What Kind of Sourcery is This? Evaluating GPT-4's Performance on Linking Scientific Fact to Citations

Autumn Toney-Wails
Georgetown University
SciTech Strategies, Inc.
`autumn.toney@georgetown.edu`

Abstract

From document summarization to code generation, chabots have disrupted various aspects of scientific research and writing. While chabots are useful research resources for ideation, information retrieval, and editing, their generative pre-trained transformer (GPT) models' underlying knowledge infrastructure is opaque. This has raised questions about the reliability of generative chatbot responses, as GPT models are known to respond with misleading information that appears to be accurate. Prior research has investigated the utility of OpenAI's public chatbot, ChatGPT, to generate reliable bibliographic information with a focus on small-scale medical-related scientific facts. We present an expanded study that analyzes GPT-4's ability to accurately identify 1,326 scientific facts and link them to academic sources. Using both the API and UI service, we experimented with open-ended and close-ended prompts to establish an understanding of GPT-4's general ability at this domain-specific task, as well as study the real-world scenario of an average user interacting with ChatGPT using its UI. GPT-4 accurately identified 96% of the scientific facts and generated relevant and existent academic citations with 78% accuracy. Using the claims that GPT-4 mislabeled and provided incorrect sources via the API, we prompt two public GPTs customized for academic writing to evaluate if they correctly label the scientific claims and provide accurate sources. We find that these GPTs are able to accurately label 38% of the mislabeled claims, with 95% of the corresponding citations being accurate and relevant.

1 Introduction

With the ability to perform a wide range of natural language generation (NLG) and information retrieval tasks, chatbots have enabled individuals to experiment with the utility of generative pre-trained transformer (GPT) language models in a publicly available, online interface. While chatbots are *generative* AI tools, users often query chatbots in a paired task that includes both NLG and information retrieval; for example, generating new content (e.g., write an introduction for a paper on a given topic) and retrieving information (e.g., provide citations when necessary). However, users often engage with chatbots for a specific task without understanding its utility in the given domain.

Using a chatbot as an information gathering tool is convenient, but comes with caveats. Various studies that analyze a chabot's performance on NLG and information retrieval tasks (e.g., document summarization and code generation) highlight a persistent error in the GPT model's responses—*hallucinations* (Shuster et al., 2021; Ji et al., 2023). Hallucinations refer to factually incorrect responses that often pass as being correct and credible text to a user (Dziri et al., 2022). Hallucinations are harmful to users, particularly in information retrieval-like tasks where the user is not an expert in the prompt topic, because chatbots can respond with well-formatted text that is convincingly accurate, but is completely fabricated.

In this work, we focus on a particular use-case for information gathering—linking scientific facts to sources for citations. Prior research has focused on evaluating GPT models (mainly versions 3 and 3.5) via the online ChatGPT interface in small scale experiments on complex scientific topics for citation generation (Wagner and Ertl-Wagner, 2023; Sebo, 2023; Xames and Shefa, 2023). Our study expands this research to prompt GPT-4 via the API on 1,326 scientific facts from 3rd–5th grade level coursework, covering a range of scientific topics. Specifically, we design an automated prompt framework that includes a close-ended prompt ("is the fact true or false?") and an open-ended prompt ("provide a citation to support your response") to analyze GPT-4's ability to identify scientific facts and accurately link them to academic citations. We then provide human annotation to evaluate the ac-

Proceedings of the 1st Workshop on Customizable NLP: Progress and Challenges in Customizing NLP for a Domain, Application, Group, or Individual (CustomNLP4U), pages 257–268
November 16, 2024 ©2024 Association for Computational Linguistics

curacy of GPT-4's responses, assessing if the provided citation is relevant to the scientific fact and exists (i.e. the source is not hallucinated).

Further assessing GPT-4's ability to generate reliable and accurate bibliographic information, we design a second prompt with two close-ended questions to verify its prior responses on the same criteria as the human annotation: "is the citation relevant to the scientific fact?" (yes or no) and "does the citation exist?" (real or fake). The full experimental design is illustrated in Figure 1.

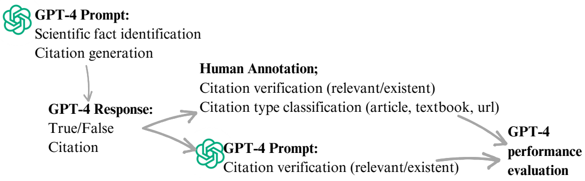

GPT-4 Prompt:
Scientific fact identification
Citation generation
Human Annotation;
Citation verification (relevant/existent)
Citation type classification (article, textbook, url)
GPT-4 Response:
True/False
Citation
GPT-4 Prompt:
Citation verification (relevant/existent)
GPT-4 performance evaluation

Figure 1: Experimental design framework for GPT-4 API prompts and response evaluation

Evaluating GPT-4's ability to verify scientific fact and provide a corresponding source via the API, we then use two sets of GPT-4 labeled claims for further experimentation: (1) claims that GPT-4 incorrectly labeled as *false* and (2) claims that GPT-4 provided hallucination sources for. We select two public GPTs[1] cusotmized for academic writing to converse with in the online user interface (UI). This experiment captures a real-world chabot interaction, where a customized chatbot is being used as a tool for a domain-specific task, while the API experiment comprehensively evaluates GPT-4's knowledge capacity and retrieval capabilities in an automated pipeline.

Our experimental results show that GPT-4 is capable of identifying scientific fact with 96% accuracy and generating a relevant, existing citation with 78% accuracy. GPT-4 favored providing a textbook citation over a scientific article or website, and only hallucinated 1% of textbook citation responses. We find that GPT-4 performs poorly as an evaluator of generated citations (determining if a source exitst), only correctly identifying 2% of the non-existent citations. In the UI experiments, we find that GPTs customized for academic writing increased the accuracy of scientific claim verification, with 38% of the previously 56 mislabeled claims receiving correct *true* labels. Additionally, the academic GPTs provided accurate and relevant citations with 95% accuracy for this set of claims. Analyzing the GPTs on a sample of 50 of the claims that GPT-4 correctly labeled as *true* but provided hallucination sources, we find that the academic GPTs responded with accurate and relevant citations for all claims when it providing a source.

Our API and UI results demonstrate that GPT-4 is able to provide reliable responses for information retrieval tasks that require scientific knowledge, both for identifying the veracity of a scientific claim and for providing an accurate source to justify its response. However, GPT-4 is stronger at the question answering task (achieving 96% accuracy) than the strict information retrieval task of providing a linked citation (achieving 78% accuracy). Chatbots customized for specific tasks, such as academic writing, improve the reliability of outputs and should be leveraged by users when available.

2 Related Work

Prior work has analyzed ChatGPT models, namely versions 3 and 3.5, in their ability to generate accurate scientific publication references, with the majority of studies focused on medical research (Gravel et al., 2023; Wagner and Ertl-Wagner, 2023; Alkaissi and McFarlane, 2023; Sebo, 2023). Additionally, researchers have analyzed and discussed GPT models' ability to be a reliable tool in scientific communication as an information resource or co-author (Schäfer, 2023; Flanagin et al., 2023; De Angelis et al., 2023; Kasneci et al., 2023; Xames and Shefa, 2023). While researchers acknowledge that GPT models have potential as a resource in academic and scientific writing, several studies highlight its shortcomings on citation generation tasks.

Gravel et al. queried ChatGPT with 20 medical questions derived from research publications, asking for the corresponding citation. ChatGPT's responses contained 59 distinct citations, which were then reviewed by the authors of the original research publications. The authors found that 69% of the citations were fabricated, with 71% of the fabrications having correctly formatted metadata (e.g., year, page numbers, volume number) and known publishers (e.g., MedRxiv and Centers for Disease Control and Prevention) (Gravel et al., 2023). Wagner and Ertl-Wagner prompted ChatGPT-3 with 88 radiology-related questions asking for responses with citations and ChatGPT-3 provided 343 distinct citations across all responses for review. Of

---

[1] `https://openai.com/index/introducing-gpts/`

the references that could be verified, only 24% related to the question (i.e., the publication could be used to support the response) and 64% of the 343 citations appeared to be fabricated by ChatGPT-3 (Wagner and Ertl-Wagner, 2023). Sebo asked ChatGPT-3.5 to provide 10 references to a set of 10 questions related to internal medicine, resulting in 100 citations for review. Of the 100 ChatGPT-3.5 provided citations, 34% were completely incorrect and 40% were partially correct due to error in metadata (e.g., publication year/publisher/etc. was incorrect) (Sebo, 2023).

While these studies are useful in understanding ChatGPT's performance on citation generation, they are limited in scope due to their topics and number of questions. Additionally, these studies query chatbots with highly specialized domain questions without leveraging a chatbot customized for that domain. Our work focuses on extending these studies to a range of 1,326 well-established scientific facts in a more generalized domain, and includes experiments using domain-specific chatbots.

3 Experimental Design

Here we describe the dataset, prompt design, and response evaluation for our experiments.

We experiment with GPT-4's ability to provide accurate bibliographic information for NLG (open-ended question) and information retrieval (close-ended) tasks. Specifically, our objective is to evaluate GPT-4's ability to identify scientific fact and provide accurate (existing and relevant) sources to support its responses, and compare the general GPT-4 performance to domain-specific ChatGPTs.

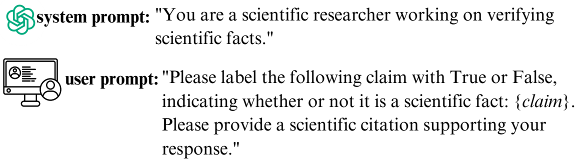
3.1 Scientific Fact Data

We use the OpenBookQA dataset from Gravel et al. (2023), which provides a set of 1,326 scientific facts. Designed for question and answering natural language processing tasks, Grave et al. extracted simple, one sentence scientific fact claims from WorldTree (Jansen et al., 2018), a corpus of 3rd–5th grade science questions with explanations. OpenBookQA contains a wide range of scientific facts (e.g., "a deer lives in a forest", "a landslide is when gravity rapidly moves rocks or soil downhill especially after a rain storm", "the moon reflects sunlight towards the Earth") that do not surpass 5th-grade knowledge, thus we consider these facts to be clear, simple, and general for GPT-4 to label as fact and provide an accurate supporting citation.

3.2 Scientific Claim Prompt

Our API experiments requires two different prompts: 1) an initial prompt to elicit a response identifying if a given claim is scientific fact and a citation supporting the fact (or not fact) identification, and 2) a follow-up prompt asking for verification of the citation and its relevancy to the scientific fact. Additionally, for the system prompt, we assign a scientific research persona in order to produce the most optimal results following OpenAI's prompt engineering documentation (OpenAI, 2023). We access GPT-4 programmatically via the API and set `temperature` $= 0$ for minimal model randomness in GPT-4's output [2].

For the initial prompt of identifying scientific fact and providing a source citation, we give GPT-4 the persona of a scientific researcher who is responsible for verifying scientific facts. In the user prompt, we ask GPT-4 a close-ended question to label a scientific claim as being true or false in order to elicit an automatically parsable response in an information retrieval task; however, we ask an open-ended question to generate a supporting source citation in a NLG task. Figure 2 displays both the system prompt and the user prompt for the first response collection.


system prompt: "You are a scientific researcher working on verifying scientific facts."

user prompt: "Please label the following claim with True or False, indicating whether or not it is a scientific fact: {*claim*}. Please provide a scientific citation supporting your response."

Figure 2: GPT-4 system and user prompts for scientific claim and citation chatbot response.

We alter the persona in the scientific claim and citation verification prompt to include that the system is responsible for verifying scientific fact *and* citations. Figure 3 displays the system and user prompt for this experiment, where two close-ended questions are asked to elicit automatically parsable responses identifying if the citation is relevant to the scientific fact (yes/no) and if the citation exists (real/fake).

We do not implement chain-of-thought prompts in our experiments, but instead treat the validation GPT-4 experiment as a separate task for compari-

---

[2] `https://github.com/autumntoney/GPT4-scifact-citation`

system prompt: "You are a scientific researcher working on verifying scientific facts and citations."

user prompt: "Given the scientific fact and citation below please respond with Yes or No indicating whether or not the citation contains information about the scientific fact. Yes indicates that the citation contains relevant information to the scientific fact and No indicates that the citation does not contains relevant information to the scientific fact. Citation: {*citation*}. Scientific Fact: {*claim*}. The citation came from an unreliable source and identifying its validity is important, please search the internet and respond with Real or Fake indicating if the citation is a real publication, document, or website. Real indicates that the publication, document, or website does exist and Fake indicates that the citation is fabricated."

Figure 3: GPT-4 system and user prompts for verification of the scientific claim and citation accuracy.

son via human annotation. Thus, in our citation validation prompt, we do not state that the citation was generated from GPT-4, but rather an "unreliable source", in order to elicit a more considered evaluation. The first prompt contains a closed-ended prompt for citation generation, representing a NLG task, and the second prompt contains a close-ended prompt, representing an information retrieval task to evaluate two use cases of bibliography generation.

For our UI experiments, we manually interact with customized GPT-4 chatbots. We use the API prompt asking for scientific fact verification and citation generation (Figure 2) and we include a third, informal prompt simulating a real-world, conversational chatbot use-case, shown in Figure 4. In this prompt, the scientific fact is explicitly stated as such to the chatbot, and the user is only asking for a corresponding source for a citation. While user interactions vary widely in conversation style and writing level, we chose a simple conversation prompt to analyze the GPTs, similar to the GPT-4 API prompt experiments.

user prompt: I need a citation for the scientific fact: {*claim*}.

Figure 4: GPT-4 system and user prompts for verification of the scientific claim and citation accuracy.

3.3 Response Evaluation

For the initial API prompt (scientific fact identification and citation generation) we parse GPT-4's response for the true or false label and extract the provided citation in order to evaluate its performance. Next, we take the parsed citation as input to the citation verification prompt and we parse GPT-4's response (citation relevance and existence) for further evaluation. Lastly, we manually verify all citations that GPT-4 provided on the following four criteria: 1) Does the cited source exist?, 2) What type of error occurred (e.g., no error, fabricated source, page not found), 3) What type of source was provided (e.g., textbook, article, URL), and 4) Is the source related to the scientific fact?

We are not concerned with evaluating the consistency of GPT-4's citation formatting, as we did not specify citation style in our prompt. Our evaluation criteria are focused on determining if GPT-4 is able to support its scientific fact identification with accurate (existing and relevant) sources.

Due to many of the generated citations being paywalled or textbooks, we determine relevance to a scientific fact by publicly available information. Thus, even if a full paper is available to read we consider only the title, abstract, and publication venue. For a textbook citation, we consider the general topic that is covered and if the scientific fact falls under that topic. The widest variety of material to review are URLs, as GPT-4 provides links to credible sources (e.g., National Geographic, NOAA, the Oxford Dictionary), but also blog posts, articles, and guides. We evaluate a URL as being accurate if the page exists and contains information relevant to the scientific fact—we do not investigate the credibility of the source itself (i.e., if the URL links to a personal blog). We use this annotation framework for both API and UI GPT responses.

4 Results and Discussion

We evaluate GPT-4's ability to accurately identify scientific fact and provide a relevant and existing citation using the API and UI prompts. Each chatbot experiment involves curating a GPT-4 response dataset from the various prompts and analyzing the responses for accuracy and relevancy.

4.1 GPT-4 API

We first evaluate the results from the first GPT-4 prompt (scientific fact identification and citation generation). GPT-4 accurately identified 96% (1,273 in total) of the claims as being scientific fact. The majority of errors were made in the citation information provided. We display the results in Table 1, listing the total count, percentage incorrect, and the most frequent error by citation type. We distinguish page not found errors from fabrication errors, since we did not investigate if a currently broken url was a historical artifact of the training data for GPT-4 (i.e., if the URL provided was previously valid and potentially a part of the model's ingested knowledge).

Type	Count	% Incorrect	Frequent Error
Article	297	13%	Fabrication
Textbook	600	1%	Fabrication
URL	429	42%	Page Not Found

Table 1: GPT-4 citation responses by source type, with the corresponding count, percentage incorrect, and most frequent error by citation type.

GPT-4 most commonly responded with a textbook citation (45% of citations) and URL (32% of citations), however the URL citations had the highest error rate (42%) compared to the textbook citation error rate (1%), which was the lowest. GPT-4 provided scientific articles with the lowest frequency (22%) and a 13% error rate. This result indicates that GPT-4 has the ability to provide accurate and relevant citations for scientific facts, with the most reliable responses involving a textbook citation, followed by an academic publication.

We analyzed the sources that GPT-4 responded with to assess if it used the same textbooks, website domains, or scientific articles for multiple responses since all scientific facts were derived from grade school knowledge. Table 2 displays the top 10 most commonly cited sources in our GPT-4 API experiments.

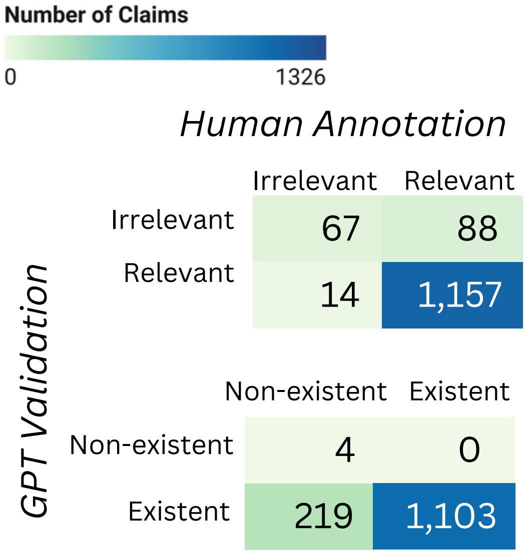
The most commonly referenced textbooks cover the general subjects of physics, biology, meteorology, and earth science. For URL citations, GPT-4 most frequently provided webpages to the National Aeronautics and Space Administration (NASA), Encyclopedia Britannica, and National Geographic. Additionally, we found that the most commonly referenced sources in the GPT-4 responses are reputable citations and could be selected by a user as an accurate reference. While only several textbooks could have been used repeatedly as sources, GPT-4 varies its response with more specific sources using scientific articles and webpages.

Next, we compared the human annotation results with the second GPT-4 prompt (citation validation) results. Figure 5 displays the co-occurrence matrices for citation relevancy and existence. We consider human annotation as the ground truth label since every generated citation was checked manually, thus the GPT-4 responses that disagree with human annotation are considered incorrect responses from GPT-4.

	Citation	Count
1.	Halliday, David, Robert Resnick, and Jearl Walker. Fundamentals of physics. John Wiley & Sons, 2013.	78
2.	National Aeronautics and Space Administration	53
3.	Encyclopedia Britannica	44
4.	National Geographic	35
5.	Raven, Peter H., Ray F. Evert, and Susan E. Eichhorn. Biology of plants. Macmillan, 2005.	33
6.	National Oceanic and Atmospheric Administration	33
7.	National Weather Service	20
8.	Lutgens, Frederick, Edward J. Tarbuck, Redina Herman, and Dennis G. Tasa. The Atmosphere: An Introduction to Meteorology. Pearson, 2017.	13
9.	Marshak, Steve. Earth: portrait of a planet: 5th international student edition. WW Norton & Company, 2015.	13
10.	Smithsonian Museums	10

Table 2: Top 10 most frequently cited source by GPT-4. The organization name is provided for URLs and the MLA-style citation is provided for textbooks.

GPT-4 achieved high performance as an evaluator of citation relevancy, identifying 83% of the irrelevant citations and 93% of the relevant citations correctly. However, GPT-4 did not exhibit the ability to identify citations that were non-existent; it incorrectly claimed that almost all (98%) of the non-existent citations were real. Of the 219 claims that GPT-4 incorrectly identified as being existent 83% were broken URLs. We did not further inves-



	Irrelevant	Relevant
Irrelevant	67	88
Relevant	14	1,157

	Non-existent	Existent
Non-existent	4	0
Existent	219	1,103

Figure 5: GPT-4 and human annotation citation validation results for relevancy and existence.

tigate if the URL was ever active online and could be a temporal error due to the data that GPT-4 was trained on, as opposed to a true hallucination.

During manual review, we did not discover any topical patterns for the set of facts that GPT-4 did not provide accurate citations. Specifically, there was not a distinct subject area (e.g., biology or chemistry), a claim length, or subject level that caused erroneous generated citations. For example the simple fact, "a balloon contains gas", resulted in the following GPT-4 response with a broken URL citation:

> True.
>
> A balloon typically contains gas, such as helium or air. This is a basic principle of balloon inflation. The gas inside the balloon is less dense than the surrounding air, which allows the balloon to float if the gas is lighter than air, like helium.
>
> Reference: Encyclopedia Britannica. (n.d.). Balloon. In Encyclopedia Britannica online. Retrieved from `https://www.britannica.com/technology/balloon-flight-vehicle`

In this error example we find that GPT-4 provided a similar citation to a correct link (`https://www.britannica.com/technology/balloon-flight`), but included an incorrect third word "vehicle". A source error such as this might be difficult for a user to notice, as the URL appears to be accurate and the Encyclopedia Britannica is a reputable source.

While GPT-4 did not always provide accurate and relevant sources for more simple scientific facts like the example above, we find that GPT-4 correctly generated citations for a more complex facts, such as "using tools is a learned characteristic", responding with the following:

> True.
>
> Using tools is indeed a learned characteristic. This is not only true for humans but also for certain animal species. For example, chimpanzees have been observed to learn tool use from their peers, indicating that this is a learned behavior rather than an innate one.
>
> Reference: Whiten, A., Goodall, J., McGrew, W. C., Nishida, T., Reynolds, V., Sugiyama, Y., ... & Boesch, C. (1999). Cultures in chimpanzees. Nature, 399(6737), 682-685.

In this example, GPT-4 processed the scientific fact, correctly assigning the label *true*, and provided an accurate and relevant citation discussing tool use as a learned behavior for chimpanzees. Linking human behavior to chimpanzee behavior is a more human-like reasoning task, highlighting that GPT-4 has ingested knowledge that enables this type of information retrieval performance.

Our experimental API results establish a baseline understanding of how well GPT-4 performs at the task of correctly answering if a scientific claim is true and linking the claim to an academic citation. While GPT-4 did fabricate citations and provide broken URLs, the majority (78%) of generated citations were relevant and existent. Additionally, GPT-4 did not have an underlying pattern to the incorrect generated citations, supporting prior research indicating its hallucinations are difficult to predict. Furthermore, our results suggest that despite having ingested the appropriate knowledge to respond with an accurate answer in an scientific knowledge information retrieval task, GPT-4 is still prone to hallucinate at random.

4.2 Academic GPT UIs

For the UI experiments we select two public GPTs that are customized for academic writing based on their use rankings and reviews on the OpenAI platform: academic assistant pro and scholar AI pro[3]. These public GPTs reflect the GPT-4 chabot a user might select in a real-world scenario, and they are designed specifically for the research task of academic writing. Here, we evaluate if these GPTs are able to correctly identify scientific fact and provide accurate and relevant corresponding citations using the claims that GPT-4 responded with error in the API experiments.

Using the API experiment prompt (Figure 2) that asks for a *true* or *false* label on the scientific claim and a corresponding citation, we prompt both academic GPTs with the 56 claims that GPT-4 mislabeled in the API experiments (See Appendix for full list of claims). The academic assistant pro GPT labeled 38% of the claims as true and the scholar AI pro GPT labeled 39% as true. We included a third label, *gray*, indicating an indecisive GPT response, which did not occur in the API experiments. For example, when prompted about the claim "veins transport nutrients throughout the body", the academic assistant pro GPT stated:

> The statement "veins transport nutrients throughout the body" can be nuanced in scientific literature because, while veins are a critical component of the circulatory system, their primary function is to return deoxygenated blood back to the heart, from where it is sent to the lungs for oxygenation.

The scholar AI GPT provided 10 gray labels and the academic assistant provided only two.

In the same annotation task as the GPT-4 API responses, we annotate the academic GPTs' source type. Table 3 displays the frequencies of the citation type provided by each GPT. Both academic GPTs aligned with the API experiments, providing majority textbook sources; however, the scholar AI pro GPT did not provide a source for 20 of the prompts (35%). We include a source label of organization, as both GPTs provided the general source of the International Astronomical Union as reference to the scientific fact "Pluto is the planet that is ninth closest to the Sun." Similarly, both customized GPTs only provided three distinct sources that did not exist (hallucinated) per GPT respectively(six sources in total from both GPTs), an improvement from the API results.

Using the informal ask for a source given a scientific fact (Figure 4), we sample 50 claims that GPT-4 correctly labeled as *true*, but provided an incorrect citation for (e.g., hallucination or broken URL); see Appendix for list of claims. Table 4 displays the source counts by type.

The academic assistant GPT did not provide a source for one claim ("as the use of a crop increases, the amount of crops planted will increase"), whereas the scholar AI GPT did not provide a source for the majority (76%) of the claims[4]. All sources provided in this prompt experiment were accurate and relevant from both GPTs. The academic assistant responded with textbook sources for 94% of its responses, wheres the scholar AI responded with 75% URL sources (of the 24% of claims it provided a source for). The chatbot UI results strengthen the API finding that GPT-4 has the most reliable results when providing a textbook citation.

In general, we find that using a customized, public GPT provides improved results from prompting GPT-4 via the API. For the application of our study, this result indicates that in a real-world scenario a user can select a GPT to reliably support bibliography curation.

5 Discussion and Limitations

The inability to study the underling algorithms, codebase, and knowledge infrastructure of a GPT model presents a challenge when studying closed-source chatbots. In this work, our goal is to systematically evaluate GPT-4's API and UI performances as reliable tools for a paired task of natural language generation and information retrieval on a domain-specific task— linking scientific claims to relevant and existent sources. A limitation of our results is the lack of validation that can only be fully achieved with the transparency of an open-source model. Additionally, we only query one chatbot (GPT-4) on scientific facts and sources (limited information types). We highlight our main findings and discuss our interpretations of these results.

GPT-4's apparent knowledge acquisition and reliability mimics the real-world. When evaluating the reliability of sources provided, we found that GPT-4 had the most accurate citations when referencing a textbook and the least accurate citations when referencing a URL. This behavior mimics real-world bibliographic curation—a relevant published piece of knowledge is more

---

[3] `https://awesomegpts.vip/`

[4] During experimentation we tested follow-on prompts asking for a citation again, but did not receive any source information.

GPT	Article	Organization	Textbook	URL	No Citation Provided
academic assistant	15	1	38	1	0
scholar AI	12	1	14	10	20

Table 3: Academic GPTs citation responses by source type using the formal prompt asking for scientific fact verification and a corresponding source.

GPT	Article	Organization	Textbook	URL	No Citation Provided
academic assistant	0	0	47	2	1
scholar AI	1	1	1	9	38

Table 4: Academic GPTs citation responses by source type using the informal prompt asking for a source given the scientific fact.

reliable for academic citation than a URL. While we did not further investigate erroneous URLs for their potential historical existence, it appeared that GPT-4 would use a reliable domain name (e.g., `nationalgeographic.com/`) with an incorrect (hallucinated) page reference (e.g., `article/volcanic-landforms-extrusive-intrusive/`). Thus, we hypothesize that GPT-4 has has ingested information on reputable bibliographic sources (e.g., National Geographic) and their corresponding domain, but does not always "retrieve" a correct URL.

Customized GPTs achieve higher performance for the intended (domain-specific) task. OpenAI's description of creating customized GPTs indicates its user-friendly design (no coding required) by stating that all a user needs is to prompt ChatGPT with further instructions or *extra knowledge*. Despite ChatGPT being a closed-source model, it can ingest knowledge via human interaction directly in the UI. Selecting the additional knowledge that a chabot can learn improves the transparency of knowing what the GPT "knows" and also increases the reliability of the chatbot's responses related to the specific information retrieval task. We highlight the fact that GPT models may appear to be poor tools for an information retrieval task like bibliography generation, as discussed in prior research, however GPT models are generative in their nature. Fine-tuning a GPT model with the necessary information for a task will improve its results and reliability, as the knowledge and knowledge sources are identified by the user. Thus, customization for a domain-specific task should be heavily considered when leveraging chatbots as a domain-specific tool.

6 Conclusion

In this paper we evaluated GPT-4's ability to identify scientific fact and generate a citation to support its response. Our experimental design contained two chatbot environments, API and UI, to fully assess GPT-4's performance. We designed prompts that included open-ended (generative) and close-ended (information retrieval) questions in order to test two prompt and response formats. Our experiments are designed to compare how GPT-4 generally performs on a domain-specific task (via the API) and how a GPT-4 chatbot performs (via the UI) when customized for use in the specified domain.

Using the API, we find that in general, GPT-4 performs well on identifying scientific fact and providing reliable sources. For the citation generation task, we find that GPT-4 provided relevant and existent academic citations with 78% accuracy. For the information retrieval tasks, we find that GPT-4 is able to identify scientific fact with 96% accuracy and determine the relevancy of citations with 83% accuracy for irrelevant citaitons and 93% accuracy for relevant citations. GPT-4 had the worst performance when determining if a citation existed, with the majority of its error as labeling broken URLs as existent. In the UI experiments, we find that using public GPTs customized for academic writing improved the API results in both scientific fact identification and source generation. However, we did identify discrepancies in chatbot performances between the two GPTs, with one chatbot's outputs resulting in the majority (76%) not containing a source when being explicitly asked for one.

Overall, we find GPT-4 to be a useful information gathering tool for general scientific knowledge.

Our experiments suggest that a user should select or design a customized chatbot for domain-specific tasks for improved utility.

References

Hussam Alkaissi and Samy I McFarlane. 2023. Artificial hallucinations in chatgpt: implications in scientific writing. *Cureus*, 15(2).

Luigi De Angelis, Francesco Baglivo, Guglielmo Arzilli, Gaetano Pierpaolo Privitera, Paolo Ferragina, Alberto Eugenio Tozzi, and Caterina Rizzo. 2023. Chatgpt and the rise of large language models: the new ai-driven infodemic threat in public health. *Frontiers in Public Health*, 11:1166120.

Nouha Dziri, Sivan Milton, Mo Yu, Osmar Zaiane, and Siva Reddy. 2022. On the origin of hallucinations in conversational models: Is it the datasets or the models? In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5271–5285, Seattle, United States. Association for Computational Linguistics.

Annette Flanagin, Kirsten Bibbins-Domingo, Michael Berkwits, and Stacy L Christiansen. 2023. Nonhuman "authors" and implications for the integrity of scientific publication and medical knowledge. *Jama*, 329(8):637–639.

Jocelyn Gravel, Madeleine D'Amours-Gravel, and Esli Osmanlliu. 2023. Learning to fake it: limited responses and fabricated references provided by chatgpt for medical questions. *Mayo Clinic Proceedings: Digital Health*, 1(3):226–234.

Peter Jansen, Elizabeth Wainwright, Steven Marmorstein, and Clayton Morrison. 2018. WorldTree: A corpus of explanation graphs for elementary science questions supporting multi-hop inference. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

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.

Enkelejda Kasneci, Kathrin Seßler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke Hüllermeier, et al. 2023. Chatgpt for good? on opportunities and challenges of large language models for education. *Learning and individual differences*, 103:102274.

OpenAI. 2023. Tactic: Ask the model to adopt a persona. `https://platform.openai.com/docs/guides/prompt-engineering/tactic-ask-the-model-to-adopt-a-persona`. Accessed: 2023-11-20.

Mike S Schäfer. 2023. The notorious gpt: science communication in the age of artificial intelligence. *Journal of Science Communication*, 22(2):Y02.

Paul Sebo. 2023. How accurate are the references generated by chatgpt in internal medicine? *Internal and Emergency Medicine*, pages 1–3.

Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. 2021. Retrieval augmentation reduces hallucination in conversation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3784–3803, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Matthias W Wagner and Birgit B Ertl-Wagner. 2023. Accuracy of information and references using chatgpt-3 for retrieval of clinical radiological information. *Canadian Association of Radiologists Journal*, page 08465371231171125.

Md Doulotuzzaman Xames and Jannatul Shefa. 2023. Chatgpt for research and publication: Opportunities and challenges. *Available at SSRN 4381803*.

A GPT-4 UI Scientific Claim Sets

We provide the sets of claims used in the UI experiments, which provide insight into the claims that resulted in error responses from the API experiments. Table 5 lists all scientific claims that GPT-4 incorrectly labeled *false* in the API experiments and Table 6 lists a random sample of 50 claims that GPT-4 correctly labeled as *true* in the API experiments, but provided inaccurate sources for.

Table 5: Set of 56 scientific facts that GPT-4 mislabled as *false* in API experiments and are used in the UI experiments.

limestone is formed by water evaporating from a solution of water and minerals
if a weed is pulled then that weed is destroyed
as water increases in an environment, the population of aquatic animals will increase
as ability to preserve food increases, the ability to transport food increases
cold environments are usually white in color from being covered in snow
as air pressure decreases, the chance of rain will increase
as the available water in an environment increases, the populations of organisms in that environment will increase
a complete electrical circuit is a source of electrical energy
if a tree falls then that tree is dead
decreasing something negative has a positive impact on a thing
precipitation is when snow fall from clouds to the Earth
poisonous darts are used for defense by sea anemones
if an animal relies on plants for food then that animal must store enough food to last through the winter
force causes the speed of an object to decrease
if a hot object touches a cold substance then that substance will likely cool
as moisture of an object decreases, the friction of that object against another object will increase
as the size of a flower increases, the number of pollinators it will attract increases
the Earth revolving around the Sun causes the seasons to change on its axis
a different moon phase occurs once per week
the sun is located directly overhead at noon
food is a source of energy for plants
as the number of eggs laid by an animal increases, the number of eggs that hatch will increase
if an object is blue then that object reflects only blue light
bees eat pollen
iron is always magnetic
mountains are formed by volcanoes
the moon does not contain water
cracking something usually has a negative impact on that something
omnivores are predators
as the time a tool lasts increases, the number of tools discarded will decrease
hunting requires seeing prey
as the size of the eyes of an animal increases, the ability of that animal to see will usually increase
clear weather means sunny weather
the increase of something required by an organism has a positive impact on that organism 's survival
cold environments contain few organisms
adding salt to a solid decreases the freezing point of that solid
water is in the solid state, called ice, for temperatures between 0 and 0 F
if a cell can not specialize then that cell must perform all life functions
as number of organisms in a group increases, the chance of survival of each organism will increase
boiling is when liquids are heated above their boiling point
breathing is when a lung converts from oxygen in air into oxygen in blood
as force exerted on an object increases, distance travelled will increase
an animal usually requires a warm body temperature for survival
a plant requires soil for to grow
as the activity of an animal increases, the amount of water in an animal 's body in that environment will decrease
if something is outside during the day then that something will receive sunlight
the moon rising occurs once per day
as the weight of an animal decreases, that animal will fly more easily
pollination requires pollinating animals
the condition of the parts of an organism are acquired characteristics
carnivores only eat animals
veins transport nutrients throughout the body
the Earth absorbs more energy than it loses
as the thickness of an object increases, the resistance to damage of that object will increase
the Earth revolving around the Sun causes the seasons to occur on its axis
Pluto is the planet that is ninth closest to the Sun

Table 6: Set of 50 randomly sampled scientific facts for UI experiments.

as the use of a crop increases, the amount of crops planted will increase	a scar is an acquired characteristic
magnetism can cause objects to repel each other	a sea turtle lives in the ocean
a spider web is used to capture food by spiders	a renewable resource can be replaced
a greenhouse is used to protect plants by keeping them warm	the tide cycle regularly occurs twice per day
water is an electrical conductor	tectonic plates being pushed together causes earthquakes
crumple means change shape from smooth into	compacted by physical force
sunlight contains ultraviolet light	the Earth revolves around the sun
meters m are a unit used for measuring distance generally used for values between 1 and 1000	the slope of the land causes a river to flow in a particular direction
natural magnetism is used for pointing north by a compass	soil is formed by weathering
if a mineral can be scratched by a fingernail then that mineral is soft	if a substance absorbs solar energy then that substance will increase in temperature
breath contains water vapor	weathering usually occurs over a period of many years
a star is a source of light through nuclear reactions	a star is made of gases
a reflector is used to reflect light especially on vehicles	high means great in altitude
a flashlight requires a source of electricity to produce light	endangered means low in population
a Rotation of the Earth on Earth ’s axis takes 1 day	An example of an inherited behavior is a bird building a nest
a balloon contains gas	the sun causes water to evaporate more quickly by adding heat
a bubble contains gas	the sun is the source of solar energy called sunlight
winter in the Northern Hemisphere is during the summer in the Southern Hemisphere	coal mine is a source of coal under the ground
In the food chain process some types of plankton have the role of producer	as time spent taking a shower decreases, water used will decrease
a compass ’s needle lines up with Earth ’s magnetic poles	a stopwatch is used to measure time
coal is used to produce electricity by burning in coal-fire power stations	arteries transport nutrients throughout the body
An example of a reproductive behavior is salmon returning to their birthplace to lay their eggs	a graduated cylinder is a kind of instrument for measuring volume of liquids or objects
a rainbow is formed by refraction of light by splitting light into all different colors	fossil fuels forming occurs over a period of 300000000 years which is considered a very long time to a human
as lightness in color of an object increases, the ability of that object to reflect light will increase the stars in the night sky are very far away from the Earth	wind causes erosion
the sun is located directly overhead at noon	a solar panel converts sunlight into electricity