AI and Ethics
https://doi.org/10.1007/s43681-023-00294-5

OPINION PAPER

Guidance for researchers and peer-reviewers on the ethical use of Large Language Models (LLMs) in scientific research workflows

Ryan Watkins[1]

Received: 18 April 2023 / Accepted: 2 May 2023

Abstract
For researchers interested in exploring the exciting applications of Large Language Models (LLMs) in their scientific investigations, there is currently limited guidance and few norms for them to consult. Similarly, those providing peer-reviews on research articles where LLMs were used are without conventions or standards to apply or guidelines to follow. This situation is understandable given the rapid and recent development of LLMs that are capable of valuable contributions to research workflows (such as OpenAI's ChatGPT). Nevertheless, now is the time to begin the development of norms, conventions, and standards that can be applied by researchers and peer-reviewers. By applying the principles of Artificial Intelligence (AI) ethics, we can better ensure that the use of LLMs in scientific research aligns with ethical principles and best practices. This editorial hopes to inspire further dialogue and research in this crucial area of scientific investigation.

Keywords Large Language Model · LLM · Research · Science · Norms · Conventions · Standards

1 Introduction

Recent advancements in Large Language Models (LLMs), such as OpenAI's ChatGPT 4 [1] and Google's LaMDA [2], have inspired developers and researchers alike to find new applications and uses for these groundbreaking tools. [3] From applications that summarize one, or one thousand, research papers, to those that let users "chat" with a research publication, many innovative techniques and creative products have been developed in the past few months. Most recently, the first wave of research articles that use LLMs in their scientific research workflows have started to show up – primarily as preprints at this stage (for instance, [4–7]). As with many new research methods, statistical techniques, or technologies, the use of new tools "in the wild" routinely precedes agreement on the norms, conventions, and standards that guide their application. LLMs are no exception, with many researchers exploring their possible applications at numerous phases of scientific research workflows. Therefore, now is the time to start establishing norms, conventions, and standards [8, 9] for the use of LLMs in scientific research, both as guidance for researchers and peer-reviewers, and as a starting place to guide future research into establishing these as foundations for applying the principles of Artificial Intelligence (AI) ethics in research practice.

The ethical use of LLMs in scientific research requires the development of norms, conventions, and standards. Just as researchers apply norms, conventions, and/or standards to hypothesis testing, regression, or CRISPR applications, researchers can benefit from guidance on how to both use, and report on their use, of LLMs in their research.[1] Similarly, for those providing peer-reviews of scientific research papers that use LLMs in their methods, guidance on current conventions and standards will be valuable. The implementation of norms, conventions, and standards plays a critical role in ensuring the ethical use of artificial intelligence (AI) in scientific research, bridging the gap between theoretical frameworks and their practical application. This is particularly relevant in research involving Large Language Models (LLMs)..

✉ Ryan Watkins
rwatkins@gwu.edu

[1] Educational Technology Leadership, George Washington University, Washington, DC, USA

[1] For example, a *norm* in international economics research is comparability (i.e.,the desire to compare statistics across countries) [10], where as a long-standing *convention* in the social sciences is to use a value of $\alpha = 0.05$ to define a statistically significant finding [11]. While IEEE's P11073-10426 is a *standard* that defines a communication framework for interoperability with personal respiratory equipment [12].

Published online: 16 May 2023

The creation and study of LLMs is a rapidly advancing field [3]. With the growing use of LLMs it is expected that the norms, conventions, and standards will evolve as new tools and techniques are introduced. Nevertheless, it is important to begin the foundation building process so that initial guidance can be systematically improved over time. In this editorial I propose an initial set of considerations that can (i) be applied by researchers to guide their use of LLMs in their workflows, and (ii) be utilized by peer-reviewers to assess the quality and ethical implications of LLMs use in the articles they review. These initial norms, conventions, and standards for what should be considered during the research process, and included in reports or articles on research that used LLMs, are a starting place with the goal of providing an ethical foundation for future dialogue on this topic.[2] The proposed foundation should ideally identify key research questions that will be explored in the coming months, such as determining the appropriate conventions for setting LLM temperature parameters and assessing potential disciplinary and field-specific variations in these conventions.

2 Framework

The following is an initial framework of proposed norms that researchers and peer-reviewers should consider when using LLMs in scientific research. While this framework is not intended to be comprehensive, it provides a foundation on which researchers can build and develop conventions and standards.

The proposed framework (which includes, context, embeddings, fine tuning, agents, ethics) was derived from the key considerations of researchers using LLMs. These considerations range from determining if LLMs are going to used in combination with other research tools and deciding when to customize LLMs with embedding models, to fine tuning the performance of LLMs and ensuring that research retains ethical rigor. As such, the proposed framework captures many unique considerations to using LLMs in the workflows of scientific research. Described first are the up-front considerations for researchers who plan to use LLMs in their workflows, followed by a checklist of questions (within the same framework) peer-reviewers should consider when reviewing articles or reports that apply LLMs in their methods.

2.1 Context

The context in which LLMs are used in research workflows is important to their appropriate and ethical application. Initial considerations of researchers should include:

- Are LLMs appropriate for the research questions and data?
- Will LLMs be used along with other methods or tools?
- Will the study be pre registered?

LLMs are not, of course, appropriate for all research questions or data types. Researchers should begin with their research question(s) and then determine if/how LLMs might be applied. LLMs may, for instance, be an appropriate component of data collection (e.g., writing interview questions), data preparation (e.g., fuzzy joining of data sets), and/or data analysis (e.g., sentiment analysis, optimizing code). For example, in analyzing qualitative data a researcher may choose to use traditional qualitative data analysis software and techniques (such as, coding or word counts with Nvivo or Atlas TI) along with a LLM for comparing semantics across samples. Within this context, the use of the LLM complements other analysis techniques, allowing the researcher to explore more diverse questions of interest. Whereas in other contexts all of the research questions may be best explored with just LLMs or another traditional method. In their reporting, researchers should describe and justify the complete methods applied in their research and the full list of LLM tools selected since each may be specialized for a different task. Likewise, if the research study was pre registered, any subsequent articles or reports should include both the pre registration URL and discussion of any changes made from the original pre registered research plan—especially when those changes are based on the testing and fine tuning of LLMs.

2.2 Embedding Models

Adding a custom embedding model(s) to complement the base LLM (such as OpenAI's ChatGPT) can enhance the value of LLMs for specific research task(s). Initial considerations of researchers should include:

- Will a custom embedding model(s) help meet the goals of the research?
- What tool(s) will be used to create the embedding model(s)?
- Will multiple embedding models created and tested (i.e., chained)?
- What size of chunks will be used in preparing the data for the embedding(s)?

[2] Research and updated guidance for using LLMs in scientific research workflows are available on the clearinghouse website: https://LLMinScience.com.

- Will overlap across chunks be permitted?
- What tool will be used for similarity matching (i.e., vector database)?
- Will the code for creating embedding model(s) be made publicly available?

While the web interface for some LLMs (such as ChatGPT) can be valuable for some research questions, many times supplemental content (in addition to a base LLM, such as GPT-3.5 or GPT-4) is important to the research. Custom embedding models allow researchers to extent the base LLM with content of their choosing. Technically, "Embeddings are vectors or arrays of numbers that represent the meaning and the context of the tokens that the model processes and generates. Embeddings are derived from the parameters or the weights of the model, and are used to encode and decode the input and output texts. Embeddings can help the model to understand the semantic and syntactic relationships between the tokens, and to generate more relevant and coherent texts" [13]. While LLMs use embeddings to create their base models (such as, GPT-4), researchers can also create embeddings with specialized content (such as a corpus of research articles on a topic, a drive of interview transcripts, or a database of automobile descriptors) to expand the inputs used by the LLM. Researchers can also chain together multiple embedding models in improve LLM performance [14].

There are numerous embedding models [algorithms] that can be used by researchers to create an embeddings file for use in their research [15]. Embedding models use a variety of algorithms to create the custom embeddings file, and therefore it is important for researchers to be transparent about their procedures in selecting and creating embeddings for use in their workflow. The preparation of data for creating the embedding model(s) can also influence the resulting embeddings and thereby the outputs of the LLMs when used in the workflow. For example, text has be divided into chunks in preparation for creating the embeddings and the size of chunks used will define the cut-off points for creating vectors. Researchers can, for instance, divide the text data into chunks of 1000 tokens, or 500 tokens. Depending on the context of the research, one dividing point for chunking may be more valuable than another. Chunking can also be done using sentence splitting in order to keep sentences together, or not. Likewise, researchers can allow for some overlap between chunks in order to maintain semantic context [16]. Each of these decisions can influence the output of the LLM when using additional embeddings, and thus should be considered in the research procedures and included in subsequent reporting.

After a embeddings are created for the additional content to be used in conjunction with the base LLM, the embeddings have to be stored in a database so that the data can be managed and searched. Vector databases (or vectorestores) are used, and there are many options researchers can choose amongst [17]. Vector databases use different heuristics and algorithms to index and search vectors, and can perform differently. Vector databases may use different neural search frameworks, such as FAISS, Jina.AI, or Haystack, and custom algorithms [18]. While the selection of a vector database mostly influences performance (i.e., speed, more than LLM outputs) it is useful for researchers to be transparent on their selection. In the future, differences in neural search frameworks, algorithms, and vector database technologies may lead to substantive differences in LLM outputs as well.

2.3 Fine Tuning

There are many Large Language Models (LLMs) available to researchers [19] and the selection of which LLM to use in a specific research workflow requires several decisions, including:

- Which language model will used (e.g., OpenAI's GPT-3.5, GPT-4, open source alternative)?
- Will multiple language models be tested for performance in the research task(s)?
- Will completion parameters be applied (e.g., temperature, presence penalty, frequency penalty, max tokens, logit bias, stops)?
- Will multiple combinations of completion parameters be tested before or during the research?
- Will systematic "prompt engineering" be done as part of the research?
- What quality review and validation checks will be performed on LLM-generated results?
- Will the LLM's performance be compared with benchmarks or standards for the field or discipline?
- Will the code for fine tuning the LLM be made publicly available?

Beyond the standard user interface and default settings offered by many LLMs (such as the ChatGPT website), by using an Application Programming Interface (API) researchers can fine tune LLMs for their research. Fine tuning can be done with or without using a embedding model(s), and is currently done primarily through setting the completion parameters (e.g., temperature) and by conducting "prompt engineering" (i.e., systematically improving LLM prompts to provide outputs with desired characteristics). Additional fine tuning options should however be expected as LLMs evolve and more competing LLMs become available to researchers.

Currently there are no conventions or standards for setting completion parameters when using LLMs in scientific research. For instance, two common parameters used to influence the outputs of LLMs are tokens and temperature.

2.3.1 Tokens

Tokens are unit of analysis of LLMs, and they are roughly equivalent to about a word, but not always. Researchers can select the number of tokens to be returned to complete a request, and the LLM will complete the request within that constraint [20]. Depending on the size of the LLM there may be limits on the total number of tokens that can be requested. There are no conventions or standards at this time for the ideal maximum number of tokens a researcher should request in order to get results, and this will routinely be dependent on the research context in which they are using the LLM. In general however, LLMs have been observed to ramble on at time (i.e., filling the maximum number of tokens) and to provide less accurate outputs toward the end when the maximum token parameter is set too high.

2.3.2 Temperature

Temperature [20] is used to provide the LLM with additional flexibility in how it completes a request. At the lowest temperature setting (e.g., 0) then the LLM is limited to selecting the next word/token that has the highest probability in the model (also see, "top p" parameter [20]). As the researcher increases the temperature (≤ 2 with OpenAI's LLMs), the LLM may select from an increasing range of probabilities for the next word/token. Setting an appropriate temperature for the unique research context is therefore important, and in the future we will hopefully have conventions (by field and/or disciplines) on appropriate temperature parameters for research.

Other completion parameters can also influence the outputs of LLMs (e.g., "presence penalty", "frequency penalty", "logit bias") and we should expect that new LLMs will expand the range of completion parameters that researchers can apply. It should be the norm, therefore, for researchers to clearly state the applied completion parameters used in their research, and describe any testing of different parameter settings done in evaluating and selecting the final parameter settings.

Prompts are the inputs provided by researchers to request a LLM response. Prompts are converted to tokens and used to inform predictions about what the following words/tokens should be in the output. Behind the curtain, LLMs are using probabilities for the various permutations and combinations of tokens/words that could follow. Changing the prompt, for instancing changing the wording of the prompt or including more prior prompts from the history of a conversation, can substantially influence the LLM's outputs [21, 22]. Prompt engineering is the systematic manipulation of prompts in order to improve outputs, and researchers should be transparent about both their prompt engineering procedures and the final prompts used to in the research.

At this time, however, "There are no reliable techniques for steering the behavior of LLMs" [3]. While transparency of research "prompt engineering" practices is essential, when using LLMs in research transparency may not lead to reproducability—and therefore limit generalizability.

2.4 Agents

The automation of LLM tasks can be important in some research contexts. If using automated LLM tools (i.e., agents) researcher considerations should include:

- Will LLM agent(s) used in the research?
- How many and in what sequence will LLM agent(s) used?
- Will the code for creating the agents be made publicly available?

Many research workflows can utilize a predetermined sequence of prompts or chains of LLMs. Other workflows, however, can't rely on predetermined sequences and/or decisions to achieve their goals. In these later cases, LLM agents can be used to make decisions about which LLMs and tools (including, for instance, internet searches [23]) to use in achieving a goal [24]. A LLM agent utilizes prompts, or LLM responses, as inputs to their (the agent's) reasoning and decisions about which LLMs or tools to utilize next. Further, LLM agents can learn from their past performance (i.e., successes or failures) leading to improved performance [25, 26]. If researchers apply LLM agents in their workflow, details on the agents and tools used in the research should be described. Any intermediate steps, and the sequence of those steps, should also be described since these are essential to how the final outputs of the LLM were achieved.

2.5 Ethics

The use of LLMs in scientific research workflows is a new area of AI ethics that requires emerging considerations for researchers, including:

- Is the organization (e.g., company, open source community) that created the LLM transparent about the choices they made in its development and fine tuning?
- How will training data for additional embedding model(s) be acquired in a transparent and ethical manner?
- What steps for data privacy and protections will be taken?
- What will be done to identify and mitigate potential biases in LLM-generated results?
- Are there any potential conflicts of interest related to the use of LLMs?

Table 1 Peer-reviewer's checklist

Context
☐ Was the study pre-registered?
☐ Were LLMs used to complement other research methods, or as the sole method?
☐ Were the research questions and data appropriate for LLM methods?
Embedding Models
☐ Were embedding(s) used in the research?
☐ Is the tool used to create the embedding model(s) provided and described?
☐ Were multiple embedding models created, tested, or used (i.e., chained)?
☐ Is the size of chunks used in preparing the data for embedding provided?
☐ Were different sizes of chunks tested for influence on the LLMs performance?
☐ Is the size of overlap permitted when creating chunks provided?
☐ Is the tool used for similarity matching (i.e., vector database) provided and described (e.g., FAISS)?
☐ Is the code for creating embedding(s) available?
Fine Tuning
☐ Which language model was used (e.g., OpenAI's GPT−3.5 model)?
☐ Were multiple language models tested for performance?
☐ Are the completion parameters applied (e.g., temperature, presence penalty, frequency penalty, max tokens, logit bias, stops) provided?
☐ Were multiple combinations of completion parameters tested?
☐ Is any "prompt engineering" described in detail?
☐ Did the researcher(s) include the final prompts used?
☐ Were quality review checks performed on LLM-generated results?
☐ Did the researcher(s) validate the LLM-generated results through experimentation or simulation?
☐ Did the researcher(s) evaluate the LLM's performance against other benchmarks or standards?
☐ Is the code for fine tuning available?
Agents
☐ Were LLM agent(s) used in the research?
☐ Were the intermediate steps of the LLM agent(s) described?
☐ Is the code for creating the agents available?
Ethics
☐ Does the researcher(s) describe ethical considerations applied when selecting an appropriate base LLM for the research?
☐ Were training data for additional embedding model(s) acquired in a transparent and ethical manner?
☐ Were proper steps for data privacy and protections taken?
☐ Did the research methods address potential biases in LLM-generated results?
☐ Did the researcher(s) disclose any conflicts of interest related to the use of LLMs?
☐ Did the researcher(s) comply with applicable institutional and/or regulatory guidelines?
☐ Were proper citations and credit given?
☐ To the extent possible are the LLM methods done in a manner that is reproducible and transparent?
☐ Were LLM outputs described in a non-anthropomorphic manner?

- Are there any applicable institutional and/or regulatory guidelines that will be followed?
- What steps will be taken for the research to be reproducible and transparent?
- Will LLM outputs be described in a non-anthropomorphic manner?

The ethical use of LLMs in research workflows is a crucial consideration that cuts across multiple disciplines. From sociology and psychology to engineering management and business, LLMs have diverse applications in research, and this necessitates attention to a range of issues. These issues include technical concerns such as data privacy and bias, as well as philosophical considerations such as anthropomorphism and the epistemological challenges posed by machine-generated knowledge. Therefore, it is essential to address ethical considerations when using LLMs in research workflows to ensure that the research remains unbiased, transparent, and scientifically rigorous. While researchers may have little control, for example, over the ethical collection of data for the initial training of an LLM (such as OpenAI's GPT-3.5), they do have choices in which LLMs to utilize in

their research and the ethical collection of data used in creating any custom embedding models used in their workflows. Likewise, while there are currently limited institutional and/or regulatory policies guiding the use of LLMs in scientific research, researchers will be responsible for adhering to those AI policies (such as the EU AI Act [27]) when they are established. In the interim, researchers must be detailed and transparent about their practices, provide proper citations and credit, and disclose any conflicts of interest.

3 Conclusions

As LLMs continue to advance, their potential uses, benefits, and limitations in scientific research workflows are emerging. This presents an opportune moment to establish norms, conventions, and standards for their application in research and reporting their use in scientific publications. In this editorial, I have proposed an initial framework and set of norms for researchers to consider, including a peer-reviewer checklist (see Table 1) for assessing research reports and articles that employ LLMs in their methods. These proposals are not meant to be definitive, as we are still in the early stages of learning about the potential uses and limitations of LLMs. Rather, it is hoped that this foundation will stimulate research questions and inform future decisions about the norms, conventions, and standards that should be applied when using LLMs in scientific research workflows.

References

1. OpenAI: GPT-4 Technical Report. https://cdn.openai.com/papers/gpt-4.pdf (2023)
2. Romal Thoppilan, E.A.: LaMDA: language models for dialog applications. arXiv preprint arXiv:2201.08239 (2022)
3. Bowman, S.R.: Eight things to know about large language models. arXiv preprint arXiv:2304.00612 (2023)
4. Crokidakis, N., de Menezes, M.A., Cajueiro, D.O.: Questions of science: chatting with ChatGPT about complex systems arXic preprint arXiv:2303.16870 (2023)
5. Wang, Z., Xie, Q., Ding, Z., Feng, Y., Xia, R.: Is ChatGPT a good sentiment analyzer? A preliminary study. arXiv preprint arXiv:2304.04339 (2023)
6. Qi, Y., Zhao, X., Huang, X.: safety analysis in the era of large language models: a case study of STPA using ChatGPT. arXiv preprint arXiv:2304.04339 (2023)
7. Khademi, A.: Can ChatGPT and bard generate aligned assessment items? A reliability analysis against human performance. arXiv preprint arXiv:2304.05372 (2023)
8. Southwood, N., Eriksson, L.: Norms and conventions. Philos. Explor. **14**(2), 195–217 (2011). https://doi.org/10.1080/13869795.2011.569748
9. Bowdery, G.J.: Conventions and norms. Philos. Sci. **8**(4), 493–505 (1941). https://doi.org/10.1086/286731
10. Mügge, D., Linsi, L.: The national accounting paradox: how statistical norms corrode international economic data. Eur. J. Int. Relat. **27**(2), 403–427 (2021). https://doi.org/10.1177/13540661209363339. (**PMID: 34040493**)
11. Johnson, V.E.: Revised standards for statistical evidence. Proc. Natl. Acad. Sci. U.S.A. **110**(48), 19313–19317 (2013)
12. Chang, M.: IEEE standards used in your everyday life-IEEE SA—standards.ieee.org. https://standards.ieee.org/beyond-standards/ieee-standards-used-in-your-everyday-life. Accessed 16 Apr 2023
13. Maeda, J.: LLM Ai Embeddings. https://learn.microsoft.com/en-us/semantic-kernel/concepts-ai/embeddings
14. Wu, T., Terry, M., Cai, C.J.: AI chains: transparent and controllable human-AI interaction by chaining large language model prompts. arXiv preprint arXiv:2110.01691 (2022)
15. Chase, H.: Text embedding models (2023). https://python.langchain.com/en/latest/modules/models/text_embedding.html?highlight=embedding
16. Chunking Strategies for LLM Applications. https://www.pinecone.io/learn/chunking-strategies/
17. Chase, H.: Vectorstores (2023). https://python.langchain.com/en/latest/modules/indexes/vectorstores.html
18. Kan, D.: Not all vector databases are made equal. Towards Data Science (2022). https://towardsdatascience.com/milvus-pinecone-vespa-weaviate-vald-gsi-what-unites-these-buzz-words-and-what-makes-each-9c65a3bd0696
19. Hannibal046: Hannibal046/Awesome-LLM: Awesome-LLM: a curated list of large language model. https://github.com/Hannibal046/Awesome-LLM
20. OpenAI: OpenAI API—platform.openai.com. https://platform.openai.com/docs/api-reference/completions/create. Accessed 16 Apr 2023
21. Si, C.: Prompting gpt-3 to be reliable. In: ICLR 2023 Proceedings. https://openreview.net/pdf?id=98p5x51L5af (2023)
22. White, J., Fu, Q., Hays, S., Sandborn, M., Olea, C., Gilbert, H., Elnashar, A., Spencer-Smith, J., Schmidt, D.C.: A prompt pattern catalog to enhance prompt engineering with ChatGPT. arXiv preprint arXiv:2302.11382 (2023)
23. Significant-Gravitas: GitHub-Significant-Gravitas/Auto-GPT: An experimental open-source attempt to make GPT-4 fully autonomous.—github.com. https://github.com/Significant-Gravitas/Auto-GPT. Accessed 16 Apr 2023 (2023)
24. Chase, H.: Agents. https://python.langchain.com/en/latest/modules/agents.html (2023)
25. Shinn, N., Labash, B., Gopinath, A.: Reflexion: an autonomous agent with dynamic memory and self-reflection. arXiv preprint arXiv:2303.11366 (2023)
26. Schick, T., Dwivedi-Yu, J., Dessì, R., Raileanu, R., Lomeli, M., Zettlemoyer, L., Cancedda, N., Scialom, T.: Toolformer: language models can teach themselves to use tools. arXiv preprint arXiv:2302.04761 (2023)
27. Union, E.: Artificial Intelligence Act (2023). https://artificialintelligenceact.eu/