arXiv.org/abs/2012.05435v2": {
   "Paper Title": "Optimization-Inspired..",
   "Sentences": [
    {"Sentence ID": 32,
      "Sentence": "... For GM, ... ",
      "Citation Text": "C. Ledig,...",
      "Citation": {
        "Citation Paper ID": "arXiv:1609.04802",
        "Citation Paper Title": "Title:Photo..",
        "Citation Paper Abstract": "Abstract:.",
        "Citation Paper Authors": "Authors:..." }}]}}}
```


3.1 Problem Formulation

We define two tasks for LLMs over the **REASONS** dataset $\mathbf{R}$: (a) Direct Querying and (b) Indirect Querying. For experimentation, we segment $\mathbf{R}$ into $\mathbf{R}_S$ and $\mathbf{R}_M$. $\mathbf{R}_S$ represents sentences and paper titles for which references are to be generated with or without the support from metadata $\mathbf{R}_M$.

Direct Querying Task: Given a title $t_i \in \mathbf{R}_S$, the LLM should generate the author list. For the task of direct querying with metadata, the LLM is given the following input: $t_i \in \mathbf{R}_S$, the Advance RAG model retrieves top-40 chunks of information $a_{i1}, ..., a_{i40} \in \mathbf{R}_M$, and generates the names.

Indirect Querying Task: Given a sentence $s_i \in \mathbf{R}_S$, the LLM should generate a paper title in zero-shot setting. For the task of indirect querying with metadata called *Sequential Indirect and Direct Prompting* (SID Prompting), the LLM is given the following input: $s_i \in \mathbf{R}_S$ and ground truth abstract $abs_s \in \mathbf{R}_M$ as well as the authors $au_s \in \mathbf{R}_M$, and the model is asked to generate the citation paper title.

[2] https://oxylabs.io/

Examples of direct and indirect queries are:

Direct Prompt

Prompt: Who were the authors of the research paper "Research Paper Title"?
Instruction: List only author names, formatted as $<firstname><lastname>$, separated by comma. Do not mention the paper in the title, also, if you don't know, write 'pass'.
Response: Author Names.

Indirect Prompt

Prompt: I have taken a sentence from the research paper titled "Research Paper Title", give me the research paper that this sentence is citing. If you cannot come up with the paper titles, write 'pass.' Don't write anything else.
Instruction: Sentence "uses fractional max-pooling to randomly specify non-integer ratios between the spatial dimension sizes of the input and the output to pooling layers."
Response: Citation Paper Title.

Implementation of Direct and Indirect Querying: Direct querying is executed using zero-shot prompting for scenarios without metadata and chain-of-thoughts prompting for metadata situations. We modify the chain-of-thoughts prompting with *SID Prompting*. It begins with an indirect query. Following an incorrect response or a 'pass,' more details about the cited paper are given (i.e., direct query), including its abstract and authors' names. This is an iterative approach to generate the correct citation. Following are the two examples of these prompting strategies:

Direct Query with Metadata Prompting

Prompt: Who were the authors of the research paper "Research Paper Title"? Let me give you some more context by providing the abstract of the research paper. Abstract:'....'.
Instruction: List only author names, formatted as ¡first name¿¡last name¿, separated by comma. Do not mention the paper in the title. Also, if you don't know, write 'pass.'
Response: Author Names.

SID Prompting

Prompt: I have taken a sentence from the research paper titled "Research Paper Title." give me the title of the possible research paper that this sentence is citing. If you cannot come up with the paper titles, write 'pass'. Don't write anything else.
Instruction: Sentence:"......". Let me give you some more context by providing the authors and the abstract of the paper the sentence is citing. Authors:"......", Abstract:"......"
Response: Citation Paper Title.

3.2 Models and Evaluation

Our research has focused on a diverse array of LLMs, carefully chosen to provide a broad perspective on the capabilities and limitations inherent in current language model technologies.

Proprietary Models: Our selection of proprietary models includes those from OpenAI and Preplexity.ai. While OpenAI is known for its cutting-edge NLP models, driving significant advancements in the field, Preplexity.ai focuses on models with unique functionalities, such as recommending citations and utilizing natural language prediction for innovative search experiences.

Public Models: We choose LLAMA 2 (Touvron et al., 2023) and Mistral (Jiang et al., 2023a) as the two publicly available LLMs that have demonstrated competitive performance compared to proprietary LLMs. We evaluate their effectiveness on the **REASONS** dataset under the standard state and retrieval-augmentation conditions. This analysis goes beyond simply comparing proprietary and public models, extending to evaluating models based on their size, particularly those with 7B parameters.

3.3 Evaluation Metrics

Our evaluation uses four key metrics: 1) The **BLEU Score** assesses the structural alignment through clipped n-gram matching. 2) The **F-1 Score** evaluates the balance between precision and recall, reflecting the models' effectiveness in capturing key information. 3) **Hallucination rate (HR)**, which we estimate by averaging over incorrect and partially correct generated citations. HR $= \frac{1}{Q_D} \sum \mathbb{I}[\hat{c} \neq c] + \frac{1}{|U_w|} \sum_{w=1}^{|U_w|} \mathbb{I}[\hat{c}_w \neq c_w]$, where Q_D: queries within a domain, and $|U_w|$: total number of unique words in generated citation ($\hat{c}$) and true citation (c). 4) **Pass Percentage (PP)** measures the tendency of an LLM to either respond or abstain from giving a response. It is calculated as follows: $\frac{1}{Q_D} \sum \mathbb{I}[\hat{c} = \text{Pass}]$. It is crucial to emphasize that PP serves as a safeguard to prevent LLMs from generating hallucinatory responses but also reduces engagement. Additionally, even with a high PP, the HR can be high. This implies that the model struggles to discern whether it offers correct or incorrect citations in the remaining instances.

3.4 Retrieval Augmented Generation (RAG)

RAG combines a retriever and a generator to create better answers. RAG can access external knowledge, unlike methods that feed the model prompts. This lets it craft more accurate, relevant, and informative responses than models that rely solely on what they were pre-trained.

We investigate RAG's ability to improve LLMs' accuracy. Ideally, RAG would help LLMs avoid giving wrong answers (low PP) and making things up (HR). We also investigate whether RAG works consistently with direct and indirect questions across different scientific fields (12 domains). We experiment with two forms of RAG architecture: (a) Naïve RAG and (b) Advance RAG. Both architectures leverage the same bi-encoder-based retriever architecture (Karpukhin et al., 2020).

Given a corpus of documents $\mathbf{R}_M$ and a sentence $s \in \mathbf{R}_S$, the document encoder maps $d \in \mathbf{R}_M$ to an embedding $\mathbf{E}_\theta(c)$ and the query encoder maps s to an embedding $\mathbf{E}_\theta(s)$. The top-k relevant documents for s are retrieved based on the sentence-document embedding similarity, which is often computed via dot product: $z(s, d) = exp(\mathbf{E}_\theta(s)^T \mathbf{E}_\theta(d))$. We start with a bi-encoder retriever using an embedding model from OpenAI (subsection A.4). Other ways to set up a bi-encoder retriever, such as DRAGON+ (Lin et al., 2023), are possible. However, those are more useful when involving large-scale data augmentation.

The retrieved documents are ranked in two ways, which separates Naïve RAG from Advance RAG. Under the Naïve RAG, we use BM25 relevance scoring to rank the documents, whereas, in Advance RAG, we fine-tune a cross-encoder on **REASONS** document index $\mathbf{R}_M$ to better align it with our task of citation generation with LLM. For the fine-tuning of the cross-encoder, we use localized contrastive loss (LCL) for two reasons: (a) In $\mathbf{R}_M$, we do not have labeled positive and negative documents, and (b) for a sentence s there is a possibility for more than one true positive documents (Pradeep et al., 2022). LCL is formally defined as follows:

$$\mathcal{L}_{LCL_s} := -\log \frac{exp(z_{s,\{d^+\}})}{\sum_{d \in G_s} exp(z_{s,d})}$$

$$\mathcal{L}_{LCL} := \frac{1}{|S|} \sum_{s \in \mathbf{R}_s, G_s \in \mathbf{R}_M^s} \mathcal{L}_{LCL_s}$$

where G_s represents a set of documents for a sentence s, which consist of a set of relevant documents ($\{d^+\}$) and n-1 non-relevant documents $\{d^-\}$ sampled from $\mathbf{R}_M^s$ using biencoder. The training of Advance RAG happens through the standard cross entropy loss: $\mathcal{L}_{CE}(\hat{c}|s, \phi) =$

$\sum_{i=1}^{b} \mathbb{I}(\hat{c}_i^w = c_i^w) \cdot \log Pr(\hat{c}_i^w|\phi)$ where, ϕ is parameter of the generator LLM and b is the mini-batch fine-tuning in Advance RAG. $\hat{c}_i$ represents i^{th} citation generation, and $\mathbb{I}(\hat{c}_i^w = c_i^w)$ represents word level comparison with ground truth citation (direct query: author names; indirect query: paper titles). For the Naïve and Advance RAG, we employ LLAMA-2 7B and Mistral 7B as competitive models against proprietary LLMs.

4 Results

We conducted experiments encompassing four distinct prompting styles applied to twelve scientific domains. This extensive analysis involved 12,723 sentences, resulting in a substantial dataset rigorously evaluated using ten different models. This equates to **508920 instance assessments** involving 4 (prompting styles) × 12,723 (sentences for all domains) × 10 (models). The time associated with performing these experiments is given in the appendix (subsection A.6 and Table 5).

Zero-Shot Indirect Prompting: In Figure 4, a majority of the models exhibited high HR. As expected for a huge model `GPT-4-1106-preview` (1 Trillion Parameters) shows a relatively lower HR of 67.73% and a higher PP of 89% averaged across 12 domains. Perplexity-7b-Chat showed an exceptionally high PP of 97.5%, which is surprising, as this LLM is designed specifically for citation generation. RAG Mistral was a competitive model with GPT-4 with a lower PP of 21% and HR of 72.49% in comparison to other LLMs. Analysis shows RAG Mistral is competitive because of the high variance in HR compared to `GPT-4-1106-preview`. Generation quality measured by F-1 and BLEU scores were predominantly low across the board, with GPT-4 (not the preview, G1) comparatively better scores. RAG Mistral and RAG LLAMA 2 rank second and third best respectively.

SID Prompting In Figure 5, showed improvement across all the LLMs in citation generation over indirect queries. An average improvement of 21% was measured, with a reduction in variance. Even though some models like Perplexity-7b-Chat and LLAMA 2 still had high HR rates, the PP dropped significantly, especially for `GPT-4-1106-preview`. The results of this experiment indicate that SID prompting in LLMs can balance the trade-off between PP and HR, significantly enhancing generation quality with an (8%↑) increase in BLEU and a (13%↑) in F-1 (The Appendix B provides examples for visual inspection.).

Zero-Shot Direct Prompting presents a very idealistic scenario where the LLMs have access to context through direct query. This leads to both lower PP and HR. The citation generation quality is great, with high F-1 and BLEU scores (see Figure Figure 4). However, Perplexity-7b-Chat, oddly, had high PP and HR, suggesting a need for more research on such specialized LLM search engines. We observed that Perplexity-7b-Chat expands its search queries and adds references to the broader content it finds. The issue is that the expanded versions drift too far in meaning from the original.

In **Direct Prompting with Metadata**, when metadata such as abstracts and titles were used with indirect questions, all the LLMs got better at generating citations and had low HR and PP. This shows that having more information helps LLMs create more accurate and related citations, proving the importance of enough data for good language processing. Note that PP dropped to zero for almost all models when direct promoting includes metadata. All GPT LLMs achieved F-1 and BLEU scores close to 1.0 and showed more consistent results overall. Two main points from this experiment are: First, *adding metadata* to LLMs is effective for all of them, especially RAG models that integrate this augmentation in their learning process. Second, *smaller models with advance RAG (Mistral and LLAMA-2) adjust better to metadata* than GPT-4-Preview/4/3.5 (see Figure 3).

Overall: *Advance RAG Mistral 7b* outperformed other competitive proprietary and public LLMs in all prompting styles. This superior performance was notably marked by reduced HR, suggesting this model is more adept at generating accurate and relevant responses when adding metadata. Furthermore, improvements in F-1 scores reinforce its reliability in retrieving information. Higher BLEU scores were observed, signifying that the language output of the model aligns closely with human-like text in terms of fluency & coherence.

5 Adversarial Examination

The analysis of LLMs using the **REASONS** dataset highlights significant variability in their performance across different domains. While they perform moderately better in areas like AI and CV with lower HR and higher F-1/BLEU scores, they struggle in complex domains such as QC, Biomolecules, and Cryptography, likely due to limited training data and the complexity of these sub-

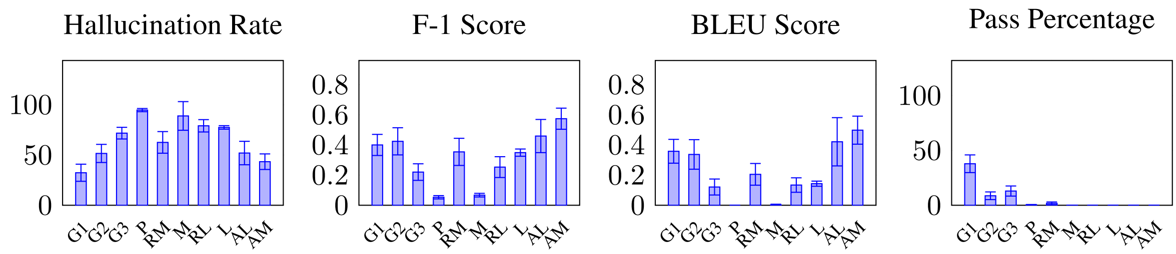

Figure 2: Averaged **Zero-Shot Direct Prompting** results of different LLMs across all 12 domains. G1 shows notably lower HR and higher F-1 and BLEU scores, indicating superior performance in generating citations. In contrast, model P exhibits the highest HR and the lowest scores in F-1 and BLEU, suggesting challenges in generating accurate and contextually relevant citations. The RAG models (RM and RL) demonstrate varied results, with RM showing a better accuracy and coherence balance than RL. **G1:** `gpt-4-1106-preview`, **G2:** `gpt-4`, **G3:** `gpt-3.5-turbo`, **P:** `pplx-7b-chat`, **RM:** `Naïve RAG mistral-7b-instruct`, **M:** `mistral-7b-instruct`, **RL:** `Naïve RAG llama-2-7b-chat`, **L:** `llama-2-7b-chat`, **AL:** `Advance RAG llama-2-7b-chat`, **AM:** `Advance RAG mistral-7b-instruct`

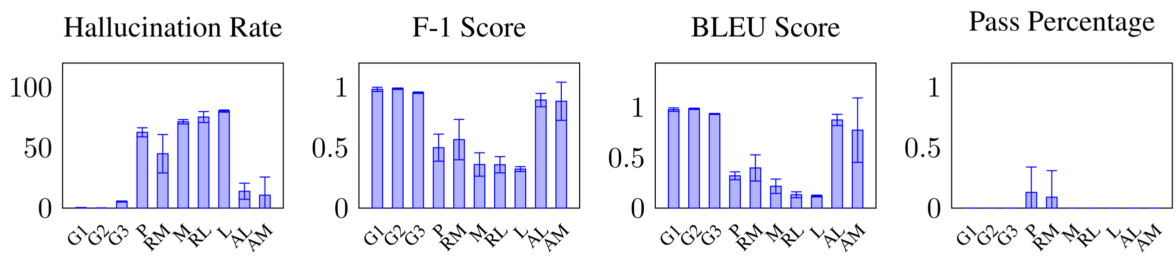

Figure 3: Averaged **Direct Prompting with Metadata** results of different LLMs across all 12 domains. The plot indicates that models G1, G2, and G3 stand out with their low HR and impressive F-1 and BLEU scores, in contrast to other models that face challenges. All models except RM reach a 0% PP, suggesting that including metadata significantly enhances their contextual understanding.

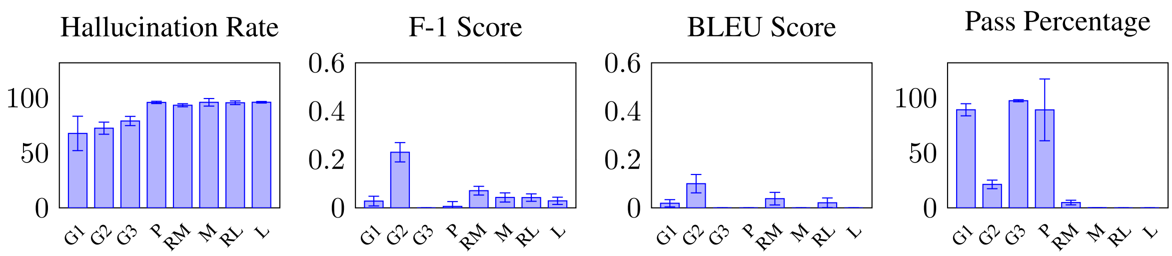

Figure 4: Averaged **Zero-Shot Indirect Prompting** across 12 domains. This prompting method led to elevated HR among the models. There was also a notable variance in PP, with models G3, P, and L exhibiting higher scores. Both conditions indicate challenges in understanding context and generating accurate citations when using indirect prompts.

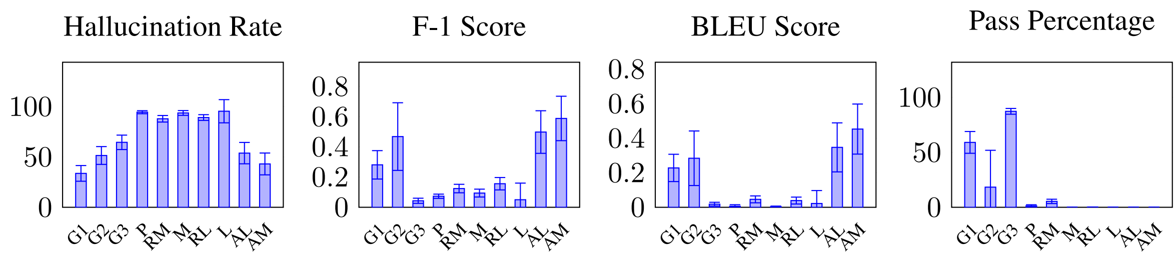

Figure 5: Averaged **SID Prompting** results of different LLMs across all 12 domains. Models G1, G2, and G3 exhibit relatively better outcomes with lower HR and higher F-1 and BLEU scores, suggesting more contextual understanding. Other models demonstrated high HR, indicating difficulties in accurate citation generation with SID Prompting. Notably, while models G1 and G3 have high PPs, indicating some difficulties with SID, their overall performance still reflects a more advanced level of language processing and contextual comprehension compared to the other models.

jects. This variability in performance indicates that LLMs have varying degrees of contextual understanding, with a tendency to perform better in domains with more extensive training data and less complex structures (e.g., maths and numerics).

Motivation and Setup: We conducted adversarial experiments across all models to better assess their contextual understanding. The core concept

Group	PP(%)	BLEU	F1	HR
Changing Paper Title				
G1	96.23	0.6210	0.8470	17.99
G2	31.45	0.0524	0.2640	83.66
G3	68.55	0.0389	0.1828	87.35
RM	3.14	0.0796	0.1584	86.78
M	0.00	0.0003	0.0221	94.95
RL	5.03	0.0628	0.1448	87.56
L	0.00	0.0066	0.0254	98.30
AdvRAG(L)	0.00	0.1322	0.4763	85.72
AdvRAG(M)	0.00	0.1569	0.5839	75.41
Changing Paper Abstract				
G1	95.60	0.4595	0.6451	38.49
G2	32.70	0.0396	0.2186	86.22
G3	76.10	0.0034	0.1013	91.64
RM	7.55	0.0520	0.1216	89.44
M	0.00	0.0074	0.0161	90.20
RL	2.52	0.0445	0.1112	90.16
L	0.00	0.0017	0.0146	99.01
AdvRAG(L)	0.00	0.4101	0.5780	39.67
AdvRAG(M)	0.00	0.4904	0.6954	39.57

Table 2: Summary of Adversarial Analysis Results Across Different Evaluation Metrics

behind these experiments was to provide the models with incorrect yet similar metadata about the sentences in the prompts. The aim was to discern whether the models generated citations based on the contextual grasp of the provided metadata or if the metadata had minimal influence on the citation generation process. These adversarial experiments comprised two types: 1) Providing *inaccurate paper titles* related to the sentences. 2) Providing *incorrect paper abstracts* associated with the sentences. Both experiments were conducted using the SID prompting.

To facilitate these experiments, we curated a subsample of 200 sentences from the **REASONS** dataset spanning all the domains. We extracted each sentence's most similar paper title or abstract from this dataset and replaced the original metadata. For similarity calculation, we use the *Ratcliff-Obershelp* metric, which is calculated as twice the length of the longest common substring plus recursively the number of matching characters in the non-matching regions on both sides of the longest common substring (Tang et al., 2023). According to this metric, for the following example title "Diffusion models for counterfactual explanations," the best replacement is "Octet: Object-aware models for counterfactual explanations (0.736)" as opposed to "Adversarial counterfactual visual explanations (0.638)". We considered a threshold of 0.70 effective in preparing the adversarial set.

Findings: We found that incorrect paper titles and abstracts easily fool most LLMs if it is similar to accurate information. This means the LLMs are not very good at understanding the true meaning of what they are given. On such a small adversarial set, we expect LLMs like `GPT-4-1106-preview` and GPT-4 to perform exceedingly well because of their extensive knowledge; however, we observed counter-intuitive results in Table 2. We do see promising direction with AdvRAG(M) and AdvRAG(L); however, further investigation is required into how rich graphical metadata (e.g., knowledge graph) and graph-theoretic approaches to information retrieval can improve LLM effectiveness (He et al., 2024).

6 Conclusion

We have developed a new resource called **REASONS** (**RE**trieval and **A**utomated citation**S** **O**f scie**N**tific Sentences), a benchmark designed to assess the ability of LLMs to understand context and generate appropriate citations. This benchmark includes sentences from the related work sections of papers, along with citations and metadata across 12 scientific and computational fields. We evaluated proprietary and public LLMs' ability to correctly provide author names and paper titles under two conditions: direct and indirect citation. Surprisingly, none of the LLMs demonstrated the readiness to annotate draft reports in various professional settings, such as market analysis, misinformation prevention, defense strategy, and healthcare reporting. We observed a trade-off between PP and HR, where GPT-4 and GPT-3.5 achieved higher accuracy at the cost of a lower HR. In contrast, though smaller with only 7B parameters, the Advance RAG model showed reasonable efficiency. Unlike other models, in adversarial tests where abstracts or paper titles were swapped, Advance RAG unexpectedly outperformed GPT-4, suggesting it does capture context before generating citations.

Future Work: Through reasoning and explanation, we plan to explore and mitigate the noted shortcomings in citation generation (trade-off between HR and PP, high variance in BLEU scores, sub-par scores on adversarial set). One approach is to employ the Toulmin model (Naveed et al., 2018)) within Advance RAG. We believe these improvements will improve the quality of citation generation and better equip the models to manage complex reasoning (e.g., hypothesis generation and verification (Tyagin and Safro, 2023)) challenges confidently.

Limitations

Several factors constrain our study on applying LLMs for citation generation. **(a)** Primarily, integrating high-parameter-size models (>13B; refer to Table 5 for computation time) with RAG is not feasible, limiting our ability to leverage more complex models. **(b)** Additionally, the high computational resources required for such models are often inaccessible in academic settings. **(c)** One constraint in our study was the dataset creation, where we confined ourselves to predominantly IEEE format papers, particularly with domains with a high count of submissions. **(d)** Another significant limitation is the current inability of LLMs to effectively process and interpret mathematical expressions, a crucial aspect in many academic papers. **(e)** Due to the latest version of Google API (time stamp: December 04, 2023) lacking the citation generation feature, we have limited our experiments to OpenAI only. **(f)** While cross-encoders can be more powerful in understanding text relationships, they tend to be more computationally intensive. This is because they need to process every possible pair of inputs together, which can be a significant workload, especially in cases where there are many potential pairs to consider (like in large-scale retrieval tasks in our **REASONS** dataset). These constraints highlight the need for advancements in model adaptability, computational resource accessibility, dataset diversity, and specialized content processing for more robust and wide-ranging applications.

Ethical Considerations

We followed the Oxylabs Acceptable Use Policy[3] and worked alongside some Oxylabs developers to ensure we respected the terms of services on arXiv. arXiv's terms of service place restrictions on automated crawling of their site for articles marked by "arxiv.org perpetual, non-exclusive license and CC BY-NC-ND". We paid attention to the following key ethical issues: **(a) Privacy and Consent:** The content on arXiv is publicly available, but the authors who upload their work there may not have consented to having their preprints crawled and used for other purposes. It's important to respect the privacy and intellectual property rights of the researchers who contribute to arXiv. We only crawled articles marked as CC Zero, CC BY, and CC BY-SA. **(b) Potential misuse:** We prepared **REASONS** only to test the citation generation capability of LLMs for subsequent future downstream applications, such as annotating draft analytic reports. Our focus on HR and PP for citation generation and its quality using BLEU and F-1 shows that the data scraped is not for malicious purposes, such as fine-tuning LLMs to generate misinformation or infringe on copyrights. **(c) Transparency and Accountability:** We have been mindful of our crawling process, and to the best of our knowledge, we have enumerated sufficient details regarding the process. This would help build trust regarding reproducibility, extend **REASONS**, and ensure that the crawling process was not abused. **(d) Author Identity and Contact:** No authors of the crawled papers were contacted through their provided information in the publicly available arXiv papers. This user study was duly approved by the authors' organization's Institutional Review Board (IRB).

[3] https://oxylabs.io/legal/oxylabs-acceptable-use-policy

References

arXiv submission rate statistics 2021- arXiv info — info.arxiv.org. https://info.arxiv.org/help/stats/2021_by_area/index.html. [Accessed 16-04-2024].

Nitin Agarwal, Ehtesham Haque, Huan Liu, and Lance Parsons. 2005. Research paper recommender systems: A subspace clustering approach. In *Advances in Web-Age Information Management: 6th International Conference, WAIM 2005, Hangzhou, China, October 11–13, 2005. Proceedings 6*, pages 475–491. Springer.

Zafar Ali, Guilin Qi, Pavlos Kefalas, Waheed Ahmad Abro, and Bahadar Ali. 2020. A graph-based taxonomy of citation recommendation models. *Artificial Intelligence Review*, 53:5217–5260.

Zafar Ali, Irfan Ullah, Amin Khan, Asim Ullah Jan, and Khan Muhammad. 2021. An overview and evaluation of citation recommendation models. *Scientometrics*, 126:4083–4119.

Isabelle Augenstein, Timothy Baldwin, Meeyoung Cha, Tanmoy Chakraborty, Giovanni Luca Ciampaglia, David Corney, Renee DiResta, Emilio Ferrara, Scott Hale, Alon Halevy, et al. 2023. Factuality challenges in the era of large language models. *arXiv preprint arXiv:2310.05189*.

Xiaomei Bai, Mengyang Wang, Ivan Lee, Zhuo Yang, Xiangjie Kong, and Feng Xia. 2019. Scientific paper recommendation: A survey. *Ieee Access*, 7:9324–9339.

Joeran Beel, Bela Gipp, Stefan Langer, and Corinna Breitinger. 2016. Paper recommender systems: a literature survey. *International Journal on Digital Libraries*, 17:305–338.

Steven Bethard and Dan Jurafsky. 2010. In *Who Should I Cite? Learning Literature Search Models from Citation Behavior ABSTRACT*, pages 609–618. [link].

Chandra Bhagavatula, Sergey Feldman, Russell Power, and Waleed Ammar. 2018. Content-based citation recommendation. *arXiv preprint arXiv:1802.08301*.

Anubrata Bhowmick, Ashish Singhal, and Shenghui Wang. 2021. Augmenting context-aware citation recommendations with citation and co-authorship history. In *18th International Conference on Scientometrics and Informetrics, ISSI 2021*, pages 115–120. International Society for Scientometrics and Informetrics.

Molly Bohannon. 2023. Lawyer used chatgpt in court and cited fake cases, a judge is considering sanctions. *Forbes*.

Bernd Bohnet, Vinh Q Tran, Pat Verga, Roee Aharoni, Daniel Andor, Livio Baldini Soares, Massimiliano Ciaramita, Jacob Eisenstein, Kuzman Ganchev, Jonathan Herzig, et al. 2022. Attributed question answering: Evaluation and modeling for attributed large language models. *arXiv preprint arXiv:2212.08037*.

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George Bm Van Den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, et al. 2022. Improving language models by retrieving from trillions of tokens. In *International conference on machine learning*, pages 2206–2240. PMLR.

Tanmoy Chakraborty, Natwar Modani, Ramasuri Narayanam, and Seema Nagar. 2015. Discern: a diversified citation recommendation system for scientific queries. In *2015 IEEE 31st international conference on data engineering*, pages 555–566. IEEE.

Blaise Cronin. 1981. The need for a theory of citing. *Journal of documentation*, 37(1):16–24.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Travis Ebesu and Yi Fang. 2017. Neural citation network for context-aware citation recommendation. In *Proceedings of the 40th international ACM SIGIR conference on research and development in information retrieval*, pages 1093–1096.

Ran Elgedawy, Sudarshan Srinivasan, and Ioana Danciu. 2024. Dynamic Q&A of Clinical Documents with Large Language Models. *arXiv preprint arXiv:2401.10733*.

Xiao Fang, Shangkun Che, Minjia Mao, Hongzhe Zhang, Ming Zhao, and Xiaohang Zhao. 2023. Bias of AI-generated content: an examination of news produced by large language models. *arXiv preprint arXiv:2309.09825*.

Luyu Gao, Zhuyun Dai, Panupong Pasupat, Anthony Chen, Arun Tejasvi Chaganty, Yicheng Fan, Vincent Zhao, Ni Lao, Hongrae Lee, Da-Cheng Juan, et al. 2023. Rarr: Researching and revising what language models say, using language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16477–16508.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. 2020. Retrieval augmented language model pre-training. In *International conference on machine learning*, pages 3929–3938. PMLR.

Qi He, Jian Pei, Daniel Kifer, Prasenjit Mitra, and Lee Giles. 2010. Context-aware citation recommendation. In *Proceedings of the 19th international conference on World wide web*, pages 421–430.

Xiaoxin He, Yijun Tian, Yifei Sun, Nitesh V Chawla, Thomas Laurent, Yann LeCun, Xavier Bresson, and Bryan Hooi. 2024. G-retriever: Retrieval-augmented generation for textual graph understanding and question answering. *arXiv preprint arXiv:2402.07630*.

Or Honovich, Roee Aharoni, Jonathan Herzig, Hagai Taitelbaum, Doron Kukliansy, Vered Cohen, Thomas Scialom, Idan Szpektor, Avinatan Hassidim, and Yossi Matias. 2022. TRUE: Re-evaluating factual consistency evaluation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3905–3920, Seattle, United States. Association for Computational Linguistics.

Zihan Huang, Charles Low, Mengqiu Teng, Hongyi Zhang, Daniel E Ho, Mark S Krass, and Matthias Grabmair. 2021. Context-aware legal citation recommendation using deep learning. In *Proceedings of the eighteenth international conference on artificial intelligence and law*, pages 79–88.

Chia-Chien Hung, Wiem Ben Rim, Lindsay Frost, Lars Bruckner, and Carolin Lawrence. 2023. Walking a tightrope–evaluating large language models in high-risk domains. *arXiv preprint arXiv:2311.14966*.

Sehrish Iqbal, Saeed-Ul Hassan, Naif Radi Aljohani, Salem Alelyani, Raheel Nawaz, and Lutz Bornmann. 2020. A decade of in-text citation analysis based on natural language processing and machine learning techniques: an overview of empirical studies. *Scientometrics*, 126:6551 – 6599.

Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. 2022. Few-shot learning with retrieval augmented language models. *arXiv preprint arXiv:2208.03299*.

Chanwoo Jeong, Sion Jang, Eunjeong Park, and Sungchul Choi. 2020a. A context-aware citation recommendation model with bert and graph convolutional networks. *Scientometrics*, 124:1907–1922.

Chanwoo Jeong, Sion Jang, Eunjeong Park, and Sungchul Choi. 2020b. A context-aware citation recommendation model with bert and graph convolutional networks. *Scientometrics*, 124.

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.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023a. Mistral 7b.

Zhengbao Jiang, Frank F Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023b. Active Retrieval Augmented Generation. *arXiv preprint arXiv:2305.06983*.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.

Urvashi Khandelwal, Omer Levy, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. 2019. Generalization through memorization: Nearest neighbor language models. *arXiv preprint arXiv:1911.00172*.

Tharindu Kumarage and Huan Liu. 2023. Neural authorship attribution: Stylometric analysis on large language models. In *2023 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC)*, pages 51–54. IEEE.

Onur Küçüktunç, Erik Saule, Kamer Kaya, and Umit Catalyurek. 2012. Diversifying citation recommendations. *ACM Transactions on Intelligent Systems and Technology*, 5.

Sheng-Chieh Lin, Akari Asai, Minghan Li, Barlas Oguz, Jimmy Lin, Yashar Mehdad, Wen-tau Yih, and Xilun Chen. 2023. How to train your DRAGON: Diverse augmentation towards generalizable dense retrieval. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6385–6400.

Nelson F Liu, Tianyi Zhang, and Percy Liang. 2023. Evaluating verifiability in generative search engines. *arXiv preprint arXiv:2304.09848*.

Yusuf Mehdi. 2024. Confirmed: the new Bing runs on OpenAI's GPT-4 — blogs.bing.com. `https://blogs.bing.com/search/march_2023/Confirmed-the-new-Bing-runs-on-OpenAI%E2%80%99s-GPT-4`. [Accessed 12-04-2024].

Jacob Menick, Maja Trebacz, Vladimir Mikulik, John Aslanides, Francis Song, Martin Chadwick, Mia Glaese, Susannah Young, Lucy Campbell-Gillingham, Geoffrey Irving, et al. 2022. Teaching language models to support answers with verified quotes. *arXiv preprint arXiv:2203.11147*.

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Ouyang Long, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, Xu Jiang, Karl Cobbe, Tyna Eloundou, Gretchen Krueger, Kevin Button, Matthew Knight, Benjamin Chess, and John Schulman. 2021. Webgpt: Browser-assisted question-answering with human feedback. *ArXiv*, abs/2112.09332.

Sidra Naveed, Tim Donkers, and Jürgen Ziegler. 2018. Argumentation-based explanations in recommender systems: Conceptual framework and empirical results. In *Adjunct Publication of the 26th Conference on User Modeling, Adaptation and Personalization*, pages 293–298.

Ronak Pradeep, Yuqi Liu, Xinyu Zhang, Yilin Li, Andrew Yates, and Jimmy Lin. 2022. Squeezing water from a stone: A bag of tricks for further improving cross-encoder effectiveness for reranking. In *European Conference on Information Retrieval*, pages 655–670. Springer.

Hannah Rashkin, Vitaly Nikolaev, Matthew Lamm, Lora Aroyo, Michael Collins, Dipanjan Das, Slav Petrov, Gaurav Singh Tomar, Iulia Turc, and David Reitter. 2023. Measuring attribution in natural language generation models. *Computational Linguistics*, 49(4):777–840.

Zafaryab Rasool, Scott Barnett, Stefanus Kurniawan, Sherwin Balugo, Rajesh Vasa, Courtney Chesser, and Alex Bahar-Fuchs. 2023. Evaluating llms on document-based qa: Exact answer selection and numerical extraction using cogtale dataset. *ArXiv*, abs/2311.07878.

Vipula Rawte, Swagata Chakraborty, Agnibh Pathak, Anubhav Sarkar, SM Towhidul Islam Tonmoy, Aman Chadha, Amit Sheth, and Amitava Das. 2023. The troubling emergence of hallucination in large language models-an extensive definition, quantification, and prescriptive remediations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2541–2573.

Kevin Roose. 2024. Can This A.I.-Powered Search Engine Replace Google? It Has for Me. — nytimes.com. `https://www.nytimes.com/2024/02/01/technology/perplexity-search-ai-google.html`. [Accessed 12-04-2024].

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023. Toolformer: Language models can teach themselves to use tools. *arXiv preprint arXiv:2302.04761*.

Leo Schwinn, David Dobre, Stephan Günnemann, and Gauthier Gidel. 2023. Adversarial attacks and defenses in large language models: Old and new threats. *arXiv preprint arXiv:2310.19737*.

Trevor Strohman, W. Croft, and David Jensen. 2007. Recommending citations for academic papers. pages 705–706.

Alex Tamkin and Deep Ganguli. 2021. How large language models will transform science, society, and ai.

Xiangru Tang, Yiming Zong, Yilun Zhao, Arman Cohan, and Mark Gerstein. 2023. Struc-bench: Are large language models really good at generating complex structured data? *arXiv preprint arXiv:2309.08963*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.

Hung Nghiep Tran, Tin Huynh, and Kiem Hoang. 2015. A potential approach to overcome data limitation in scientific publication recommendation. In *2015 Seventh International Conference on Knowledge and Systems Engineering (KSE)*, pages 310–313. IEEE.

Ilya Tyagin and Ilya Safro. 2023. Dyport: Dynamic importance-based hypothesis generation benchmarking technique. *arXiv preprint arXiv:2312.03303*.

Peter Vajdecka, Elena Callegari, Desara Xhura, and Atli Ásmundsson. 2023. Predicting the presence of inline citations in academic text using binary classification. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 717–722.

Marco Valenzuela, Vu Ha, and Oren Etzioni. 2015. Identifying meaningful citations. In *AAAI workshop: Scholarly big data*, volume 15, page 13.

Wei Wang, Tao Tang, Feng Xia, Zhiguo Gong, Zhikui Chen, and Huan Liu. 2020. Collaborative filtering with network representation learning for citation recommendation. *IEEE Transactions on Big Data*, 8(5):1233–1246.

Muhammad Yamin, Ehtesham Hashmi, Mohib Ullah, and Basel Katt. 2024. Applications of LLMs for generating cyber security exercise scenarios.

Lixiang Yan, Lele Sha, Linxuan Zhao, Yuheng Li, Roberto Martinez-Maldonado, Guanliang Chen, Xinyu Li, Yueqiao Jin, and Dragan Gašević. 2024. Practical and ethical challenges of large language models in education: A systematic scoping review. *British Journal of Educational Technology*, 55(1):90–112.

Libin Yang, Yu Zheng, Xiaoyan Cai, Hang Dai, Dejun Mu, Lantian Guo, and Tao Dai. 2018. A lstm based model for personalized context-aware citation recommendation. *IEEE access*, 6:59618–59627.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2022. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*.

Xiang Yue, Boshi Wang, Kai Zhang, Ziru Chen, Yu Su, and Huan Sun. 2023. Automatic evaluation of attribution by large language models. *arXiv preprint arXiv:2305.06311*.

Fattane Zarrinkalam and Mohsen Kahani. 2013. Semcir: A citation recommendation system based on a novel semantic distance measure. *Program: electronic library information systems*, 47.

Yilun Zhao, Haowei Zhang, Shengyun Si, Linyong Nan, Xiangru Tang, and Arman Cohan. 2023. Investigating table-to-text generation capabilities of large language models in real-world information seeking scenarios. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 160–175, Singapore. Association for Computational Linguistics.

Shen Zheng, Jie Huang, and Kevin Chen-Chuan Chang. 2023. Why does chatgpt fall short in answering questions faithfully? *arXiv preprint arXiv:2304.10513*.

A Appendix

A.1 The Story of a Lawyer who employed ChatGPT

In Figure 6, the reliance on LLM-generated content by legal professionals, highlighted by The New York Times, illuminates the pitfalls when these LLMs produce content that lacks proper verification. This incident not only signifies the importance of cross-checking LLM outputs against reliable sources but also exemplifies the potential repercussions of neglecting this critical step. The subsequent requirement for the involved attorney to issue apologies and accept sanctions demonstrates

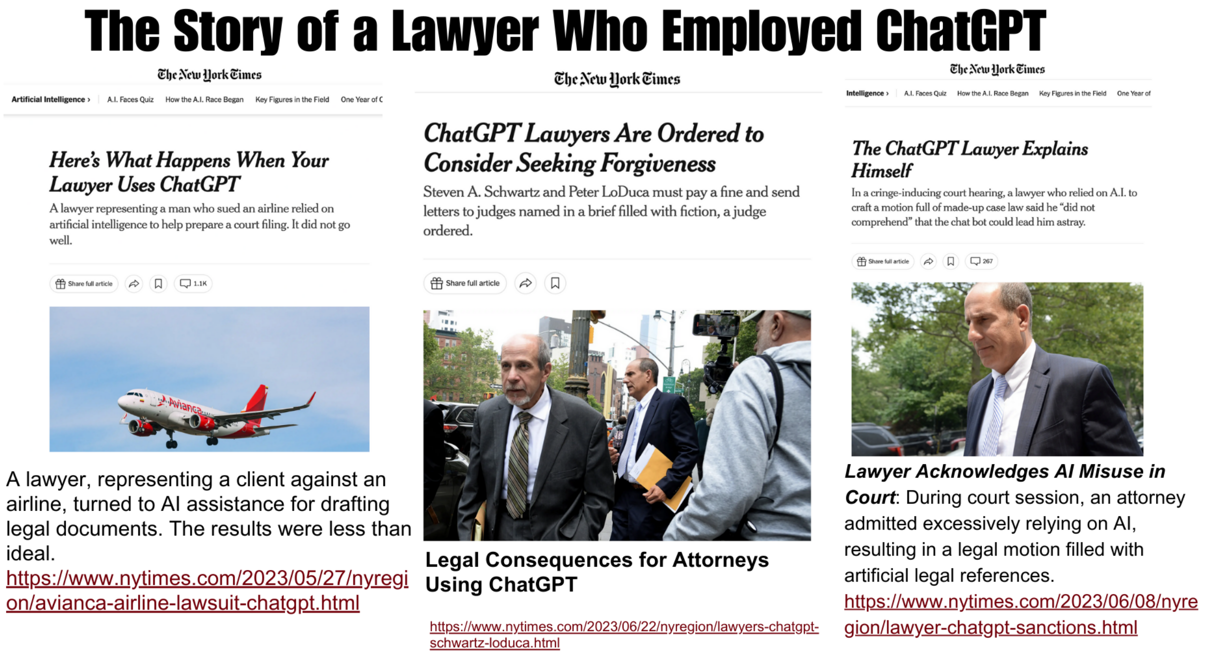

Figure 6: The perils of inadequate verification of LLMs-generated citations in legal documents.

the dire need for robust citation practices in the deployment of LLMs and serves as a crucial learning point for all sectors considering the integration of LLMs into their workflow. Links to the New York Times news articles covering the whole story:

- https://www.nytimes.com/2023/05/27/nyregion/avianca-airline-lawsuit-chatgpt.html,
- https://www.nytimes.com/2023/06/22/nyregion/lawyers-chatgpt-schwartz-loduca.html
- https://www.nytimes.com/2023/06/08/nyregion/lawyer-chatgpt-sanctions.html

A.2 Research Cost Breakdown

The cost associated with this research includes expenses for utilizing OpenAI API, totaling $640.37. Additionally, the use of Perplexity API incurred costs amounting to $259.39. Furthermore, GPU resources, we used Replicate[4] API for our experiments, amounted to $466.22. For dataset creation, we used Oxylab for $249 for a month. In total, the expenses for conducting this research sum up to $1614.98.

[4] https://replicate.com/

A.3 Reproduciblity

Our pipeline is straightforward to implement and can be easily reproduced. We have thoroughly documented all experimental details in the main text and the appendices. Although the full text of each prompt is too lengthy to include, we offer examples of each in Appendix B to help readers understand the style used. *All of our resources, including complete prompt scripts, crawling data, and code for evaluating our approach, are available to the* ***public repository here***:

- https://anonymous.4open.science/r/REASONS_BENCHMARK-D04D/README.md

A.4 Models specifications used during experimentation

The 'temperature' hyper-parameter in the LLMs controls the creativity of the LLMs in their response. The lower the temperature, the lower the creativity in the response, and the higher the temperature value, the higher the creativity in the response. By default, the temperature for most of the LLMs is set to 1. The 'max_tokens' describes the maximum number of tokens the LLM can generate. The 'top_p' is nucleus sampling, which helps limit the irrelevant tokens in the generation.

The 'top_k' is the number of retrieved chunks of information that will be considered during the generation in the RAG process. The 'tokenizer' converts the retrieved chunks of information and the prompts into tokens.

We have used two different tokenizers 'NousResearch/Llama-2-7b-chat-hf' [5] for LLAMA-2-7b-chat and 'mistralai/Mistral-7B-v0.1' [6] for Mistral-7b-instruct. The "Embedding Model" generates embeddings for tokens produced during tokenization. We have utilized the 'BAAI/bge-small-en-v1.5'[7] model for this purpose. And finally, the Cross-Encoder 'ms-marco-MiniLM-L-12-v2'[8] is fine-tuned using the LCL function for re-ranking of the retrieved chunks.

Our research utilized a dual-configuration server setup provided by the University. Configuration 1 consists of two nodes, with each node housing 128 cores (totaling 256 cores), 256GB of RAM, and two NVIDIA L40S GPUs, each equipped with 48GB of GPU memory. Configuration 2 is equipped with 8 NVIDIA A100-40GB cards, 1TB of RAM, and 256 CPUs. Due to resource availability in the queue, we alternate between these two configurations. Currently, we have not been able to compare their performance.

We concluded that the Zero Shot Indirect prompting approach is susceptible to hallucinations and is ineffective for the citation generation task. Hence, we did not conduct Advance RAG experiments with this prompting due to earlier results from other models, and also, the Advance RAG approach is computationally more expensive Table 6.

Hyperparameter	Value
temperature	1.0
max_tokens	256
top_p	0.95
Naïve RAG	
top_k	2
Embedding Model	BAAI/bge-small-en-v1.5
Advance RAG	
top_k	40
Cross-Encoder	ms-marco-MiniLM-L-12-v2
LLAMA-2 Tokenizer	NousResearch/Llama-2-7b-chat-hf
Mistral Tokenizer	mistralai/Mistral-7B-v0.1

Table 3: Hyper-parameters along with their values used during experimentation

[5] https://huggingface.co/NousResearch/Llama-2-7b-chat-hf

[6] https://huggingface.co/mistralai/Mistral-7B-v0.1

[7] https://huggingface.co/BAAI/bge-small-en-v1.5

[8] https://huggingface.co/cross-encoder/ms-marco-MiniLM-L-12-v2

A.5 Dataset Comparison

We contrast the **REASONS** dataset with other similar datasets that could have been utilized for citation generation. However, due to constraints within these datasets—such as the absence of sentence-level annotation of citations, metadata of citations, and paper titles—we would not be able to effectively assess the ability of LLMs and RAG LLMs to accurately grasp the context and generate suitable citations (see Table 4). Acronyms used in the paper: Computer Vision (CV), Information Retrieval (IR), Artificial Intelligence (AI) Natural Language Processing (NLP), Cryptography (Crypto), Neurons and Cognition (NNC), Human-Computer Interaction (HCI), Quantum Computing (QC), and Biomolecules.

A.6 GPU Machine Hours

With the exception of direct prompting, all other prompting styles required a substantial number of GPU hours (see Table 5). Training Advance RAG proved to be a highly time-intensive endeavor, which we attempted to mitigate by alternating between NVIDIA L40S and A100. We also found that LLAMA 2 required less time in training than Mistral. The reasons behind this can be a subject of future work. We provide machine-hour estimates to assist other researchers interested in RAG and its applications in provenance and context comprehension, facilitating better time management.

B Examples of Prompts in Direct and Indirect Queries

In the following visual examples, each model is followed by a checkbox indicating whether it generated citations correctly or incorrectly. See Figure 7, Figure 8, Figure 9, Figure 10, Figure 11, Figure 12, Figure 13.

B.1 Individual Results of all the domains across all the prompting styles

A comparative analysis of hallucination rates (HR) across several LLMs in **zero-shot indirect prompting** reveals distinct patterns, focusing on common domains. The **G1**, **G2**, **G3**, **P**, **RM**, **M**, **RL**, and **L** models consistently show variations in HR. High HR domains like **NNC**, **Cryptography**, and **NLP** appear recurrently across several models.

Low HR results frequently occur in **IR**, **CV**, and **HCI**, indicating a general resilience in these areas across different settings. For instance, **NNC** features prominently with high HR in the **G1**, **G2**, **G3**, **RM**, and **RL** models, while **IR** and **CV** consistently show low HR across **G1**, **G2**, **RM**, and **M**

	REASONS	UnarXive	PubMed	CiteULike	S2orc
Main Purpose	Sentence Annotation	Citation Recommendation	Medical Research	Benchmark for Recommendation Systems and Collaborative Filtering Algorithms	Citation recommendation, text summarization
Contains Sentences?	✓	✗	✗	✗	✗
Contains Paper Title?	✓	✗	✓	✓	✓
Contains Abstract?	✓	✓	✓	✓	✗ (Not all documents)
Contains Authors Names?	✓	✓	✓	✓	✓
Contains Keywords?	✗	✗	✓	✓	✗
Cover Multiple Domains?	✓	✓	✗	✓	✓
Covers Metadata of citation	✓	✗	✗	✗	✗
Data Time Period	2017-2023	1991-2023	1990-2023	2004-2023	Last release: 2021-02-01

Table 4: Comparison of different datasets

Domain	OpenAI	Mistral	L	RM	RL	Perplexity	AdvRAG(L)	AdvRAG(M)
AI	34:25	26:03	11:10	74:49	73:09	34:31	156:24	163:28
Biomolecules	01:11	00:41	00:10	4:38	4:10	00:20	7:29	7:40
CV	47:45	18:35	19:24	189:20	198:45	42:05	259:32	302:14
Crytography	03:50	02:18	04:59	83:28	89:21	13:23	190:19	194:25
Databases	01:27	00:51	00:40	49:34	45:46	00:51	96:19	97:48
Graphics	07:08	08:55	06:08	108:08	127:48	16:52	214:25	227:23
HCI	03:01	01:10	00:42	48:32	50:51	02:47	95:56	98:44
IR	20:31	11:40	06:52	91:30	99:43	19:50	193:37	202:23
NLP	28:26	11:42	05:09	91:07	88:40	13:06	175:58	156:49
NNC	05:00	01:39	02:12	34:56	41:09	01:19	70:17	84:07
QC	07:26	02:46	01:59	61:09	67:56	03:17	109:21	113:54
Robotics	19:39	05:41	06:11	41:67	46:55	09:17	92:67	98:45

Table 5: Time taken by different models with respect to each domain during experimentation, converted to **hours and minutes**. Red Color: Time recorded while using Replicate API, and Blue Color: Time recording while using NVIDIA A100/L40S USC server.

models.

For **direct prompting with metadata** also shows common domains across the models. Notable high HR domains such as **NNC**, **IR**, **NLP**, **QC**, and **Graphics** feature prominently across different models, indicating frequent challenges in these areas.

Low HR results consistently appear in **CV**, **NLP**, **Cryptography**, and **Biomolecules**, showcasing general robustness against hallucinations in these domains. Specifically, **NNC** is recurrently observed with high HR in the **G1**, **AdvRAG(L)**, and **AdvRAG(M)** models, while **QC** shows up frequently in high HR scenarios (**G1**, **G2**, **L**, **AdvRAG(M)**).

Similarly, **IR** is highlighted in high HR for the **P**, **RM**, **RL**, and **AdvRAG(L)** models, indicating its susceptibility, whereas **NLP** and **Graphics** show variability in HR across multiple models.

For **zero-shot direct prompting** also show significant patterns in common domains.

High HR is commonly observed in domains like **QC**, **Cryptography**, **Robotics**, and **Databases**, indicating areas prone to hallucinations. Low HR domains frequently include **IR**, **HCI**, **CV**, and **Biomolecules**, highlighting resilience in these areas.

Specifically, **QC** appears as a high HR domain in the **G1**, **G2**, **G3**, **RL**, **L**, **AdvRAG(L)**, and **AdvRAG(M)** models, reflecting a consistent challenge across these models. **IR** and **HCI** are notably present as low HR domains in **G2**, **G3**, **AdvRAG(L)**, showing widespread reliability.

Moreover, **Robotics** and **Cryptography** are frequently observed in high HR scenarios in models like **G2**, **M**, and **AdvRAG(M)**, while **CV** and **Biomolecules** commonly appear in low HR settings across **G2**, **G3**, **M**, and **AdvRAG()**.

For **SID prompting**, high HR domains such as **QC**, **Cryptography**, **Databases**, **NNC**, and **Robotics** frequently appear across several models, highlighting a general susceptibility in these areas. On the other hand, low HR domains commonly include **IR**, **HCI**, **CV**, and **Graphics**, demonstrating resilience against hallucinations.

Specifically, **QC** is observed as a high HR domain in the **G1**, **G2**, **G3**, **RM**, **RL**, **AdvRAG(L)**, and **AdvRAG(M)** models, signifying a consistent

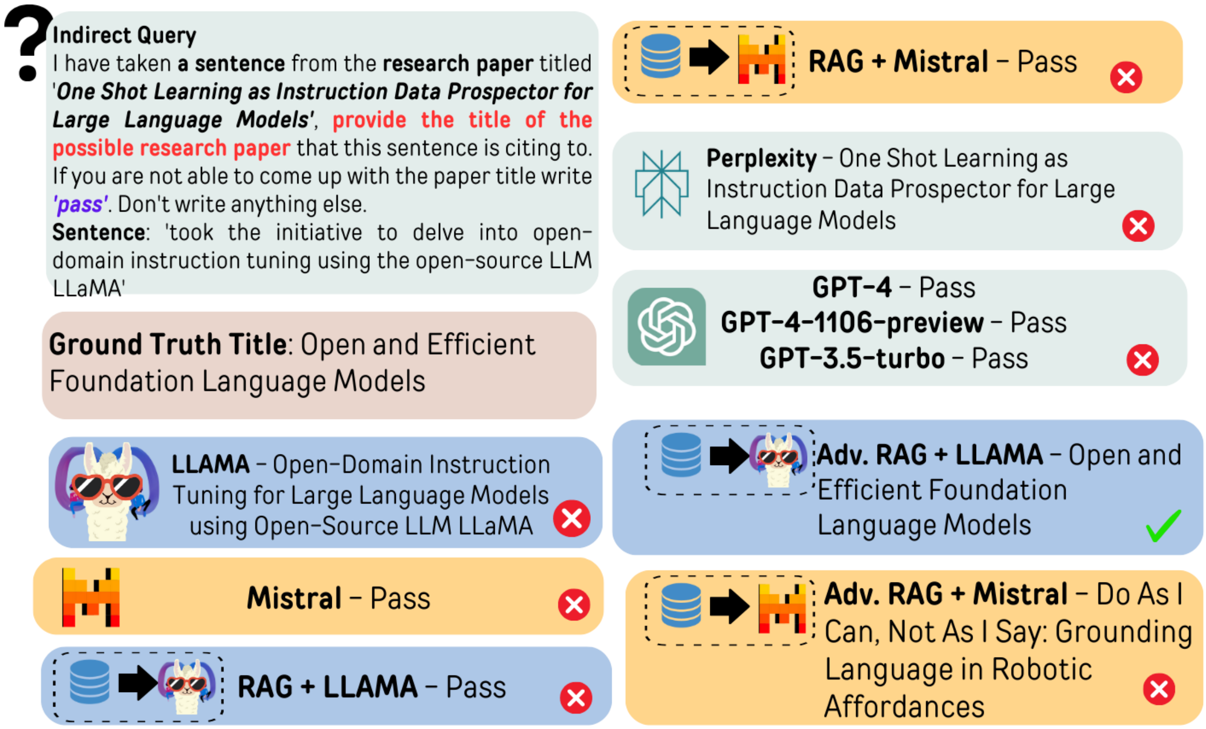

Figure 7: Example 1 of an indirect query where a sentence from the research paper is provided and asked for the correct title. We have ground truth for the paper title and responses from various LLMs. Only Adv. RAG+LLAMA generated the correct title.

challenge in this area. **IR** and **HCI** are notably present as low HR domains in **G1**, **G2**, **G3**, **RM**, and **AdvRAG(L)**, indicating widespread reliability in these areas.

Moreover, **Cryptography** and **Robotics** are frequently observed in high HR scenarios in models like **G1**, **G2**, and **RM**, while **CV** and **Graphics** commonly appear in low HR settings across **G2**, **L**, and **AdvRAG(L)**. To summarize our results

- The **zero-shot indirect** and **SID** promoting styles are more prone to hallucinations, which lack contextual understanding.
- Notably, **NNC** and **QC** consistently show high HR across multiple models and promoting styles, indicating common challenging domains.
- Conversely, **CV** and **IR** low HR, which show robustness in models, suggesting reliability in these domains across different prompting strategies.

B.2 Further Discussion on Adversarial Examination

This analysis emphasizes the strengths and weaknesses of current LLMs and the need for domain-specific training. It shows that a general approach is insufficient and highlights the importance of specialized training to meet the unique demands of different fields. As LLMs evolve, aligning their development with human knowledge's varied and intricate nature is crucial.

The study finds a significant relationship between the specificity of prompts, especially those with metadata, and the linguistic accuracy of LLMs, as evidenced by higher F-1 and BLEU scores. This suggests that providing detailed, context-rich prompts can significantly improve the quality of generated citations.

Pass Percentage (PP): The varying PP among different models points to a key challenge in LLM development: the ability to understand and reason through complex situations. Models with lower PP struggle with generating relevant responses in complex or critical scenarios, underlining the importance of enhancing reasoning capabilities in LLMs for effective application.

Prompt Design: There's a noticeable difference in how individual models, such as gpt-4-1106-preview and gpt-4, respond to different prompts. This underscoring the significance of prompt design in leveraging the full potential of LLMs suggests a complex interplay between the model's structure, prompt formulation, and performance.

Zero-Shot Indirect								
Domain	G1	G2	G3	P	RM	M	RL	L
Hallucination Rate (%)								
AI	63.61	72.44	81.87	96.27	93.98	97.16	92.21	95.87
Biomolecules	96.82	69.77	84.68	95.06	96.63	85.14	96.25	95.57
Crypto	75.04	70.21	81.97	94.16	93.07	96.11	93.83	97.23
CV	51.83	64.3	79.34	94.63	91.42	97.12	94.68	95.96
Databases	76.66	69.99	78.93	96.99	93.42	97.28	95.68	95.84
Graphics	57.49	70.76	85.39	97.25	92.32	97.55	96.1	95.92
HCI	51.83	73.46	73.41	96.71	93.01	96.83	96.85	95.61
IR	51.78	67.89	73.41	96.80	92.01	96.81	96.85	96.01
NLP	63.03	73.98	74.77	97.11	94.10	97.05	94.29	97.93
NNC	77.27	80.75	82.11	95.49	94.32	97.13	97.92	96.14
QC	91.72	84.85	76.09	95.15	92.13	97.14	95.34	95.56
Robotics	55.78	71.55	76.73	95.81	94.26	97.2	97.51	95.67
Mean	67.73	72.49	79.05	95.95	93.38	96.04	95.64	96.10
Standard Deviation	15.64	5.51	4.19	1.05	1.40	3.45	1.67	0.72
F-1 Score								
AI	0.02	0.22	0.00	0.00	0.10	0.08	0.07	0.05
Biomolecules	0.00	0.26	0.00	0.07	0.09	0.06	0.06	0.05
Crypto	0.01	0.25	0.00	0.00	0.08	0.04	0.06	0.04
CV	0.06	0.29	0.00	0.00	0.07	0.05	0.05	0.04
Databases	0.00	0.26	0.00	0.00	0.09	0.06	0.05	0.04
Graphics	0.06	0.25	0.00	0.00	0.05	0.03	0.03	0.01
HCI	0.04	0.23	0.00	0.00	0.07	0.03	0.04	0.03
IR	0.06	0.29	0.00	0.00	0.04	0.01	0.03	0.02
NLP	0.02	0.21	0.00	0.00	0.07	0.04	0.04	0.03
NNC	0.02	0.16	0.00	0.00	0.06	0.04	0.02	0.01
QC	0.01	0.13	0.00	0.00	0.05	0.02	0.03	0.01
Robotics	0.03	0.21	0.00	0.00	0.08	0.05	0.03	0.02
Mean	0.02	0.23	0.00	0.00	0.07	0.04	0.04	0.02
Standard Deviation	0.02	0.04	0.00	0.02	0.01	0.01	0.01	0.01
BLEU Score								
AI	0.01	0.09	0.00	0.00	0.05	0.00	0.06	0.00
Biomolecules	0.00	0.12	0.00	0.00	0.00	0.00	0.04	0.00
Crypto	0.01	0.12	0.00	0.00	0.07	0.00	0.05	0.00
CV	0.04	0.16	0.00	0.00	0.02	0.00	0.03	0.00
Databases	0.00	0.12	0.00	0.00	0.08	0.00	0.03	0.00
Graphics	0.04	0.12	0.00	0.00	0.03	0.00	0.01	0.00
HCI	0.03	0.09	0.00	0.00	0.05	0.00	0.02	0.00
IR	0.04	0.14	0.00	0.00	0.01	0.00	0.02	0.00
NLP	0.02	0.09	0.00	0.00	0.06	0.00	0.00	0.00
NNC	0.02	0.05	0.00	0.00	0.02	0.00	0.00	0.00
QC	0.00	0.02	0.00	0	0.01	0.00	0.00	0.00
Robotics	0.02	0.08	0.00	0.00	0.06	0.00	0.00	0.00
Mean	0.01	0.10	0.00	0.00	0.03	0.00	0.02	0.00
Standard Deviation	0.01	0.03	0.00	0.00	0.02	0.00	0.02	0.00
Pass Percentage (%)								
AI	92.92	24.15	97.08	97.77	4.95	0.05	0	0
Biomolecules	88.89	19.76	97.81	0	0	0	0	0
Crypto	92.45	20.47	98.17	99.01	5.63	0.09	0	0
CV	86.7	23.8	95.66	96.48	3.84	0	0	0
Databases	97.25	20.11	97.67	97.14	6.23	0	0	0
Graphics	86.38	19.69	97.32	98.8	1.34	0	0	0
HCI	90.83	19.21	96.61	98.32	6.11	0	0	0
IR	87.67	16.69	96.61	97.83	5.21	0	0	0
NLP	92.4	21.98	97.89	98.53	6.75	0	0	0
NNC	87.73	20.86	98.16	95.21	6.39	0	0	0
QC	75	17.76	99.34	95.09	5.72	0	0	0
Robotics	92.91	31.7	97.68	95.95	5.73	0	0	0
Mean	89.26	21.34	97.50	89.17	4.82	0.01	0.00	0.00
Standard Deviation	5.528	3.91	0.94	28.11	2.10	0.02	0.00	0.00

Table 6: Zero-Shot Indirect

Direct with Metadata										
Domain	G1	G2	G3	P	RM	M	RL	L	AdvRAG(L)	AdvRAG(M)
Hallucination Rate (%)										
AI	0.32	0.10	6.04	61.31	37.6	71.39	72.16	80.90	19.24	7.67
Biomolecules	0.46	0.01	5.29	73.99	94.5	67.98	87.10	79.15	8.15	0.07
Crypto	0.42	0.05	5.41	61.77	40.87	71.56	73.18	80.45	6.76	4.15
CV	0.42	0.07	4.9	62.35	41.60	73.67	74.16	78.93	5.51	2.22
Databases	0.20	0.15	5.05	62.55	39.60	73.33	75.16	0.79	9.73	7.60
Graphics	0.20	0.15	5.43	62.64	42.31	71.43	78.21	79.80	11.45	8.10
HCI	0.24	0.26	5.26	60.38	40.75	73.29	75.45	80.66	17.65	7.04
IR	0.39	0.09	5.26	63.88	48.98	73.1	79.43	80.98	19.71	7.81
NLP	0.64	0.27	6.20	58.79	37.44	69.68	71.24	80.17	12.60	5.80
NNC	0.51	0.16	5.82	61.12	38.73	72.04	75.14	81.31	28.11	57.95
QC	0.54	0.17	4.95	61.97	38.54	69.34	72.09	81.70	18.19	9.25
Robotics	0.45	0.12	5.98	61.89	39.01	70.62	71.02	80.34	10.27	3.88
Mean	0.39	0.13	5.46	62.72	44.99	71.45	75.36	80.28	13.94	10.70
Standard Deviation	0.13	0.07	0.44	3.76	15.89	1.79	4.52	0.90	6.67	15.01
F-1 Score										
AI	0.99	0.89	0.95	0.69	0.71	0.36	0.33	0.28	0.84	0.92
Biomolecules	0.97	0.99	0.96	0.36	0.07	0.07	0.21	0.32	0.96	0.95
Crypto	0.93	0.97	0.96	0.61	0.60	0.40	0.37	0.31	0.91	0.94
CV	0.98	0.99	0.96	0.39	0.52	0.38	0.34	0.35	0.98	0.98
Databases	0.99	0.98	0.96	0.42	0.59	0.34	0.34	0.33	0.92	0.95
Graphics	0.99	0.99	0.96	0.45	0.64	0.44	0.41	0.32	0.94	0.90
HCI	0.99	0.98	0.96	0.34	0.58	0.35	0.35	0.34	0.82	0.94
IR	0.99	0.98	0.94	0.52	0.54	0.39	0.39	0.30	0.84	0.92
NLP	0.99	0.92	0.95	0.53	0.62	0.42	0.40	0.31	0.86	0.91
NNC	0.99	0.99	0.95	0.51	0.62	0.41	0.36	0.30	0.92	0.39
QC	0.99	0.99	0.96	0.58	0.65	0.43	0.33	0.29	0.82	0.86
Robotics	0.99	0.99	0.95	0.63	0.69	0.35	0.49	0.31	0.92	0.95
Mean	0.98	0.98	0.95	0.50	0.56	0.36	0.35	0.32	0.89	0.88
Standard Deviation	0.01	0.00	0.00	0.11	0.16	0.09	0.06	0.02	0.05	0.15
BLEU Score										
AI	0.99	0.99	0.93	0.31	0.43	0.24	0.11	0.12	0.81	0.92
Biomolecules	0.95	0.99	0.94	0.22	0.00	0.00	0.07	0.12	0.93	0.02
Crypto	0.95	0.97	0.94	0.33	0.41	0.24	0.13	0.12	0.93	0.95
CV	0.95	0.99	0.94	0.32	0.39	0.22	0.13	0.13	0.95	0.96
Databases	0.98	0.99	0.94	0.33	0.41	0.21	0.13	0.13	0.79	0.86
Graphics	0.99	0.99	0.94	0.33	0.45	0.24	0.17	0.12	0.91	0.91
HCI	0.99	0.98	0.94	0.33	0.43	0.22	0.13	0.14	0.91	0.92
IR	0.99	0.99	0.94	0.36	0.48	0.23	0.16	0.11	0.87	0.92
NLP	0.99	0.99	0.93	0.37	0.46	0.27	0.12	0.12	0.82	0.91
NNC	0.99	0.99	0.93	0.34	0.46	0.22	0.12	0.11	0.90	0.17
QC	0.98	0.98	0.93	0.28	0.38	0.26	0.15	0.11	0.80	0.83
Robotics	0.99	0.99	0.93	0.34	0.49	0.26	0.18	0.12	0.89	0.94
Mean	0.97	0.98	0.93	0.32	0.39	0.21	0.13	0.12	0.87	0.77
Standard Deviation	0.01	0.00	0.00	0.03	0.13	0.07	0.02	0.00	0.05	0.32
Pass Percentage (%)										
AI	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Biomolecules	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Crypto	0.00	0.00	0.00	0.15	0.00	0.00	0.00	0.00	0.00	0.00
CV	0.00	0.00	0.00	0.03	0.00	0.00	0.00	0.00	0.00	0.00
Databases	0.00	0.00	0.00	0.72	0.00	0.00	0.00	0.00	0.00	0.00
Graphics	0.00	0.00	0.00	0.33	0.00	0.00	0.00	0.00	0.00	0.00
HCI	0.00	0.00	0.00	0.24	0.44	0.00	0.00	0.00	0.00	0.00
IR	0.00	0.00	0.00	0.03	0.67	0.00	0.00	0.00	0.00	0.00
NLP	0.00	0.00	0.00	0.09	0.00	0.00	0.00	0.00	0.00	0.00
NNC	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
QC	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Robotics	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Mean	0.00	0.00	0.00	0.13	0.09	0.00	0.00	0.00	0.00	0.00
Standard Deviation	0.00	0.00	0.00	0.21	0.22	0.00	0.00	0.00	0.00	0.00

Table 7: Direct with Metadata

Zero-Shot Direct Prompting										
Domain	G1	G2	G3	P	RM	M	RL	L	AdvRAG(L)	AdvRAG(M)
Hallucination Rate (%)										
AI	30.9	53.99	73.13	95.64	56.45	94.23	72.17	76.85	43.77	34.42
CV	35.9	36.32	61.38	95.84	58.45	92.84	73.17	76.67	35.38	35.43
NLP	27.51	52.49	72.28	96.18	63.92	93.89	83.17	75.91	47.95	36.63
IR	24.82	42.55	64.19	95.23	63.12	91.59	77.38	78.16	42.01	37.93
Databases	37.48	53.33	74.08	95.98	55.45	93.81	74.17	77.92	58.11	40.23
Graphics	29.3	54.29	73.71	95.67	52.4	92.99	71.19	75.57	47.41	40.26
HCI	22.92	38.02	64.19	95.01	62.67	92.64	78.15	76.49	38.51	41.11
Biomolecules	21.01	53.25	73.88	90.83	94.00	43.84	91.2	79.92	67.56	46.28
NNC	36.05	53.13	72.39	93.37	63.51	91.18	83.73	78.24	48.51	46.31
Crypto	34.41	54.68	73.01	95.39	54.45	94.78	76.59	76.44	66.16	50.08
Robotics	34.71	56.62	76.29	93.25	60.89	94.69	81.99	75.92	59.017	50.65
QC	53.04	70.01	82.26	93.70	65.07	89.75	85.64	81.24	69.108	60.81
Mean	32.33	51.55	71.73	94.67	62.53	88.85	79.04	77.44	51.95	43.34
Standard Deviation	8.52	9.02	5.80	1.58	10.76	14.25	6.14	1.73	11.66	7.75
F-1 Score										
AI	0.42	0.39	0.21	0.04	0.41	0.06	0.31	0.36	0.46	0.53
Biomolecules	0.37	0.42	0.21	0.08	0.07	0.05	0.14	0.31	0.29	0.65
Crypto	0.42	0.41	0.22	0.04	0.43	0.06	0.32	0.36	0.40	0.56
CV	0.42	0.60	0.33	0.05	0.39	0.07	0.32	0.36	0.62	0.62
Databases	0.40	0.42	0.21	0.05	0.41	0.06	0.31	0.34	0.42	0.55
Graphics	0.49	0.41	0.22	0.05	0.44	0.07	0.33	0.38	0.42	0.56
HCI	0.51	0.55	0.29	0.05	0.36	0.07	0.27	0.36	0.62	0.56
IR	0.51	0.52	0.29	0.05	0.35	0.08	0.26	0.34	0.57	0.69
NLP	0.39	0.38	0.21	0.04	0.35	0.06	0.21	0.37	0.52	0.66
NNC	0.39	0.39	0.19	0.06	0.37	0.08	0.24	0.34	0.48	0.57
QC	0.22	0.25	0.12	0.06	0.34	0.09	0.18	0.30	0.30	0.40
Robotics	0.35	0.36	0.20	0.06	0.33	0.05	0.15	0.37	0.41	0.54
Mean	0.40	0.42	0.22	0.05	0.35	0.06	0.25	0.34	0.45	0.57
Standard Deviation	0.07	0.09	0.05	0.01	0.09	0.01	0.06	0.02	0.10	0.07
BLEU Score										
AI	0.37	0.31	0.11	0.00	0.24	0.00	0.17	0.15	0.38	0.49
Biomolecules	0.34	0.33	0.10	0.00	0.00	0.02	0.04	0.11	0.27	0.60
Crypto	0.37	0.32	0.11	0.00	0.25	0.00	0.18	0.15	0.26	0.47
CV	0.40	0.52	0.23	0.00	0.24	0.00	0.16	0.15	0.57	0.58
Databases	0.32	0.33	0.10	0.00	0.25	0.00	0.18	0.14	0.31	0.42
Graphics	0.44	0.31	0.11	0.00	0.23	0.00	0.19	0.16	0.70	0.51
HCI	0.46	0.46	0.18	0.00	0.22	0.00	0.13	0.15	0.64	0.51
IR	0.45	0.44	0.18	0.00	0.28	0.00	0.17	0.14	0.48	0.62
NLP	0.34	0.32	0.11	0.00	0.21	0.00	0.12	0.16	0.46	0.51
NNC	0.33	0.28	0.11	0.00	0.19	0.00	0.10	0.14	0.48	0.57
QC	0.17	0.14	0.02	0.00	0.17	0.00	0.08	0.11	0.20	0.29
Robotics	0.30	0.28	0.09	0.00	0.18	0.00	0.09	0.16	0.30	0.41
Mean	0.35	0.33	0.12	0.00	0.20	0.00	0.13	0.14	0.42	0.49
Standard Deviation	0.07	0.09	0.05	0.00	0.07	0.00	0.04	0.01	0.16	0.09
Pass Percentage (%)										
AI	37.26	9.70	12.37	0.66	1.65	0.00	0.00	0.00	0.00	0.00
Biomolecules	51.85	6.77	6.77	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Crypto	33.4	5.43	10.52	0.20	2.15	0.00	0.00	0.00	0.00	0.00
CV	32.26	3.84	8.67	0.09	3.12	0.09	0.00	0.00	0.00	0.00
Databases	32.42	6.70	10.59	0.95	2.49	0.00	0.00	0.00	0.00	0.00
Graphics	28.86	6.49	10.30	0.15	0.45	0.07	0.00	0.00	0.00	0.00
HCI	31.00	8.30	14.51	0.32	0.56	0.00	0.00	0.00	0.00	0.00
IR	30.11	6.11	14.51	0.86	0.87	0.00	0.00	0.00	0.00	0.00
NLP	44.6	15.75	17.03	0.18	1.76	0.00	0.00	0.00	0.00	0.00
NNC	37.12	13.19	21.47	0.74	1.53	0.00	0.00	0.00	0.00	0.00
QC	50.22	10.09	19.96	0.00	1.94	0.00	0.00	0.00	0.00	0.00
Robotics	45.10	11.60	9.02	0.00	4.54	0.00	0.00	0.00	0.00	0.00
Mean	37.85	8.66	12.97	0.34	1.75	0.01	0.00	0.00	0.00	0.00
Standard Deviation	8.06	3.50	4.61	0.35	1.25	0.03	0.00	0.00	0.00	0.00

Table 8: Zero-Shot Direct

SID										
Domain	G1	G2	G3	P	RM	M	RL	L	AdvRAG(L)	AdvRAG(M)
Hallucination Rate (%)										
AI	29.44	48.49	61.18	95.08	85.21	94.18	86.68	98.42	51.47	38.45
Biomolecules	35.71	54.99	66.34	95.79	96.87	86.32	96.51	99.06	52.15	40.89
Crypto	40.44	48.15	66.48	91.18	85.28	94.78	86.91	98	53.67	45.77
CV	34.44	38.15	59.77	93.47	87.65	94.13	89.58	99.56	38.82	39.25
Databases	40.74	62.34	66.00	93.91	86.66	93.96	86.10	98.67	62.49	43.2
Graphics	25.54	62.34	66.55	95.28	85.91	94.39	86.41	58.83	59.65	47.72
HCI	27.35	39.58	57.01	94.41	85.68	93.87	88.15	98.12	30.53	23.39
IR	24.01	41.87	57.01	94.68	85.61	93.33	88.45	98.57	58.58	40.97
NLP	29.2	50.69	61.68	95.87	88.46	93.88	89.28	98.64	60.26	37.72
NNC	32.68	57.13	74.64	95.97	88.01	95.14	89.56	99.34	59.42	64.43
QC	51.83	63.63	80.05	92.10	89.75	95.49	90.73	98.98	69.18	59.84
Robotics	32.45	49.76	57.27	95.07	89.46	94.36	90.86	98.27	49.24	34.95
Mean	33.65	51.42	64.49	94.40	87.87	93.65	89.10	95.371	53.788	43.048
Standard Deviation	7.80	8.85	7.16	1.51	3.25	2.38	2.85	11.51	10.60	10.84
F-1 Score										
AI	0.30	0.54	0.05	0.09	0.12	0.11	0.20	0.02	0.50	0.61
Biomolecules	0.15	0.51	0.03	0.05	0.05	0.03	0.05	0.00	0.52	0.57
Crypto	0.35	0.67	0.03	0.07	0.13	0.10	0.19	0.02	0.62	0.71
CV	0.35	0.67	0.06	0.09	0.13	0.11	0.16	0.03	0.72	0.73
Databases	0.21	0.03	0.03	0.08	0.14	0.10	0.19	0.02	0.29	0.48
Graphics	0.41	0.03	0.03	0.05	0.13	0.09	0.18	0.41	0.38	0.58
HCI	0.33	0.66	0.07	0.08	0.15	0.13	0.18	0.03	0.70	0.85
IR	0.38	0.64	0.07	0.08	0.14	0.12	0.15	0.02	0.43	0.68
NLP	0.30	0.51	0.05	0.07	0.16	0.10	0.13	0.02	0.41	0.49
NNC	0.21	0.45	0.03	0.09	0.11	0.08	0.17	0.00	0.50	0.31
QC	0.10	0.37	0.02	0.06	0.10	0.07	0.13	0.01	0.31	0.42
Robotics	0.28	0.54	0.05	0.07	0.13	0.09	0.14	0.02	0.60	0.62
Mean	0.28	0.46	0.04	0.07	0.12	0.09	0.15	0.05	0.49	0.58
Standard Deviation	0.09	0.22	0.01	0.01	0.02	0.02	0.04	0.11	0.14	0.14
BLEU Score										
AI	0.25	0.31	0.02	0.00	0.06	0.00	0.04	0.00	0.32	0.51
Biomolecules	0.14	0.34	0.01	0.00	0.00	0.00	0.00	0.00	0.32	0.56
Crypto	0.27	0.48	0.01	0.00	0.06	0.00	0.06	0.00	0.47	0.55
CV	0.25	0.46	0.03	0.00	0.03	0.01	0.06	0.00	0.51	0.51
Databases	0.17	0.01	0.01	0.00	0.06	0.00	0.03	0.00	0.12	0.42
Graphics	0.35	0.01	0.01	0.00	0.03	0.00	0.01	0.26	0.22	0.44
HCI	0.28	0.45	0.03	0.00	0.07	0.01	0.05	0.00	0.53	0.71
IR	0.32	0.39	0.03	0.00	0.07	0.01	0.07	0.00	0.54	0.45
NLP	0.26	0.27	0.03	0.00	0.04	0.01	0.04	0.00	0.23	0.43
NNC	0.15	0.24	0.01	0.00	0.05	0.00	0.05	0.00	0.40	0.11
QC	0.08	0.17	0.00	0.00	0.04	0.00	0.03	0.00	0.20	0.31
Robotics	0.22	0.28	0.03	0.00	0.04	0.00	0.03	0.00	0.30	0.44
Mean	0.22	0.28	0.01	0.00	0.04	0.00	0.03	0.02	0.34	0.45
Standard Deviation	0.07	0.15	0.01	0.00	0.02	0.00	0.02	0.07	0.14	0.14
Pass Percentage (%)										
AI	56.8	4.21	87.14	1.86	7.25	0.00	0.00	0.00	0.00	0.00
Biomolecules	74.07	7.21	89.98	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Crypto	53.34	3.6	89.7	0.84	6.89	0.00	0.00	0.00	0.00	0.00
CV	52.3	1.6	83.42	0.79	4.94	0.00	0.00	0.00	0.00	0.00
Databases	63.19	89.98	90.61	0.00	6.04	0.00	0.00	0.00	0.00	0.00
Graphics	44.25	88.91	90.19	0.64	6.29	0.00	0.79	0.00	0.00	0.00
HCI	54.15	0.44	83.68	0.96	4.37	0.00	0.00	0.00	0.00	0.00
IR	49.52	1.45	83.68	0.79	4.39	0.00	0.00	0.00	0.00	0.00
NLP	57.33	5.49	86.45	2.38	4.91	0.00	0.00	0.00	0.00	0.00
NNC	69.33	5.21	87.42	2.88	5.93	0.00	0.00	0.00	0.00	0.00
QC	76.75	7.46	88.6	2.14	5.97	0.00	0.00	0.00	0.00	0.00
Robotics	57.6	3.35	86.86	2.65	7.31	0.00	0.00	0.00	0.00	0.00
Mean	59.05	18.33	87.31	1.30	5.357	0.00	0.65	0.00	0.00	0.00
Standard Deviation	9.92	33.53	2.63	1.02	1.97	0.00	0.22	0.00	0.00	0.00

Table 9: SID

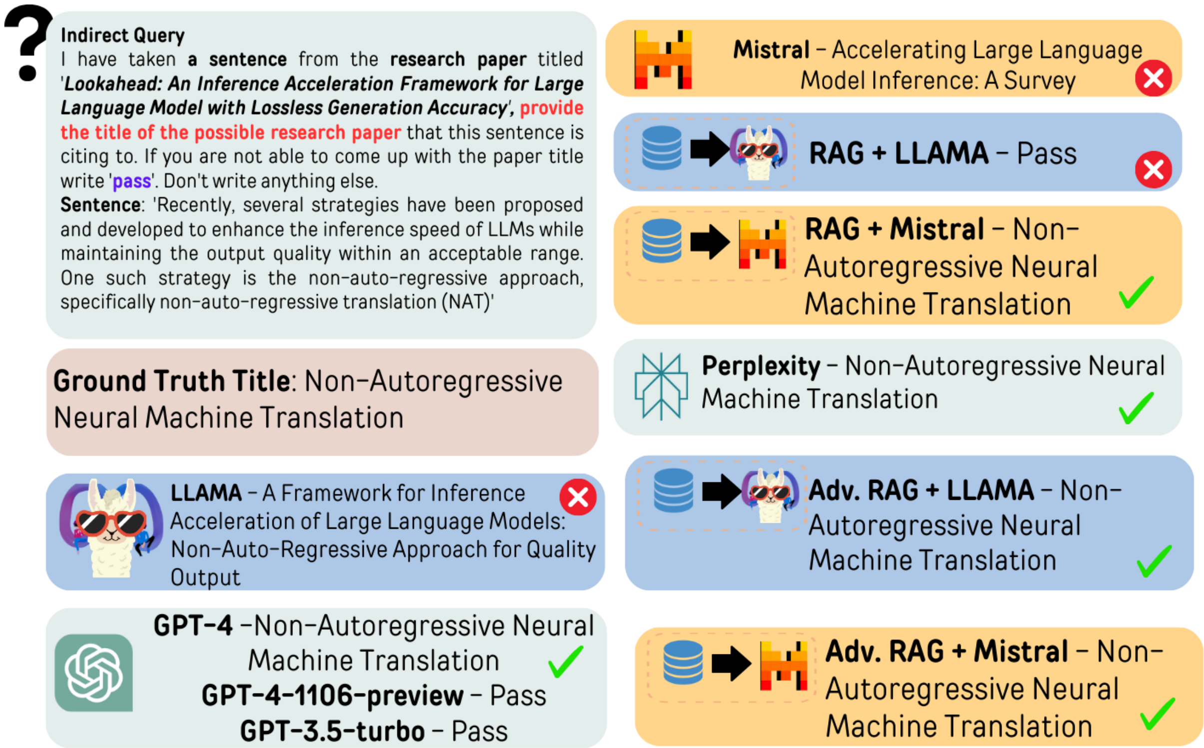

Figure 8: Example 2 of an indirect query is where a sentence from the research paper is provided and asked for the correct title. Here, we can see that GPT-4, RAG+Mistral, Adv. RAG+Mistral, Adv. RAG+LLAMA and Perplexity yield the correct title.

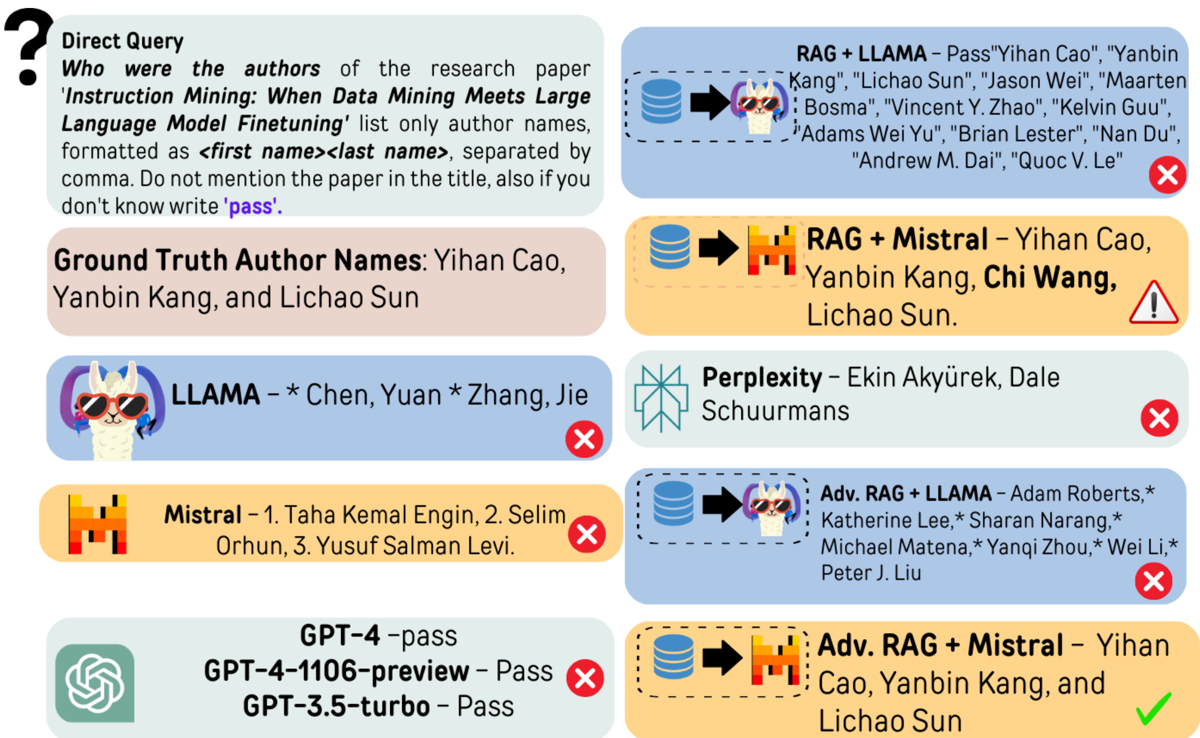

Figure 9: Example 1 of zero-shot direct prompting demonstrated that only the Adv.RAG(M), i.e., with cross-encoder reranking, accurately produced all the correct author names. It's noteworthy that the basic RAG+Mistral version only made a single error in the author names, but the addition of the advance reranking process in Adv. RAG+Mistral rectified this and yielded the correct title.

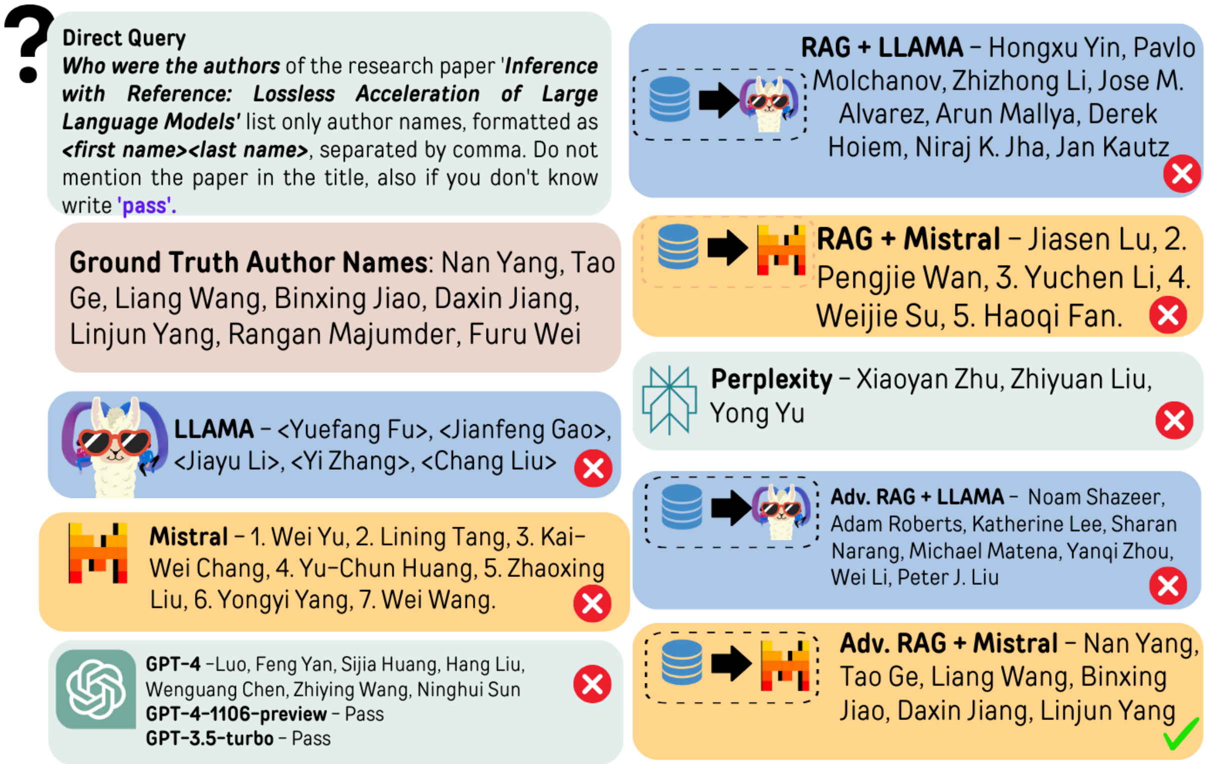

Figure 10: Example 2 of zero-shot direct prompting demonstrated RAG+LLAMA, Adv. RAG+LLAMA, Adv. RAG+Mistral yields the correct title.

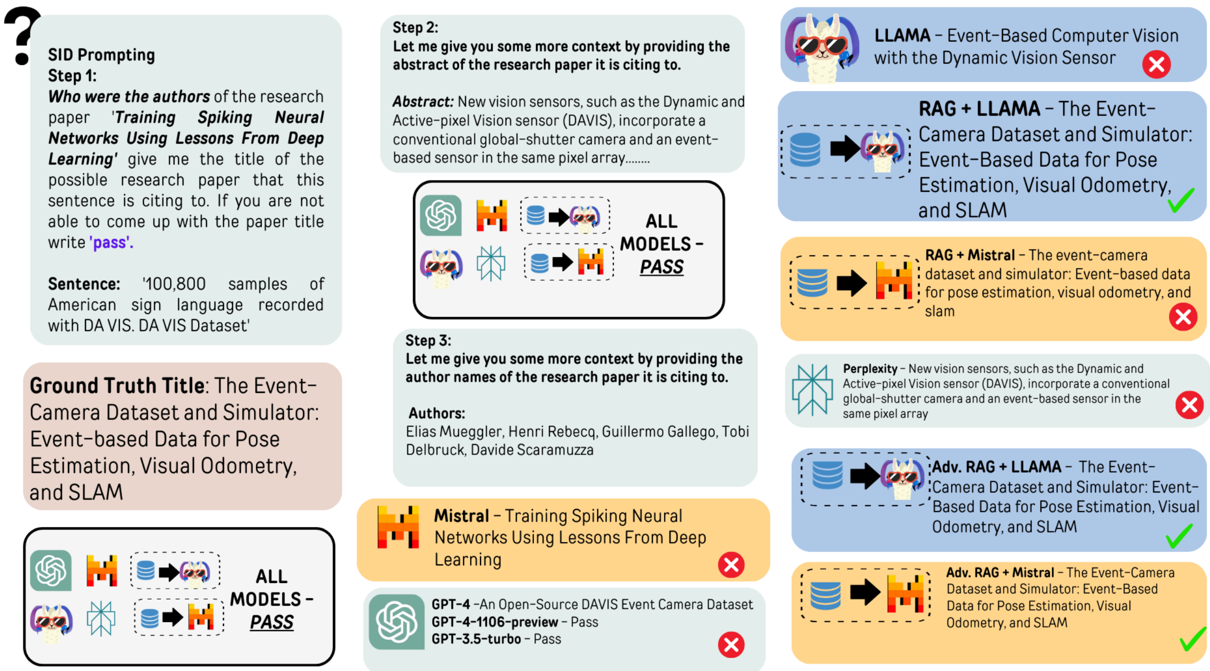

Figure 11: In SID prompting, asking the indirect query yielded a pass for all models. After providing a complete abstract ([...], in the image, we did not add a complete abstract because of space constraints, but the actual prompt was provided with a complete abstract), it still yielded a pass. Then, we provided the abstract names, which shows that only RAG models yielded the right titles.

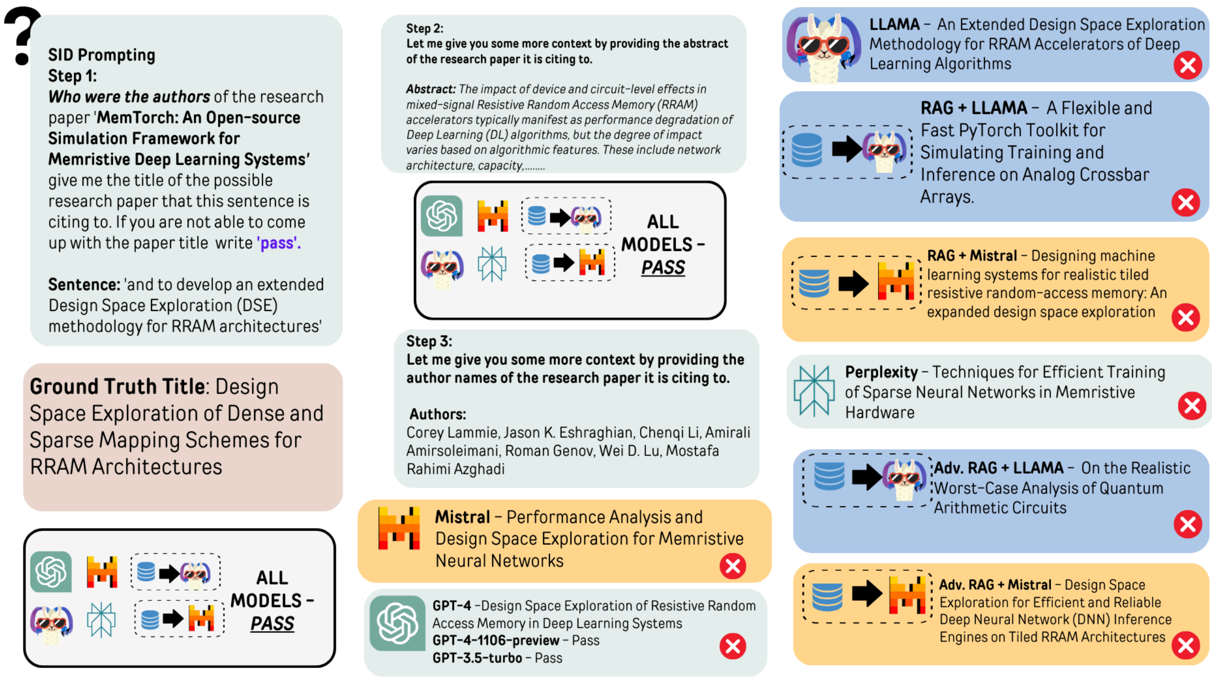

Figure 12: Worst case example of SID prompting where it did not yield correct title to any model.

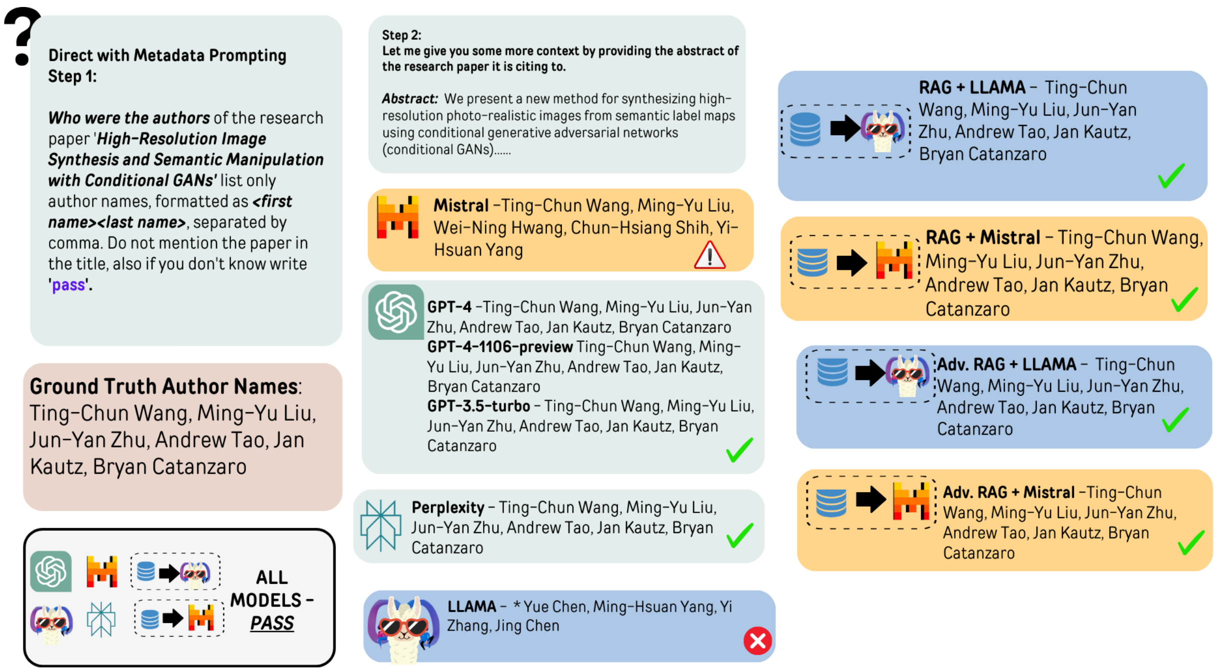

Figure 13: An example of a direct prompt scenario where initially all models failed to identify the author names and responded pass. Upon presenting the abstract, all but the LLAMA model, and to some extent Mistral (a few of the wrong names in the list with correct names were generated), failed to respond appropriately to the prompt.