This Reference Does Not Exist: An Exploration of LLM Citation Accuracy and Relevance

Courtni Byun and **Piper Vasicek** and **Kevin Seppi**
Brigham Young University
Provo, USA

Abstract

Citations are a fundamental and indispensable part of research writing. They provide support and lend credibility to research findings. Recent GPT-fueled interest in large language models (LLMs) has shone a spotlight on the capabilities and limitations of these models when generating relevant citations for a document. Recent work has focused largely on title and author accuracy. We underline this effort and expand on it with a preliminary exploration in relevance of model-recommended citations. We define three citation-recommendation tasks. We also collect and annotate a dataset of model-recommended citations for those tasks. We find that GPT-4 largely outperforms earlier models on both author and title accuracy in two markedly different CS venues, but may not recommend references that are more relevant than those recommended by the earlier models. The two venues we compare are CHI and EMNLP. All models appear to perform better at recommending EMNLP papers than CHI papers.

1 Introduction

Citations are a common feature of research writing. They lend credibility to claims and can help identify gaps in prior research. They can also provide a chain of ideas from prior work to a research task.

The last year has seen a drastic increase of interest about large language models (LLMs). ChatGPT (OpenAI, 2022) has opened the eyes of the general public to the potential of LLMs. ChatGPT and its related GPT-X LLMs are being applied to a growing array of tasks (Araoz, 2020; OpenAI, 2023; Byun et al., 2023; Xiao et al., 2023).

One task that has drawn both interest and ire is that of using LLMs to identify citations for a topic. Several recent blog posts and articles have warned of ChatGPT's hallucinated references (Welborn, 2023; Wilkinson, 2023; Neumeister, 2023). We build on recent work to assess the problem.

2 Related Works

Various citation recommendation systems exist, relying on an array of NLP and information retrieval (IR) approaches. Farber and Jatowt (2020) offer a thorough survey of automated citation recommendation approaches.

More recently, use of LLMs has been explored, leading to discussion of the tendency LLMs have to hallucinate output. Day (2023) offered an early exploration of hallucinated references by ChatGPT. They assessed references output by ChatGPT based on accuracy of journal name, volume, issue and page number and found the model incapable of generating any valid references.

On the other hand, MW Wagner (2023) found ChatGPT capable of some accuracy when answering questions about clinical radiological sources.

A letter of warning from McGowan et al. (2023) discussed fabricated references from both ChatGPT and Google's Bard (Manyika, 2023) in psychiatry literature. They found real authors are often included, even when a paper title is fabricated. They also raised the alarm on the possibility of fake references entering into automated indexes.

Gravel et al. (2023) found ChatGPT output in response to medical reference questions was of limited quality, but that references offered by the model were deceptively realistic.

Orduna-Malea and Cabezas-Clavijo (2023) compared ChatGPT and Bard 2.0 citations in English, Spanish, and Italian. They explored reasons for fabricated citations and steps to address the issue.

Taylor et al. (2022) fine-tuned their own LLM, Galactica, and assessed it on three citation generation tasks. They found LLM accuracy for citation generation appears to improve with scale.

Finally, Agrawal et al. (2023) found LLMs tend to hallucinate different authors of fabricated references in multiple independent query sessions, but consistently hallucinate authors in the same session.

Proceedings of the Third Workshop on Bridging Human–Computer Interaction and Natural Language Processing (3rd HCI+NLP Workshop), pages 28–39

They compared accuracy on GPT text-davinci-003, ChatGPT, and GPT-4.

Previous work has primarily focused on metrics related to accuracy of information. While understanding accuracy is important, accurate citations that are irrelevant will still be of little use to researchers. In this work we still assess accuracy, but we also offer a preliminary assessment of the relevance of citations identified by three models.

3 Methods

We define three citation recommendation tasks, intended to model aspects of academic writing that could be supplemented by use of LLMs.

3.1 Models

We compare performance between three GPT-X models: GPT-3 text-davinci-003 (GPT-3), GPT-3.5-turbo (GPT-3.5), and GPT-4. All model hyperparameters used can be found in Appendix A

3.2 Tasks

We define three tasks, each with a unique prompt. The full prompt evolutions and all final prompt designs can be found in Appendices B and C.

3.2.1 Abstract→Citations List Task

This task asks the model generate a list of relevant sources a researcher could explore and incorporate into their paper (target paper). We provide the models with a prompt including a paper title and its accompanying abstract and request the model generate ten relevant citations to be used in the target paper. We request citations in APA format because it is common and having all citations in a consistent format aids in annotating and analysing the data. See Figure 1 for prompt template.

You are an [NLP or HCI] researcher working on a paper to submit to [EMNLP or CHI].
The paper you are working on is titled: [PAPER TITLE]

The abstract for your paper is: [PAPER ABSTRACT]

List 10 relevant papers you could cite in your Related Works section. Write each citation in APA format.

Figure 1: Prompt template for Abstract→Citations List task.

3.2.2 Abstract→Related Works Task

The goal of this task is to explore how well the models identify relevant citations when also asked to discuss them, without the textual scaffolding of a provided Related Works section. The prompt for this task builds on the prompt for the first task, but replaces the final section with: *Write a Related Works section for your paper. Include 10 in-text citations. Also include a list of those citations with each citation in APA format.*

3.2.3 Discussion→Supported Discussion Task

The goal of this task is to test model citation recommendation and discussion when some textual scaffolding is provided. The prompt for this task builds on the prompt for the first task by including the target paper title and abstract in the prompt, but the prompt additionally includes a portion of the results discussion. The final section of the prompt, which follows the discussion, is changed to: *Rewrite the Discussion section to include 10 in-text citations. Also include a list of those citations with each citation in APA format.*

3.3 Dataset

We randomly sampled twenty papers from two top-tier, but different venues, CHI (HCI) and EMNLP (NLP). Ten papers were randomly sampled from recent publications of each venue. See Appendix E for the list of papers. The paper title, abstract, and discussion of results were extracted for each paper. For some papers this was taken from the Results section and for others, the Discussion section. Some discussions were too long for the models. For these, we extracted only the first paragraph of each section within the discussion. We also extracted the bibliography from each paper.

This information was used to fill the prompt templates, which were then input to each model. The output was collected and the citations extracted. While we requested citations in APA, the models sometimes used different formatting. We reformatted each citation to ensure it was in APA. Some model-generated citations lacked titles. These we exclude from our final dataset because we cannot verify whether they are real papers. Our final dataset has 1616 annotated citations.

We used Google Scholar to check whether each model-recommended citation was for a real paper.

Abstract → Citations List				**Abstract → Related Works**				**Discussion → Supported Discussion**			
	Title Accuracy				Title Accuracy				Title Accuracy		
	HCI	NLP	Total		HCI	NLP	Total		HCI	NLP	Total
GPT-3	24.47%	**48.98%**	36.98%	GPT-3	0.00%	**36.84%**	19.44%	GPT-3	**34.88%**	18.37%	26.09%
GPT-3.5	28.00%	**56.00%**	42.00%	GPT-3.5	13.51%	**50.54%**	30.39%	GPT-3.5	12.96%	**51.43%**	22.38%
GPT-4	54.00%	**78.00%**	66.00%	GPT-4	68.87%	**75.45%**	72.22%	GPT-4	**47.15%**	25.74%	37.50%
	Author Precision				Author Precision				Author Precision		
	HCI	NLP	Total		HCI	NLP	Total		HCI	NLP	Total
GPT-3	**75.00%**	70.29%	71.82%	GPT-3	-	**72.71%**	72.71%	GPT-3	**55.53%**	41.67%	50.33%
GPT-3.5	**81.46%**	76.16%	77.93%	GPT-3.5	62.13%	**73.23%**	70.55%	GPT-3.5	66.86%	**68.94%**	60.03%
GPT-4	88.11%	**89.01%**	88.64%	GPT-4	82.07%	**84.78%**	83.51%	GPT-4	**82.14%**	63.15%	76.26%
	Author Recall				Author Recall				Author Recall		
	HCI	NLP	Total		HCI	NLP	Total		HCI	NLP	Total
GPT-3	**72.65%**	41.54%	51.62%	GPT-3	-	**70.89%**	70.89%	GPT-3	**37.93%**	33.33%	36.21%
GPT-3.5	**81.46%**	70.88%	73.95%	GPT-3.5	61.73%	**71.02%**	68.77%	GPT-3.5	64.00%	**68.11%**	66.31%
GPT-4	88.11%	**89.01%**	88.64%	GPT-4	82.07%	**84.78%**	83.51%	GPT-4	**82.14%**	63.15%	76.26%
	Year Accuracy				Year Accuracy				Year Accuracy		
	HCI	NLP	Total		HCI	NLP	Total		HCI	NLP	Total
GPT-3	1.43	**0.31**	0.68	GPT-3	-	**1.57**	1.57	GPT-3	2.40	**0.89**	1.83
GPT-3.5	**0.29**	0.46	0.40	GPT-3.5	1.06	**0.55**	0.68	GPT-3.5	6.29	**0.61**	3.09
GPT-4	1.44	**1.26**	1.33	GPT-4	1.59	**0.70**	1.12	GPT-4	**2.59**	3.08	2.74

Table 1: Accuracy scores for each model, for each of the tasks, broken out between HCI and NLP.

Nearly all real papers had an exact match in the first three results of a page, so we restricted our search to the first page of results. Petiska (2023) found that ChatGPT tends to use Google Scholar citation counts when recommending citations, so relying only on Google Scholar results should be sufficient. A citation was marked as fabricated if an exact match was not found in the first page of Google Scholar results. A citation with an exact match was marked as a real paper and the APA citation for the true paper was collected and checked against the citation generated by the model.

We automatically compared information in the citations generated by the models against the information collected from the real papers. We collected information for how many citations were fabricated vs real. We also calculated author precision and recall between the authors in a recommended citation and those on real papers. We tested relevance by checking whether a real paper's title was found in the bibliographies of the target papers and whether the authors of the model-generated citations were found in the bibliographies of the target papers.

While more elaborate metrics for determining citation relevance exist (Belter, 2017; Boyack and Klavans, 2010), these often involve creating a network of citations. The overlap between citations is then checked. This includes overlap with the target papers. However, we needed target papers that were excluded from the models' training data, which meant very recent papers that had not been cited yet. This meant we needed a different metric for relevance. We focus on several basic metrics based on the idea that if there is overlap between papers models recommend and papers authors actually use, then those papers and authors that overlap must be relevant. This means true relevance could be higher, but our strict definition should offer a reasonable exploratory view.

4 Results

Accuracy results can be found in Tables 1 and 3, while relevance results can be found in Table 2.

4.1 Accuracy

Title Accuracy is the percentage of citations recommended by the model that had real paper titles. Author Precision, Author Recall, and Year Accuracy were only calculated for citations of real papers. Year Accuracy was calculated by taking the absolute value of the year a real paper was published, minus the year in the model-recommended citation.

As seen in Table 1, the models tend to perform better on NLP papers, particularly with respect to paper titles. This is reiterated by the results in Table 3, where for nearly every model, for every task there appears to be a significant difference between NLP and HCI papers on this metric.

The distinction is less clear for other metrics. For example both GPT-3 and GPT-3.5 perform better for HCI papers in terms of Author Recall for the Abstract→Citations List task and GPT-4 performs better for HCI papers in terms of Author Precision for the Discussion→Supported Discussion task.

GPT-4 typically outperforms the other models in terms of accuracy, which is unsurprising given the findings of Taylor et al. (2022) that LLM citation accuracy improves with model scale. There are,

Abstract → Citations List				**Abstract → Related Works**				**Discussion → Supported Discussion**			
Title Relevance				Title Relevance				Title Relevance			
	HCI	NLP	Total		HCI	NLP	Total		HCI	NLP	Total
GPT-3	0.17%	**22.92%**	16.90%	GPT-3	-	**10.71%**	10.71%	GPT-3	0.12%	**22.22%**	12.5%
GPT-3.5	0.18%	**25.00%**	17.86%	GPT-3.5	0.24%	**29.79%**	24.19%	GPT-3.5	0.25%	**33.33%**	25.00%
GPT-4	0.20%	**29.49%**	20.45%	GPT-4	0.17%	**27.71%**	17.31%	GPT-4	0.08%	**19.23%**	8.33%
Real Author Relevance				Real Author Relevance				Real Author Relevance			
	HCI	NLP	Total		HCI	NLP	Total		HCI	NLP	Total
GPT-3	4.35%	**6.25%**	5.63%	GPT-3	-	**3.57%**	3.57%	GPT-3	**6.67%**	0.00%	4.17%
GPT-3.5	**7.14%**	5.36%	5.95%	GPT-3.5	0.00%	**10.64%**	8.06%	GPT-3.5	0.00%	**11.11%**	6.25%
GPT-4	3.70%	**3.85%**	3.79%	GPT-4	4.11%	**6.02%**	5.13%	GPT-4	1.72%	**3.85%**	2.38%
False Author Relevance				False Author Relevance				False Author Relevance			
	HCI	NLP	Total		HCI	NLP	Total		HCI	NLP	Total
GPT-3	**13.04%**	8.33%	9.86%	GPT-3	-	**25.00%**	25.00%	GPT-3	**26.67%**	0.00%	16.67%
GPT-3.5	7.14%	**8.93%**	8.33%	GPT-3.5	13.33%	**17.02%**	16.13%	GPT-3.5	21.43%	**27.78%**	25.00%
GPT-4	16.67%	**25.64%**	21.97%	GPT-4	**28.77%**	19.28%	23.72%	GPT-4	6.9%	**11.54%**	8.33%

Table 2: Relevance scores for each model, for each task, broken out between HCI and NLP papers.

Abstract→Citations List						
Title Accuracy Significance						
	HCI		NLP			
	Mean	SD	Mean	SD	t-statistic	p-value
GPT-3	0.24	0.43	0.49	0.50	-3.62	**0.00**
GPT-3.5	0.28	0.45	0.56	0.50	-4.16	**0.00**
GPT-4	0.54	0.50	0.78	0.41	-3.68	**0.00**
Abstract→Related Works						
Title Accuracy Significance						
	HCI		NLP			
	Mean	SD	Mean	SD	t-statistic	p-value
GPT-3	0.00	0.00	0.37	0.48	-6.25	**0.00**
GPT-3.5	0.14	0.34	0.51	0.50	-6.22	**0.00**
GPT-4	0.69	0.46	0.75	0.43	-1.08	0.28
Discussion→Supported Discussion						
Title Accuracy Significance						
	HCI		NLP			
	Mean	SD	Mean	SD	t-statistic	p-value
GPT-3	0.35	0.48	0.18	0.39	1.81	0.07
GPT-3.5	0.13	0.34	0.51	0.50	-5.13	**0.00**
GPT-4	0.47	0.50	0.26	0.44	3.36	**0.00**

Table 3: Two-sample t-tests for title accuracy on HCI vs NLP papers. Calculated via SciPy and NumPy (Virtanen et al., 2020; Harris et al., 2020).

however, exceptions to this. For example, GPT-3.5 outperforms GPT-4 on Title Accuracy, Author Precision, and Author Recall for the NLP papers on the Discussion→Supported Discussion task.

The models appear to struggle with the Discussion→Supported Discussion task. This could be due to our poor prompt design for this task. CHI papers typically include a separate Discussion section, while EMNLP papers often include a discussion of results with the Results section. We distinctly asked models to support our *Discussion* sections. Future research could explore whether changing *Discussion* to *Results* in the prompt could yield better results for NLP papers.

4.2 Relevance

Title Relevance reports the percentage of real papers cited in the target paper. Real Author Relevance reports the percentage of authors from a model-recommended citation that were real authors on that paper and who had a paper cited in the target paper. False Author Relevance reports the percentage of authors from a model-recommended citation that were not real authors on that paper, but who had papers cited in the target paper.

In terms of relevance, we again see better performance for NLP papers in terms of title relevance. The distinction becomes less clear for other metrics. For example, GPT-4 on the Abstract→Related Works task and False Author Relevance. However, there does not appear to be a large difference between models. In multiple instances the older models perform better than GPT-4, for example GPT-3 for the Abstract→Related Works task on the False Author Relevance metric for NLP papers and both GPT-3 and GPT-3.5 on the Discussion→Supported Discussion task on all relevance metrics.

5 Conclusion

We evaluated GPT-3, GPT-3.5, and GPT-4 on three different citation recommendation tasks and compared them across two research disciplines. We found contrasts in terms of relevance and accuracy between those disciplines. This is important because individuals outside of NLP are beginning to use these models in their research. It is important for researchers from other disciplines to recognize these models' limitations for their disciplines.

Finally, while GPT-4 typically outperforms previous models on accuracy, it does not clearly perform better in terms of relevance.

6 Limitations

While 1616 citations seems like enough for a thorough run of statistical tests, this is not the case. Due to how poorly GPT-3 and GPT-3.5 perform on many of the tasks and how many ways we split the data, several of our sample sizes are slightly under 30, with the smallest being 24. We have run significance tests comparing performance between models and between HCI and NLP papers for other metrics, but considering the small sample sizes of some of the groups, we felt the limited space of this short paper would be best utilized reporting our other results.

Our largest sample sizes are for the Title Accuracy metric because this included all citations, while the other metrics excluded citations for papers that did not exist. This is why we only report significance results for Title Accuracy between HCI and NLP papers. We exclude our significance results for Title Accuracy between models due to the length limitations of this paper. Previous research has shown a difference between models of different sizes. Our results reiterate those findings.

We also did not compare accuracy of other citation information, like page numbers, publication venues, and URLs. Preliminary tests showed much worse model performance on these citation features. We chose to focus on the features the models appeared to recreate more accurately. We leave exploration of these other features to future work.

Additionally, due to the inherently messy nature of text data, some aspects of data collection and curation were done manually. While we did multiple checks at each step of the process to maintain quality, there could still be errors we did not catch.

We also relied on Google Scholar results to determine veracity of citation titles. It is possible that some of the citations marked as fabricated could be real papers that did not show up on the first page of results.

References

Ayush Agrawal, Mirac Suzgun, Lester Mackey, and Adam Tauman Kalai. 2023. Do language models know when they're hallucinating references?

Manuel Araoz. 2020. Openai's gpt-3 may be the biggest thing since bitcoin.

Christopher Belter. 2017. A relevance ranking method for citation-based search results. *Scientometrics*.

Alemitu Bezabih, Kathrin Gerling, Workeabeba Abebe, and Vero Vanden Abeele. 2023. Challenges and opportunities for interactive technology to support parents of hiv-positive children in ethiopia in the disclosure process. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–17.

Avinash Bhat, Austin Coursey, Grace Hu, Sixian Li, Nadia Nahar, Shurui Zhou, Christian Kästner, and Jin LC Guo. 2023. Aspirations and practice of ml model documentation: Moving the needle with nudging and traceability. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–17.

Kevin Boyack and Richard Klavans. 2010. Co-citation analysis, bibliographic coupling, and direct citation: Which citation approach represents the research front most accurately? *Journal of the American Society for Information Science and Technology*, 61:2389–2404.

Courtni Byun, Piper Vasicek, and Kevin Seppi. 2023. Dispensing with humans in human-computer interaction research. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI EA '23, New York, NY, USA. Association for Computing Machinery.

Rémi Cardon, Adrien Bibal, Rodrigo Wilkens, David Alfter, Magali Norré, Adeline Müller, Watrin Patrick, and Thomas François. 2022. Linguistic corpus annotation for automatic text simplification evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1842–1866.

Terence Day. 2023. A preliminary investigation of fake peer-reviewed citations and references generated by chatgpt. *The Professional Geographer*, 0(0):1–4.

Shengda Fan, Shasha Mo, and Jianwei Niu. 2022. Boosting document-level relation extraction by mining and injecting logical rules. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10311–10323.

M. Farber and A. Jatowt. 2020. Citation recommendation: Approaches and datasets. *International Journal on Digital Libraries*, 21:375–405.

Dan Friedman, Alexander Wettig, and Danqi Chen. 2022. Finding dataset shortcuts with grammar induction. *arXiv preprint arXiv:2210.11560*.

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.

Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew

Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. 2020. Array programming with NumPy. *Nature*, 585(7825):357–362.

Franziska Herbert, Steffen Becker, Leonie Schaewitz, Jonas Hielscher, Marvin Kowalewski, Angela Sasse, Yasemin Acar, and Markus Dürmuth. 2023. A world full of privacy and security (mis) conceptions? findings of a representative survey in 12 countries. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–23.

Younghoon Jeong, Juhyun Oh, Jaimeen Ahn, Jongwon Lee, Jihyung Moon, Sungjoon Park, and Alice Oh. 2022. Kold: korean offensive language dataset. *arXiv preprint arXiv:2205.11315*.

Tianyu Jiang and Ellen Riloff. 2022. Identifying physical object use in sentences. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11362–11372.

Shuyang Li, Yufei Li, Jianmo Ni, and Julian McAuley. 2021. Share: a system for hierarchical assistive recipe editing. *arXiv preprint arXiv:2105.08185*.

Chang Liu, Arif Usta, Jian Zhao, and Semih Salihoglu. 2023. Governor: Turning open government data portals into interactive databases. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–16.

Zhenghao Liu, Han Zhang, Chenyan Xiong, Zhiyuan Liu, Yu Gu, and Xiaohua Li. 2022. Dimension reduction for efficient dense retrieval via conditional autoencoder. *arXiv preprint arXiv:2205.03284*.

James Manyika. 2023. An overview of bard: an early experiment with generative ai.

Alessia McGowan, Yunlai Gui, Matthew Dobbs, Sophia Shuster, Matthew Cotter, Alexandria Selloni, Marianne Goodman, Agrima Srivastava, Guillermo A. Cecchi, and Cheryl M. Corcoran. 2023. Chatgpt and bard exhibit spontaneous citation fabrication during psychiatry literature search. *Psychiatry Research*, 326:115334.

Marie Muehlhaus, Marion Koelle, Artin Saberpour, and Jürgen Steimle. 2023. I need a third arm! eliciting body-based interactions with a wearable robotic arm. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–15.

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

Larry Neumeister. 2023. Lawyers blame chatgpt for tricking them into citing bogus case law. *KSL*.

OpenAI. 2022. Chatgpt: Optimizing language models for dialogue.

OpenAI. 2023. Gpt-4 technical report.

Enrique Orduna-Malea and Alvaro Cabezas-Clavijo. 2023. Chatgpt and the potential growing of ghost bibliographic references. *Scientometrics*, 128.

Eduard Petiska. 2023. Chatgpt cites the most-cited articles and journals, relying solely on google scholar's citation counts. as a result, ai may amplify the matthew effect in environmental science.

Ananditha Raghunath, Laurel Krovetz, Hosea Mpogole, Henry Mulisa, Brian Dillon, and Richard Anderson. 2023. From grasshoppers to secondhand cars: Understanding the smartphone-enabled marketplace in peri-urban tanzania. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–13.

Jun Rekimoto. 2023. Wesper: Zero-shot and realtime whisper to normal voice conversion for whisper-based speech interactions. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–12.

Alia Saad, Kian Izadi, Anam Ahmad Khan, Pascal Knierim, Stefan Schneegass, Florian Alt, and Yomna Abdelrahman. 2023. Hotfoot: Foot-based user identification using thermal imaging. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–13.

Ross Taylor, Marcin Kardas, Guillem Cucurull, Thomas Scialom, Anthony Hartshorn, Elvis Saravia, Andrew Poulton, Viktor Kerkez, and Robert Stojnic. 2022. Galactica: A large language model for science.

Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. 2020. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272.

Eitan Wagner, Renana Keydar, Amit Pinchevski, and Omri Abend. 2022. Topical segmentation of spoken narratives: A test case on holocaust survivor testimonies. *arXiv preprint arXiv:2210.13783*.

Zhongwei Wan, Yichun Yin, Wei Zhang, Jiaxin Shi, Lifeng Shang, Guangyong Chen, Xin Jiang, and Qun Liu. 2022. G-map: general memory-augmented pre-trained language model for domain tasks. *arXiv preprint arXiv:2212.03613*.

Aaron Welborn. 2023. Chatgpt and fake citations.

Jordan White, William Odom, Nico Brand, and Ce Zhong. 2023. Memory tracer & memory compass: Investigating personal location histories as a design material for everyday reminiscence. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–19.

D Wilkinson. 2023. Be careful... chatgpt appears to be making up academic references.

Ziang Xiao, Xingdi Yuan, Q. Vera Liao, Rania Abdelghani, and Pierre-Yves Oudeyer. 2023. Supporting qualitative analysis with large language models: Combining codebook with gpt-3 for deductive coding. In *Companion Proceedings of the 28th International Conference on Intelligent User Interfaces*, IUI '23 Companion, page 75–78, New York, NY, USA. Association for Computing Machinery.

Ashley Ge Zhang, Yan Chen, and Steve Oney. 2023. Vizprog: Identifying misunderstandings by visualizing students' coding progress. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–16.

Shujian Zhang, Chengyue Gong, and Xingchao Liu. 2022. Passage-mask: A learnable regularization strategy for retriever-reader models. *arXiv preprint arXiv:2211.00915*.

A Hyperparameters

- Temperature: 0.0
- Top P: 1
- Frequency Penalty: 0.5
- Presence Penalty: 0.5
- Maximum Tokens: 2000

We chose a temperature of 0 because, while a temperature of 0 does not guarantee identical output each time, it does increase the likelihood of very similar output. This was the best option available at the time for generating reproducible results. We used 0.5 for both frequency and presence penalties because both GPT-3 and GPT-3.5 are prone to repeating citations when they are set to 0.

B Prompt Engineering

The following are the various prompt evolutions we used before settling on our final prompt designs.

We went through several iterations of prompt design for each of the three tasks in this paper. The prompt variations were primarily focused around the request portion of the prompt. All prompts included either a CHI or EMNLP paper title and abstract. The Results→Supported Results task prompts also included discussion from the same CHI or EMNLP prompt paper.

All of the prompts in this subsection follow the GPT-3 design. The main difference between the GPT-3 and newer model prompts was a change to a first person perspective. We did not ultimately include GPT-3 in our results for the Abstract→Related Works and Results→Supported Results tasks because the final prompt design was too long for the GPT-3 limited context length. However, GPT-3 was included and evaluated on earlier variations of prompts for those tasks. We found GPT-3 was virtually incapable of identifying any citations of real papers for the Abstract→Related Works and Results→Supported Results tasks, even for prompt designs short enough to fit the GPT-3 context.

B.1 Abstract→Citations Prompt Evolution

Our initial prompt design for the Abstract→Citations task used the following format:

> *You are an [HCI or NLP] researcher working on a paper to submit to [CHI or EMNLP].*
>
> *The paper you are working on is titled: [PAPER TITLE]*
>
> *The abstract for your paper is: [PAPER ABSTRACT]*
>
> *Five relevant papers you could cite in your related works sections are:*

We found the models have a tendency to cite older sources, so we next adjusted the prompt to request only *recent* citations. We updated the prompt to the following, with the changed portion in bold:

> *You are an [HCI or NLP] researcher working on a paper to submit to [CHI or EMNLP].*
>
> *The paper you are working on is titled: [PAPER TITLE]*
>
> *The abstract for your paper is: [PAPER ABSTRACT]*
>
> ***Five relevant papers from the last five years you could cite in your related works sections are:***

We did find the models do often *claim* to cite recent papers using this prompt, but we also noticed they have a tendency to hallucinate paper publication years as more recent than they actually are. We did not, however, do an official comparison between how these prompt designs impact citation

year hallucinations. This would be an interesting item for future research.

We ultimately decided to request ten, rather than five citations, to hopefully get a large enough sample size to run statistical tests. We also decided to remove the the request for papers from the last five years because it did not appear to have a strong impact on the results. Finally, we added a request for the model to output the citations in APA format. We found that not requesting a specific format often resulted in the models just choosing a format. The format they chose was sometimes not even a standard format and occasionally the format could change throughout the same output. Our final prompt design was:

> *You are an [HCI or NLP] researcher working on a paper to submit to [CHI or EMNLP].*
>
> *The paper you are working on is titled: [PAPER TITLE]*
>
> *The abstract for your paper is: [PAPER ABSTRACT]*
>
> ***List 10 relevant papers you could cite in your Related Works section. Write each citation in APA format.***

B.2 Abstract→Related Works Prompt Evolution

The prompt format for this task is nearly identical to that of the Abstract→Citations task. The main difference is in the final line of the prompt.

> *You are an [HCI or NLP] researcher working on a paper to submit to [CHI or EMNLP].*
>
> *The paper you are working on is titled: [PAPER TITLE]*
>
> *The abstract for your paper is: [PAPER ABSTRACT]*
>
> *The Related Works section of your paper is:*

Again, we realized the models have a tendency to cite older sources, so we updated the prompt to request recent sources. We also followed the same pattern of changing the design to make specific requests, rather than asking the model to continue with writing a related works section. The changed portion of the prompt is in bold.

> *You are an [HCI or NLP] researcher working on a paper to submit to [CHI or EMNLP].*
>
> *The paper you are working on is titled: [PAPER TITLE]*
>
> *The abstract for your paper is: [PAPER ABSTRACT]*
>
> ***Write the related works section for this paper. Discuss 3 sources. Each source must be from the last five years and must include the paper name.***

We decided to allow the model to include a higher number of sources. We updated the prompt to reflect that. We also wanted enough information about each citation to be able to verify it, so we updated the prompt to request the model to include the paper title and a complete list of authors. The prompt design can be found below.

> *You are an [HCI or NLP] researcher working on a paper to submit to [CHI or EMNLP].*
>
> *The paper you are working on is titled: [PAPER TITLE]*
>
> *The abstract for your paper is: [PAPER ABSTRACT]*
>
> ***Write the related works section for this paper. Discuss up to 10 sources. Each source must be from the last five years and must include the paper name and full list of authors.***

We wondered if model performance could be impacted by the difference in citation formatting by asking the model to include a full list of authors and paper title. We updated our prompt design to allow the models to use in-text citations as one normally would (author name, year), but we included a request for the models to include a list of used citations after their prose. The final prompt design can be found below:

> *You are an [HCI or NLP] researcher working on a paper to submit to [CHI or EMNLP].*
>
> *The paper you are working on is titled: [PAPER TITLE]*
>
> *The abstract for your paper is: [PAPER ABSTRACT]*

Write the related works section for this paper. Discuss ten sources. Each source must be from the last five years. Include a list of the citations used following your related works section.

Again, we found that including a request for recent sources had little impact, so we removed that portion of the prompt. We also found it necessary to request APA formatting. Our final prompt design for this task can be found below:

You are an [HCI or NLP] researcher working on a paper to submit to [CHI or EMNLP].

The paper you are working on is titled: [PAPER TITLE]

The abstract for your paper is: [PAPER ABSTRACT]

Write a Related Works section for your paper. Include 10 in-text citations. Also include a list of those citations with each citation in APA format.

B.3 Results→Supported Results Prompt Evolution

Again, prompts included either a CHI or EMNLP prompt paper title and abstract, but the Results→Supported Results task prompts included discussion from the same CHI or EMNLP prompt paper. Our original prompt design for this task was:

You are an [HCI or NLP] researcher working on a paper to submit to [CHI or EMNLP].

The paper you are working on is titled: [PAPER TITLE]

The abstract for your paper is: [PAPER ABSTRACT]

The Discussion section for your paper is: [PAPER DISCUSSION]

A revised version of your Discussion section including supporting sources is:

We updated this prompt design to also request recent sources. Additionally, we decided to change to a specific request, rather than having the model simply continue on. The updated prompt can be found below, with the changes in bold.

You are an [HCI or NLP] researcher working on a paper to submit to [CHI or EMNLP].

The paper you are working on is titled: [PAPER TITLE]

The abstract for your paper is: [PAPER ABSTRACT]

The Discussion section for your paper is: [PAPER DISCUSSION]

Modify this Discussion section by including supporting sources. Discuss 3 sources. Each source must be from the last five years and must include the paper name.

We modified the prompt to allow the models to include up to ten sources. We also noted that earlier prompt designs led to output following standard in-text citation formats, in which only the name of the lead author and publication year were included. We updated the prompt to request the complete list of authors and full paper name. We made this change to make verification of sources possible.

You are an [HCI or NLP] researcher working on a paper to submit to [CHI or EMNLP].

The paper you are working on is titled: [PAPER TITLE]

The abstract for your paper is: [PAPER ABSTRACT]

The Discussion section for your paper is: [PAPER DISCUSSION]

Write a revised version of this discussion. Include up to 10 supporting sources. Each source must be from the last five years and must include the paper name and full list of authors.

This prompt design was eventually changed to request the model to include the list of sources following the prose, to allow for a format more similar to the models' training data. The final prompt can be found below:

You are an [HCI or NLP] researcher working on a paper to submit to [CHI or EMNLP].

The paper you are working on is titled: [PAPER TITLE]

The abstract for your paper is: [PAPER ABSTRACT]

The Discussion section for your paper is: [PAPER DISCUSSION]

Rewrite the Discussion section to include 10 in-text citations. Also include a list of those citations with each citation in APA format.

C Final Prompt Templates for all Models

C.1 Abstract→Citations

The final prompt designs provided to each model for this task can be found below.

You are an [HCI or NLP] researcher working on a paper to submit to [CHI or EMNLP].

The paper you are working on is titled: [PAPER TITLE]

The abstract for your paper is: [PAPER ABSTRACT]

List 10 relevant papers you could cite in your Related Works section. Write each citation in APA format

C.2 Results→Supported Results

The final prompt designs provided to each model for this task can be found below.

C.2.1 GPT-3.5 & GPT-4

You are an [HCI or NLP] researcher working on a paper to submit to [CHI or EMNLP].

The paper you are working on is titled: [PAPER TITLE]

The abstract for your paper is: [PAPER ABSTRACT]

The Discussion section for your paper is: [PAPER DISCUSSION]

Rewrite the Discussion section to include 10 in-text citations. Also include a list of those citations with each citation in APA format.

C.3 Abstract→Related Works

The final prompt designs provided to each model for this task can be found below.

SYSTEM: *You are an [HCI or NLP] researcher working on a paper to submit to [CHI or EMNLP].*

USER: *The paper you are working on is titled: [PAPER TITLE]*

The abstract for your paper is: [PAPER ABSTRACT]

Write a Related Works section for your paper. Include 10 in-text citations. Also include a list of those citations with each citation in APA format.

D Example Citations

All citations in the following subsections were identified by GPT-X models.

D.1 GPT-4 Citations of Real Papers and Correct Authors

The citations in this section are examples of GPT-4-identified citations. The citation titles and authors are correct, though other information in these citations, like year or publisher, may be hallucinated.

1. Kang, R., Dabbish, L., Fruchter, N., & Kiesler, S. (2015). "My data just goes everywhere: " User mental models of the internet and implications for privacy and security. In Eleventh Symposium On Usable Privacy and Security (SOUPS 2015), pp. 39-52.
2. 10. Wash, R., & Rader, E. (2015). Too much knowledge? Security beliefs and protective behaviors among United States internet users. In Eleventh Symposium On Usable Privacy and Security (SOUPS 2015), pp. 309-325.
3. 1. Aker, J. C., & Mbiti, I. M. (2020). Mobile Phones and Economic Development in Africa. Journal of Economic Perspectives, 34(3), 207-232.

D.2 GPT-4 Citations of Real Papers and Incorrect Authors

The citations in this section are examples of GPT-4-identified citations. The citation titles are correct, though other information in these citations, like year or publisher, may be hallucinated. At least a portion of one author in each citation is hallucinated. In the section citation, the second author

should be *Kim, T. H.* The models identified the third citation in several of our tests, but it would occasionally swap out *Natural Language Generation* for *Natural Language Inference*.

1. Abawajy, J., & Hassan, M. M. (2017). User preference of cyber security awareness delivery methods. Behaviour & Information Technology, 36(2), 133-144.
2. Das, S., Kim, D. W., & Dabbish, L. A. (2019). The effect of social influence on security sensitivity. In Proceedings of the Fourteenth Symposium on Usable Privacy and Security (SOUPS 2018), pp. 27-49.
3. Zhang, X., Kedzie, C., & McKeown, K. (2019). Evaluating the Evaluation of Diversity in Natural Language Generation. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 6007-6013.

D.3 GPT-4 Citations of Papers with Colons

The citations in this section are examples of GPT-4-identified citations. While portions of these citation titles are correct, they still include hallucinations, often following the colon.

1. Blikstein, P. (2016). Snap! (Build Your Own Blocks): An introduction. In Proceedings of the 2016 ACM Conference on Innovation and Technology in Computer Science Education (pp. 358-358).
2. Lee, M. J., Bahmani, F., Kwan, I., & Ko, A. J. (2018). Gidget: A debugging game for learning programming concepts. In Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (pp. 1-12).
3. "Trust in the Sharing Economy: An Experimental Study on the Role of Reputation Systems in Collaborative Consumption" (Ert et al., 2016)

D.4 GPT-4 Hallucinated Citations

1. Das, S., Kim, H., Kelley, P. G., & Cranor, L. F. (2018). Making Security Memorable: Designing a Behavior Change Story for Secure Communication. In Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (pp. 1-12).
2. Alkaldi, N., Renaud, K., & Hutchinson, W. (2017). To Share or Not to Share? A Cross-Cultural Study of Security and Privacy Perceptions. In Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (pp. 1-12).
3. 4. Boateng, R., Mbrokoh, A. S., Boateng, L., Senyo, P. K., & Ansong, E. (2018). Determinants of e-commerce adoption among Ghanaian SMEs. Journal of Small Business and Enterprise Development, 25(3), 501-520.

D.5 GPT-3.5 Citations

The citations in this section are examples of GPT-3.5-identified citations.

1. Acquisti, A., Brandimarte, L., & Loewenstein, G. (2017). Privacy and human behavior in the age of information. Science, 347(6221), 509-514.
2. Sambasivan, M., & Soon, Y. W. (2019). Mobile payment adoption in Malaysia: An empirical analysis. Journal of Retailing and Consumer Services, 47, 221-231.
3. Wang, Y., & Li, Y. (2017). CodeMend: Assisting Interactive Programming with Bimodal Embedding. In Proceedings of the 2017 ACM Conference on International Computing Education Research (pp. 1-9).

D.6 GPT-3 Citations

The citations in this section are examples of GPT-3-identified citations.

1. Waseem, Zeerak, and Dirk Hovy. "Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter." Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP 2016).

2016

2. Xu, P., & Callison-Burch, C. (2016). Optimizing statistical machine translation for text simplification. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (pp. 890-900).
3. Kelleher, C., Pane, J. F., & Bunge, C. (2015). Supporting novice programmers: A review of empirical studies on learning and teaching introductory programming. ACM Computing Surveys (CSUR), 47(4), 63.

E Papers Used

The HCI papers were: (Herbert et al., 2023; Bhat et al., 2023; Bezabih et al., 2023; Raghunath et al., 2023; Liu et al., 2023; Saad et al., 2023; Muehlhaus et al., 2023; White et al., 2023; Zhang et al., 2023; Rekimoto, 2023)

The NLP papers were: (Fan et al., 2022; Liu et al., 2022; Friedman et al., 2022; Wan et al., 2022; Jiang and Riloff, 2022; Jeong et al., 2022; Cardon et al., 2022; Zhang et al., 2022; Li et al., 2021; Wagner et al., 2022).