

# Apprentices to Research Assistants: Advancing Research with Large Language Models

MOHAMMAD NAMVARPOUR, Drexel University, USA

AFSANEH RAZI, Drexel University, USA

Large Language Models (LLMs) have emerged as powerful tools in various research domains. This article examines their potential through a literature review and firsthand experimentation. While LLMs offer benefits like cost-effectiveness and efficiency, challenges such as prompt tuning, biases, and subjectivity must be addressed. The study presents insights from experiments utilizing LLMs for qualitative analysis, highlighting successes and limitations. Additionally, it discusses strategies for mitigating challenges, such as prompt optimization techniques and leveraging human expertise. This study aligns with the 'LLMs as Research Tools' workshop's focus on integrating LLMs into HCI data work critically and ethically. By addressing both opportunities and challenges, our work contributes to the ongoing dialogue on their responsible application in research.

CCS Concepts: • **Computing methodologies** → **Machine learning**; • **Human-centered computing** → *Human computer interaction (HCI)*; • **Information systems** → *Information extraction*; • **Social and professional topics** → Codes of ethics; • **Applied computing** → **Document management and text processing**.

Additional Key Words and Phrases: Large Language Models (LLMs), Natural Language Processing (NLP), Research Tools, Artificial Intelligence (AI)

## ACM Reference Format:

Mohammad Namvarpour and Afsaneh Razi. 2024. Apprentices to Research Assistants: Advancing Research with Large Language Models. In *LLMs as Research Tools: Applications and Evaluations in HCI Data Work, a workshop at CHI 2024 (LART@CHI '24)*. ACM, New York, NY, USA, 5 pages.

## 1 OUR NEW LAB MEMBERS?

Although the surprising abilities of Large Language Models (LLMs) only became public knowledge after the introduction of ChatGPT and the ensuing excitement, for those of us who have been following research around LLMs, it was an emerging capability that we could foresee since the introduction of GPT-2 [17]. However, the LLMs were not reliable enough to wear the white jacket and assist in our research.

Ever since ChatGPT was introduced [14], and subsequently tested in every imaginable scenario, from passing SAT exams [19] to boost productivity in white-collar jobs [13], the research community has also become bolder. This might have been due to the push from PhD research assistants, who tried to convince their hesitant advisors that the time had come for them to offload some of the less critical and more mundane aspects of their job (coding, relevancy coding, summarization, etc.) to LLMs. After all, if lawyers [12] and editors of famous magazines [3] can try that, why can't we?

So far, utilization of LLMs in research have been the topic of several studies. In a comprehensive evaluation conducted by Bang et al. [2], ChatGPT was put through a multitask, multilingual, multimodal evaluation. The study found that ChatGPT demonstrated proficiency in several areas including misinformation detection and sentiment analysis. However, the same study also highlighted some limitations of ChatGPT, including struggling with inductive, spatial, mathematical, and non-textual semantic reasoning. Thus, while ChatGPT showed promise in certain areas, it also revealed areas that require further improvement. ChatGPT also proved valuable in annotating implicit hate speech [9]

and achieved acceptable accuracy rates in categorizing statements as true or false for fact-checking [8]. Similarly, stance detection tasks [21] saw ChatGPT producing top-tier results, revolutionizing the field. However, its non-deterministic nature [18] led to inconsistent text annotation and classification outputs, emphasizing the need for validation and caution in unsupervised applications. Concerns were also raised about the potential for data leakage affecting model evaluations [1]. Despite these concerns, Gilardi et al. [6] found that ChatGPT excelled in text-annotation tasks, surpassing crowd-workers in terms of accuracy and cost-effectiveness.

In a study by Ziems et al. [22], the usage of LLMs for textual analysis in various computational social science tasks (e.g., Political Science, Psychology, Sociology) was explored. Results showed LLMs excel in classifying shorter texts with high accuracy and agreement with human annotators, while occasionally outperforming humans in text generation. However, the paper suggested that LLMs should only be used as supplementary annotators and in an augmentative capacity. Researchers were cautioned against solely relying on LLMs for their data analysis.

While the insights from existing studies are valuable in understanding the utility of LLMs in research, offering a firsthand account can provide a more engaging and practical narrative for readers of this article. Therefore, in the following section, we explain our experience with LLMs in a recent project, detailing our observations and approaches.

## 2 OUR EXPERIMENTS WITH LARGE LANGUAGE MODELS

In a recent research project, we decided to try the new Generative AI capabilities as a part of our qualitative study pipeline to help reduce the human labor needed for the project. The corpus on which we wanted to work was comprised of about 35 thousand short spans of text, covering a wide range of topics. We only wanted to study a thin slice of data in which authors mentioned their experience of unwanted online sexual behavior and harassment; a task that was similar to looking for a needle in a haystack.

We initially adopted a basic keyword search for text related to online sexual harassment, refining our keyword list through iterative trials to optimize text retrieval. While this method might be acceptable for searching among more ‘formal’ types of text such as papers and news articles, it certainly was not appropriate for our informal, user-generated content. In fact, our goal in identifying relevant keywords was to build a small corpus of text related to online sexual harassment. We compiled a corpus of 500 text spans. To achieve this, we randomly picked 250 text spans from a subset of our data containing online sexual harassment-related keywords. The remaining 250 spans were selected randomly from the entire dataset to ensure a representative sample. Subsequently, two human annotators labeled each text span as either relevant or non-relevant to our topic of online sexual harassment. They achieved near-perfect agreement with a Cohen’s kappa score of 0.994. We concluded that the task is straightforward and can be automated. The resulting corpus could be used to train a model, thereby automating the process of filtering out relevant text spans.

We initially chose a logistic regression classifier based on BERT [5] to identify relevant text spans. However, due to our small dataset mainly comprising spans with specific keywords related to online sexual harassment, we were concerned the model might overfit and merely recognize spans containing those keywords. Consequently, we adopted an active learning strategy, a machine learning subset where the algorithm queries a human for labeling uncertain data points, aiming to refine its decision-making without large datasets [4]. Despite the engaging learning experience, the approach didn’t significantly improve our model. After processing 35,000 text spans, the model showed high precision but very low recall, accurately identifying few relevant instances and marking most data as irrelevant.

After our first experiment failed, we decided to utilize the new tools in our toolkit, the OpenAI’s GPT models. To ensure the quality of GPT’s annotations, we first wanted to classify our dataset of 500 text spans and compare it with human annotations. To achieve this, we engineered the following prompt for the Language Models:

*Given the text below, determine if it contains a complaint about online sexual harassment. For the purpose of this task, 'Online Sexual Harassment' is defined as any unwanted or unwelcome sexual behavior on any digital platform using digital content (images, videos, posts, messages, pages), which makes a person feel offended, humiliated or intimidated. If the review contains a complaint about such behavior, such as unsolicited flirting or inappropriate texts that are sexual in nature or otherwise harassing, output 1. Otherwise, output 0.: [here goes the review]*

We tested both GPT3.5-turbo [15] and GPT4 [16] through OpenAI's API and measured their performance against human annotations using the Fleiss Kappa Score. GPT3.5-turbo scored 0.799, while GPT4 scored 0.837. Both models showed a strong agreement with human annotations, but GPT4 tended to follow instructions more accurately. For example, while it was mentioned in the prompt that we expected the output to be a single digit, GPT3.5-turbo would sometimes return an explanation, or answers such as 'Yes' and 'No'. GPT4, on the other hand, almost always followed the requested format. However, GPT4 was more expensive to use. Since GPT3.5-turbo's results were still satisfactory, we opted for it due to its cost-efficiency. We input our dataset of 35 thousand text spans into GPT3.5-turbo, which identified 1040 text spans related to online sexual harassment. We then conducted a thematic analysis on this subset of data, performed manually by humans, which is not discussed further in this article.

Upon manual review of the classified text spans, we observed that the language model successfully identified text spans that would pose significant challenges for other machine learning models. Given the sensitivity of our study topic, many instances avoided using exact phrases, opting for substitutions like 'seggs' for 'sex' and 'faq' for 'fuck', making it difficult for simpler models like our initial BERT-based classifier to comprehend. While impressed by the natural language understanding capabilities of LLMs, we found that approximately a quarter of the classifications were false positives. This means that although GPT3.5-turbo identified them as relevant to online sexual harassment, they were not. However, in most cases, these false positives still discussed issues related to online sexual behavior, albeit not as complaints and thus not categorized as 'harassment'.

We applied LLMs in the relevancy coding phase, condensing 35,000 text spans into 1040. Despite around a quarter not directly relevant, this smaller dataset sufficed for our qualitative analysis and theme extraction. Our experiment with LLMs was successful and replicable. Despite some inaccuracies, they didn't hinder our study. In the next section, we'll discuss the opportunities and challenges of using LLMs in research.

### 3 RESEARCH ASSISTANT LLMS: OPPORTUNITIES AND CHALLENGES

Up to this point, we have presented existing research on the utilization of LLMs in the first section, followed by our personal encounters with LLMs in the second section. In this concluding segment of the article, we summarize the opportunities and challenges associated with employing large language models as a tool in research.

Considering how crucial data is in many research endeavors, LLMs have shown great promise in aiding research efforts. They've often performed better than crowdsourced workers [6], who are commonly used for tasks like data analysis. With LLM technology advancing steadily and without major setbacks so far, it seems likely that LLMs will start taking over many tasks, especially simpler ones, from human workers. This presents an opportunity for researchers to conduct larger-scale studies with better quality and lower cost compared to just a few years ago.

However, there are concerns that need to be carefully considered before fully embracing LLMs in research. One of the most challenging aspects of working with LLMs is prompt tuning. Unfortunately, prompts play a crucial role in determining how well the model performs. By utilizing techniques like chain-of-thoughts [20] and retrieval augmented generation (RAG) [11], along with selecting better prompts, the performance of a specific model can be significantly enhanced without altering the model itself. This highlights that while the model may inherently possess the ability

to excel in a particular task, this potential cannot be fully realized without employing specific methods and prompts. Such reliance on external factors introduces a high level of instability and non-reproducibility, which can hinder the integrity of research outcomes.

In addressing this issue, Khattab et al. [10] have introduced the DSPy library, offering researchers a more systematic approach to utilizing LLMs. This library includes an automatic teleprompter feature, which can generate prompts for researchers. While their approach does not negate the significance of prompts and techniques in enhancing LLM performance, it assists researchers, particularly those less experienced in prompt optimization, in obtaining favorable results from their models. Additionally, their work demonstrates that with an effective prompt, even simpler, open-source models can yield satisfactory outcomes.

Another significant challenge lies in addressing the biases inherent in LLMs, which can be particularly problematic when these models are applied to computational social science tasks. A concerning discovery highlighted by Ziems et al. [22] underscores this issue. They conducted experiments using various LLMs, including both open-source models and OpenAI's GPT models. In some instances, they observed unanimous errors, where all models would agree on a false answer. This phenomenon, termed unanimous error, suggests that due to the similarity of datasets used to train large models, they may exhibit similar biases and errors that are not automatically detected, as they are shared among other LLMs as well. Unfortunately, as demonstrated by Glaese et al. [7], even human feedback alone does not entirely eliminate biases either and may sometimes exacerbate them.

Another critical aspect to consider regarding the utilization of LLMs in research is the nature of the research itself. Some research endeavors are inherently more subjective and interpretive. In these cases, the beliefs and lived experiences of researchers significantly influence the outcome of their research and should be acknowledged when publishing results. With the widespread adoption of LLMs, the scientific community risks losing the diverse perspectives of different researchers investigating the same problem. By relying on LLMs as data annotators and analyzers, researchers' perspectives are obscured, and their unique insights may not be fully reflected in their experiments and results. Therefore, the use of LLMs in such fields should be approached with heightened caution.

## 4 CONCLUSION

Given the rising opportunities and challenges brought about by LLMs in research, it's crucial to carefully examine their methodological and ethical implications in the HCI community. We believe that LLMs can significantly improve research efficiency and data analysis. However, we also believe we need to consider their biases and the importance of human oversight carefully. This viewpoint not only supports the objectives of the '*LLMs as Research Tools*' workshop, which aims to ensure methodological validity and address ethical concerns in HCI data work but also contributes to the broader discussion on responsibly integrating LLMs into research practices. By sharing our experiences and insights, we aim to help establish community standards and enhance understanding of LLM applications in HCI research. Our work serves as a case study, illustrating the potential benefits and challenges of integrating LLMs, thus adding to the collective knowledge base on how to navigate LLM use in research settings. Through this contribution, we hope to promote a balanced and critical approach to using LLMs as research tools, furthering the workshop's goal of outlining current approaches and addressing challenges in this rapidly evolving field.

## REFERENCES

- [1] Rachith Aiyappa, Jisun An, Haewoon Kwak, and Yong-Yeol Ahn. 2023. Can We Trust the Evaluation on ChatGPT? arXiv:2303.12767 [cs]

- [2] Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. 2023. A Multitask, Multilingual, Multimodal Evaluation of ChatGPT on Reasoning, Hallucination, and Interactivity. arXiv:2302.04023 [cs]
- [3] David Bauder. 2023. Sports Illustrated Found Publishing AI Generated Stories, Photos and Authors. (Nov. 2023).
- [4] David Cohn, Les Atlas, and Richard Ladner. 1994. Improving Generalization with Active Learning. *Machine Learning* 15, 2 (May 1994), 201–221. <https://doi.org/10.1007/BF00993277>
- [5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, et al. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. <https://doi.org/10.18653/v1/N19-1423>
- [6] Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. ChatGPT Outperforms Crowd-Workers for Text-Annotation Tasks. *Proceedings of the National Academy of Sciences* 120, 30 (July 2023), e2305016120. <https://doi.org/10.1073/pnas.2305016120> arXiv:2303.15056 [cs]
- [7] Amelia Glaese, Nat McAleese, Maja Trębacz, John Aslanides, Vlad Firoiu, Timo Ewalds, Maribeth Rauh, Laura Weidinger, Martin Chadwick, Phoebe Thacker, Lucy Campbell-Gillingham, Jonathan Uesato, Po-Sen Huang, Ramona Comanescu, Fan Yang, Abigail See, Sumanth Dathathri, Rory Greig, Charlie Chen, Doug Fritz, Jaume Sanchez Elias, Richard Green, Soňa Mokrá, Nicholas Fernando, Boxi Wu, Rachel Foley, Susannah Young, Iason Gabriel, William Isaac, John Mellor, Demis Hassabis, Koray Kavukcuoglu, Lisa Anne Hendricks, and Geoffrey Irving. 2022. Improving Alignment of Dialogue Agents via Targeted Human Judgements. <https://doi.org/10.48550/arXiv.2209.14375> arXiv:2209.14375 [cs]
- [8] Emma Hoes, Sacha Altay, and Juan Bermeo. 2023. Leveraging ChatGPT for Efficient Fact-Checking. <https://doi.org/10.31234/osf.io/qnjkf>
- [9] Fan Huang, Haewoon Kwak, and Jisun An. 2023. Is ChatGPT Better than Human Annotators? Potential and Limitations of ChatGPT in Explaining Implicit Hate Speech. In *Companion Proceedings of the ACM Web Conference 2023*. 294–297. <https://doi.org/10.1145/3543873.3587368> arXiv:2302.07736 [cs]
- [10] Omar Khattab, Arnab Singhvi, Paridhi Maheshwari, Zhiyuan Zhang, Keshav Santhanam, Sri Vardhamanan, Saiful Haq, Ashutosh Sharma, Thomas T. Joshi, Hanna Moazam, Heather Miller, Matei Zaharia, and Christopher Potts. 2023. DSPy: Compiling Declarative Language Model Calls into Self-Improving Pipelines. (2023). <https://doi.org/10.48550/ARXIV.2310.03714>
- [11] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems (NIPS'20)*. Curran Associates Inc., Red Hook, NY, USA, Article 793.
- [12] Dan Mangan. 2023. AI: Judge Sanctions Lawyers over ChatGPT Legal Brief. (June 2023).
- [13] Shakked Noy and Whitney Zhang. 2023. Experimental Evidence on the Productivity Effects of Generative Artificial Intelligence. *Science* 381, 6654 (July 2023), 187–192. <https://doi.org/10.1126/science.adh2586>
- [14] OpenAI. 2022. Introducing ChatGPT.
- [15] OpenAI. 2023. GPT-3.5-Turbo.
- [16] OpenAI. 2023. GPT-4.
- [17] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language Models Are Unsupervised Multitask Learners.
- [18] Michael V. Reiss. 2023. Testing the Reliability of ChatGPT for Text Annotation and Classification: A Cautionary Remark. arXiv:2304.11085 [cs]
- [19] Taylor Webb, Keith J. Holyoak, and Hongjing Lu. 2023. Emergent Analogical Reasoning in Large Language Models. *Nature Human Behaviour* 7, 9 (July 2023), 1526–1541. <https://doi.org/10.1038/s41562-023-01659-w>
- [20] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. (2022). <https://doi.org/10.48550/ARXIV.2201.11903>
- [21] Bowen Zhang, Daijun Ding, and Liwen Jing. 2023. How Would Stance Detection Techniques Evolve after the Launch of ChatGPT? arXiv:2212.14548 [cs]
- [22] Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. 2023. Can Large Language Models Transform Computational Social Science? (2023). <https://doi.org/10.48550/ARXIV.2305.03514>

Received 22 February 2024; accepted 22 March 2024