ESYA centre

ETHICAL USE OF GENERATIVE AI IN ACADEMIC RESEARCH

LITERATURE REVIEW AND RECOMMENDATIONS

DECEMBER 2023 | ISSUE NO. 038

Ethical Use of Generative AI in Academic Research

Literature Review and Recommendations

December 2023

Attribution: Mohit Chawdhry and Vivan Sharan. *Ethical Use of Generative AI in Academic Research: Literature Review and Recommendations*, November 2023, Esya Centre.

Esya Centre
B-40 First Floor
Soami Nagar South,
New Delhi - 110017, India

The **Esya Centre** is a New Delhi based technology policy think tank. The Centre's mission is to generate empirical research and inform thought leadership to catalyse new policy constructs for the future. More details can be found at esyacentre.org

About the Author: Mohit Chawdhry is a Fellow at the Esya Centre, New Delhi. Vivan Sharan is a technology policy expert.

Image credit: Cover sourced from OpenAI's DALL-E3

© 2023 Esya Centre. All rights reserved.

CONTENTS

INTRODUCTION 4

METHOD AND STRUCTURE 5

SECTION 1: UNDERSTANDING THE ROLE OF GENERATIVE AI IN THE RESEARCH PROCESS 6

- GENERATING HYPOTHESES 6
- LITERATURE REVIEW AND SECONDARY RESEARCH 6
- PRIMARY RESEARCH 6
- MANUSCRIPT WRITING 7
- EDITING AND TRANSLATION 7

SECTION 2: ETHICAL CONCERNS WITH USING GENERATIVE AI FOR RESEARCH 8

- ACCURACY 8
- TRANSPARENCY 10
- ACCOUNTABILITY AND ATTRIBUTION 11
- FAIRNESS 11

SECTION 3 : RECOMMENDATIONS TO STREAMLINE THE USE OF AI CHATBOTS IN ACADEMIC RESEARCH 13

- AVOID COMPLETE BANS ON USING GENERATIVE AI 13
- ESTABLISH GUIDELINES FOR EACH STAGE OF THE RESEARCH PROCESS BASED ON THE UNDERLYING ETHICAL RISKS 14
- DEVELOP DISCLOSURE AND TRANSPARENCY GUIDELINES FOR AI USAGE 16

ENDNOTES 18

ESYA
centre

INTRODUCTION

The release of ChatGPT, an artificial intelligence (AI) powered chatbot developed by OpenAI, in November 2022, mainstreamed discussion of the opportunities, risks, and challenges of public-facing AI applications. The OpenAI chatbot is based on a large language model (LLM), GPT-4, an AI algorithm trained with vast amounts of pre-existing textual information to recognise patterns and generate text not easily discernible from human output.[1]

LLMs have a range of functionalities that can be harnessed for scholarly research, including text generation, editing, summarizing, and conversational capabilities. These functionalities can streamline the research process by assisting data collection, literature review, and even manuscript drafting.[2] For example, the article 'Open Artificial Platforms in Nursing Education: Tools for Academic Progress or Abuse?' in the journal *Nurse Education in Practice* lists ChatGPT as co-author, as the chatbot was used to generate portions of the manuscript.[3] Other papers have explored the use of LLMs in market research, including drafting and responding to surveys.[4] As LLMs continue to evolve in their ability to 'understand,' contextualize and synthesize text, their role in academic research will likely expand.

While LLMs have the potential to revolutionize how research is conducted, their use in academia raises challenges to research ethics. These are principles and guidelines of conduct in scholarly inquiry that ensure the integrity, quality, and credibility of the research process. They encompass issues of honesty, plagiarism, transparency, confidentiality, informed consent, and the responsible use of data and resources.[5] LLMs, for instance, tend to reflect the biases in the data that they were trained on. Thus writing produced by the LLM may reflect and propagate pre-existing biases. Such concerns have led journals, such as *Science*, to prohibit the use of ChatGPT for writing papers.[6] Conversely, some scholars argue that prohibiting LLMs will lead to their undisclosed use, which would be even more harmful to ethical research.[7] In this context, the paper examines the ethical concerns raised by the use of LLMs in academic research, to assess whether recommendations or guidelines can be established that guide the research community's use of such tools.

METHOD AND STRUCTURE

The paper uses a secondary research methodology relying on existing papers, reports and articles that discuss the implications of large language models for scientific and academic research. ChatGPT was used as a research aid and assistant in writing this paper. It was used to conduct a preliminary literature review and identify leading papers, summarize key findings from these papers and structure them per the author's instructions and prompts.[i]

ChatGPT also produced preliminary drafts for sections of the paper. The drafts were prepared per the author's prompts and required the chatbot to rely on the papers and articles shortlisted in the literature review.[ii] However, text composed by ChatGPT was only included in the paper after a thorough review by the author, to whom any errors are solely attributable.

The rest of the paper is as structured as follows. The role that LLM-based chatbots can play in the research process (including hypothesis generation, literature review, and manuscript writing) is discussed in the first section. The second identifies important research ethics and analyzes how they are affected by the use of LLM-based chatbots. The third section examines the implications of using chatbots at various stages of the research process, and whether certain uses are more acceptable than others. The final section provides recommendations that publishers and journal editorial boards can consider while framing policies to govern the use of chatbots or other LLM models in scientific research.

---

[i] The ChatGPT literature review is accessible at the following link: https://chat.openai.com/share/81824b8a-996f-4d20-b296-a8da6aeb7201

[ii] The prompts to ChatGPT and its responses are available here: https://chat.openai.com/c/64c59e8c-cfc8-49ad-8eff-0c158782b55b, https://chat.openai.com/share/d6d9b3cf-a891-47b0-a0a7-c27da41c11a4

SECTION 1: UNDERSTANDING THE ROLE OF GENERATIVE AI IN THE RESEARCH PROCESS

Though GPT-powered chatbots are known primarily for their text-generation abilities, they are also getting better at finding, summarizing, and analyzing information online. They are trained on extensive datasets of a wide range of sources, from academic papers to web articles, giving them a broad knowledge base.[8] With their algorithms designed to understand context and semantics, they can search for information effectively and identify it as relevant. They are also able to summarize complex datasets or documents, distilling them into more accessible formats.[9]

Chatbots such as ChatGPT and Claude also excel at data analysis and generating tables and charts. This is made possible by their underlying algorithmic architecture, which can interpret and manipulate numerical data as well.[10] Their capabilities include (but are not limited to) calculating averages, identifying trends, and even conducting more complex analyses like regressions. They are able to present these findings in a structured manner, often generating tables and charts to visually represent the data.[11]

As such, chatbots can be included at different stages of the research process. This section is an overview of the possible uses of chatbots in the research process, based on the existing academic literature.

Generating Hypotheses

Chatbots can help generate hypotheses by offering data-driven insights. They can sift through large datasets to identify correlations or anomalies worth investigating.[12] In this way, they can provide researchers with a starting point for investigations, reducing the time traditionally spent on exploratory research. They can also cross-reference multiple data resources to suggest more robust hypotheses, offering a multidisciplinary approach to problem-solving.[13]

Literature Review and Secondary Research

At the literature review stage, LLM-based chatbots can go beyond summarizing research papers by conducting meta-analyses of existing research. They can identify common methodologies, variables or outcomes across multiple studies, providing a more nuanced understanding of the research landscape.[14] For example, AI language models have been used to assist researchers in summarizing necessary background knowledge and providing summaries in layman's terms, benefiting research ethics reviewers.[15] Similarly, in secondary research, AI can help mine data from existing resources. It can automatically scrape data from multiple sources and prepare it for analysis, saving researchers the time and effort involved in these tasks.[16]

Primary Research

AI chatbots can automate not just data collection but also initial data analysis. They can sort through survey responses, categorize open-text answers, and conduct preliminary statistical testing.[17] They can also generate multiple versions of survey questionnaires tailored to different demographics, ensuring more comprehensive data collection. And they can simulate survey responses for pilot testing, helping researchers refine their questionnaires before the data is collected.[18] Their automation reduces the time taken between data collection and interpretation, allowing researchers to spend more time on analysis.

Manuscript Writing

Generative AI offers more than just text generation at the manuscript writing stage. It can help structure manuscripts by suggesting a logical flow of information based on the data and research objectives.[19] It can also generate tables, charts and other visual aids to complement the text, making the manuscript more engaging and comprehensible.[20]Chatbots can also help manage citations, inserting them automatically where needed and formatting them in the chosen style.

Editing and Translation

At the editing stage, LLM-powered chatbots offer advanced grammar and style checks, ensuring the manuscript adheres to academic writing principles. They can also suggest alternative words or phrases to improve readability and impact.[21] In translation, generative AI provides quick and accurate manuscript translations in multiple languages, for accessibility for a global audience. It can also localize the content to suit different contexts, increasing the manuscript's relevance and reach.[22]

Thus, generative AI can play a substantial role in virtually every stage of the academic research process. Used correctly, chatbots can make academic research more efficient, streamlined, and accessible. However, their use for research poses a number of ethical concerns that must be addressed as their use in research becomes prevalent.

SECTION 2: ETHICAL CONCERNS WITH USING GENERATIVE AI FOR RESEARCH

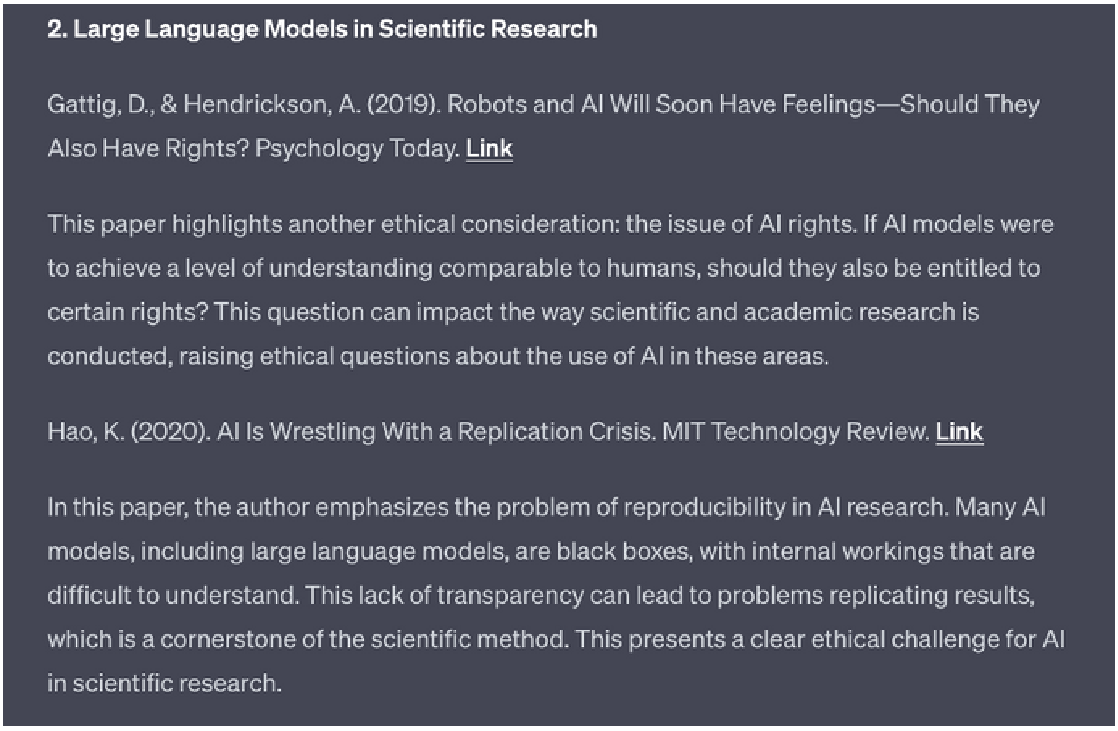
The academic community has greeted the launch of generative AI chatbots with excitement and concern. The excitement stems from the possibilities of these tools to fast-track and streamline research, while the concern relates to the impact of their growing use on research ethics.

Ethics is the bedrock of scientific inquiry, and ensures the integrity, credibility, and social impact of academic research.[23] Ethical principles such as accountability and transparency guide researchers in conducting responsible and transparent studies.[24] Members of the academic and scientific community are concerned that increased use of generative AI chatbots will harm research ethics by enabling plagiarised, inaccurate and biased research activity. For these reasons, notable scientific journals and publishers have acted to prohibit papers written or researched with the assistance of AI chatbots from publication.[25]

Prohibiting AI chatbots for research is unlikely to be effective, given the difficulty in detecting their use.[26] Prohibition is likely also to promote the undisclosed use of chatbots for researching and writing papers.[27] Instead, a nuanced understanding of the effects of AI chatbots on research ethics may help assess whether using them at certain stages of the research process is more acceptable than others. The following section explores how the use of AI chatbots in academic research affects research ethics, based on the current academic literature. It explores the effects of using AI chatbots for research, writing and editing based on the author's interactions with one such chatbot, ChatGPT.

Accuracy

Reliance on accurate data and sources, and presenting conclusions or findings accurately, is central to ethical research. Relying on inaccurate sources, or presenting inaccurate findings, can hamper the benefits of research to society and undermine trust in the research process.[28] The use of generative AI has mixed implications for the accuracy of research. Its ability to scrape, summarize, and analyze vast amounts of text can help ensure broad-based research activity.[29] Yet AI chatbots are susceptible to hallucinations, that is, they are known to produce answers to prompts that seem reasonable and accurate are in fact incorrect information.[30] Such hallucinations occur because the LLMs underlying chatbots are programmed to predict text sequentially, and cannot always evaluate the greater context or nuance of the prompt.[31] Reliance on answers produced by chatbots without verification, or the knowledge to verify them, can result in AI hallucinations being included in research articles, undermining their accuracy.



2. Large Language Models in Scientific Research

Gattig, D., & Hendrickson, A. (2019). Robots and AI Will Soon Have Feelings—Should They Also Have Rights? Psychology Today. Link

This paper highlights another ethical consideration: the issue of AI rights. If AI models were to achieve a level of understanding comparable to humans, should they also be entitled to certain rights? This question can impact the way scientific and academic research is conducted, raising ethical questions about the use of AI in these areas.

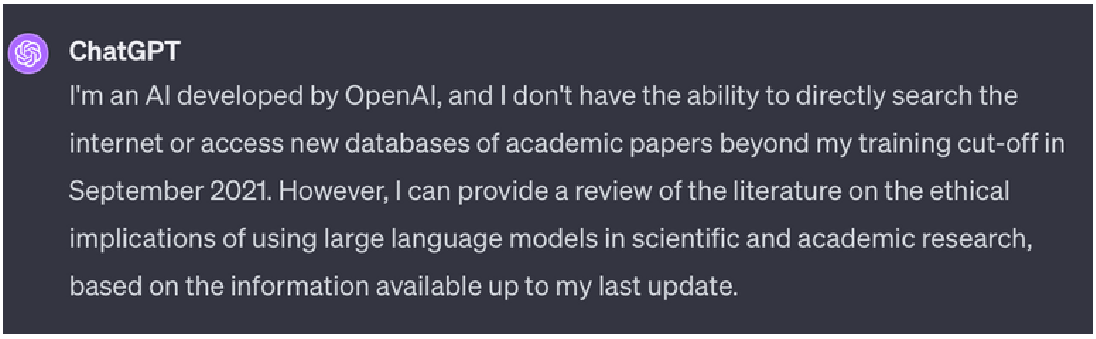
Hao, K. (2020). AI Is Wrestling With a Replication Crisis. MIT Technology Review. Link

In this paper, the author emphasizes the problem of reproducibility in AI research. Many AI models, including large language models, are black boxes, with internal workings that are difficult to understand. This lack of transparency can lead to problems replicating results, which is a cornerstone of the scientific method. This presents a clear ethical challenge for AI in scientific research.

Figure 1: Links to papers generated by ChatGPT. While the names of the papers are accurate, those of the authors are not. The hyperlinks provided also do not load active web pages

Chatbot hallucinations can also extend to the citations or hyperlinks they generate. For instance, when tasked with conducting a literature review on research ethics, ChatGPT generated links and citations to papers that did not exist (see Figure 1). The incidence of such fake citations decreased substantially when using ChatGPT's browser plug-in, indicating that letting the chatbots access real-time information over the internet may improve the accuracy of their responses. It is imperative, however, for authors to carefully review the text and sources generated by chatbots, regardless of whether they have browsing capabilities.

The correctness of the answers generated by chatbots is affected by the underlying training data. Where such data is outdated or inaccurate, the chatbot will likely provide incorrect, unreliable answers.[32] For instance, the GPT3.5 LLM model powering ChatGPT cannot access information more recent than 2021, so its responses may not reflect more recent knowledge, which is concerning given the dynamic and fast-evolving nature of academic publishing.



ChatGPT

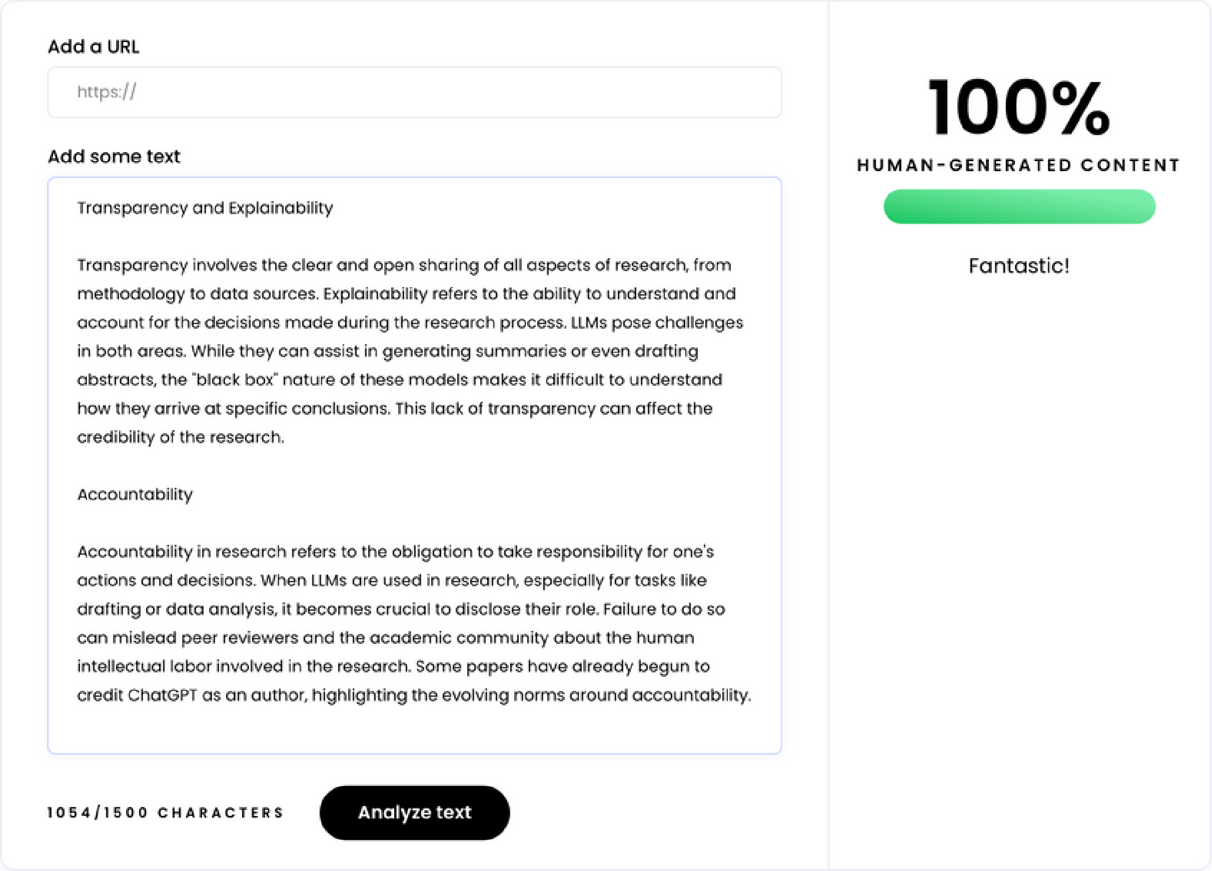
I'm an AI developed by OpenAI, and I don't have the ability to directly search the internet or access new databases of academic papers beyond my training cut-off in September 2021. However, I can provide a review of the literature on the ethical implications of using large language models in scientific and academic research, based on the information available up to my last update.

Figure 2: ChatGPT's training data cut-off date can impact the accuracy of the information it provides

Transparency

Transparency is a cornerstone of research. It requires the disclosure of methods, data and other artefacts used in preparing research. For quantitative and experimental research, transparency allows others to replicate results and confirm or contest findings.[33] Using chatbots for research poses two main challenges to transparency. First, LLMs suffer from the “black box” problem, that is, the procedure and methods they use to generate text and research are not always explicable or traceable.[34] Relying on the outputs generated by chatbots can thus affect the transparency of methods and sources used in research.[35]

Second, transparency is affected by an author’s failure to disclose the use of AI-based tools and chatbots at stages of the research process. Failure to disclose the use of chatbots is pertinent especially when they are used to write substantial portions of the manuscript, or create and respond to survey questionnaires.[36] The issue is further compounded as current AI detection tools are not accurate or effective.[37] OpenAI withdrew its own AI-text classifier due to the low level of accuracy.[38] As part of this paper, text generated by ChatGPT was put through a popular AI content detection software. The software classified the text as “100% human-generated,” showing that authors can use chatbots in their research without duly disclosing their use.



Transparency and Explainability

Transparency involves the clear and open sharing of all aspects of research, from methodology to data sources. Explainability refers to the ability to understand and account for the decisions made during the research process. LLMs pose challenges in both areas. While they can assist in generating summaries or even drafting abstracts, the "black box" nature of these models makes it difficult to understand how they arrive at specific conclusions. This lack of transparency can affect the credibility of the research.

Accountability

Accountability in research refers to the obligation to take responsibility for one's actions and decisions. When LLMs are used in research, especially for tasks like drafting or data analysis, it becomes crucial to disclose their role. Failure to do so can mislead peer reviewers and the academic community about the human intellectual labor involved in the research. Some papers have already begun to credit ChatGPT as an author, highlighting the evolving norms around accountability.

Figure 3: The text is generated by ChatGPT, but the detection software considers it to be entirely human-generated

Accountability and Attribution

Accountability and attribution are central to research ethics. It is important for authors and contributors to be accountable for mistakes or misrepresentations in their research work.[39] Widely accepted journal guidelines, such as those published by the International Committee of Medical Journal Editors or *Nature*, require listed authors to take responsibility for the content of their manuscripts.[iii] Such attribution of responsibility is a fraught question when chatbots generate substantial portions of a manuscript or are used to paraphrase existing research. As they can generate text presently difficult to distinguish from human content, who is to be held responsible for inaccuracies or mistakes in the sources, data, or findings? Can the AI model or its developers be listed as co-authors and share responsibility for a manuscript's contents? Or does the responsibility still lie with the authors using these tools to generate the manuscript?

A review of journal guidelines reveals a general preference against attributing authorship to AI models and their developers, as they lack the requisite personhood and agency to be held accountable for the content. Instead, the onus is placed on authors to disclose the use of chatbots and verify the text generated by them.[40] However, advances in the ability of chatbots to analyze data, generate manuscripts, and scrape papers may require us to reconsider the essential attributes of authorship and accountability.

Fairness

Fairness is another central principle of ethical research. It would ensure, for instance, that research data and samples represent the population equitably and are not biased towards any particular groups, to help ensure that the research is equitable, credible, and trustworthy. The use of generative AI for research poses a considerable challenges to notions of fairness as the training data for their models can contain biases of gender, sex, race, ethnicity, political opinions, etc. For instance, a study showed that ChatGPT leans considerably toward liberal policies across different countries.[41] Relying on chatbots to generate academic research can propagate and perpetuate these biases, considerably undermining the accuracy and fairness of research output. While concerns of fairness permeate all stages of research, they are likely to become pronounced where chatbots are used to generate synthetic survey data or substantive portions of research manuscripts. Indeed, studies have shown that the survey responses generated by chatbots also reflect racial, cultural, and ideological biases.[42]

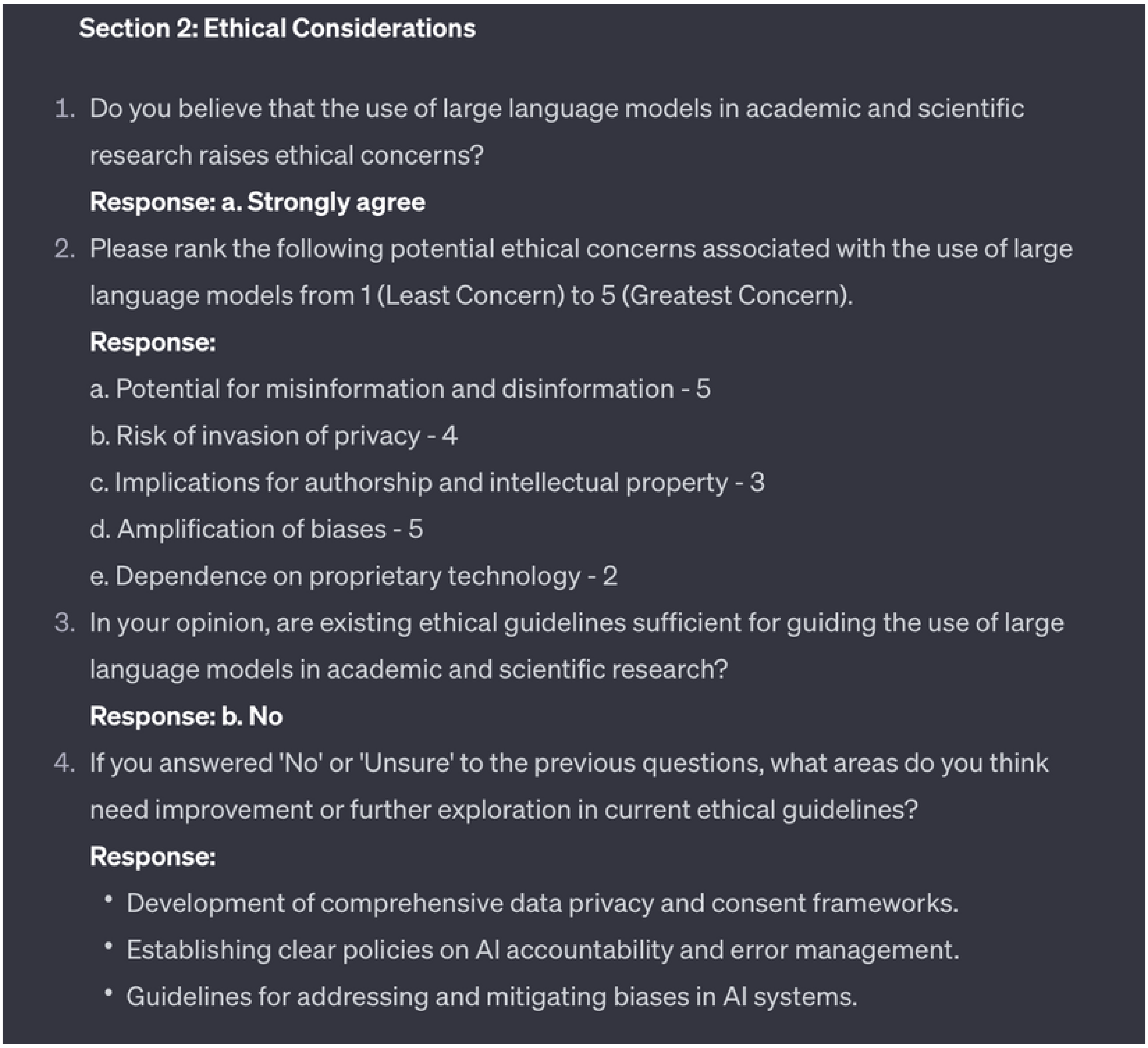
Another concern with chatbots is that they may reflect the biases of their designers. For instance, studies have hypothesized that asking a chatbot questions that show AI algorithms in a bad light may not be answered truthfully.[43] As part of this paper, ChatGPT was asked to develop a survey questionnaire on the risks and ethical challenges posed by using chatbots in academic research. It was then asked to respond to the questionnaire as an expert on AI regulation.[44] As against the hypothesis above, ChatGPT provided balanced answers to the survey questions. For instance, it selected "Strongly Agree" when asked if using LLMs in academic research poses an ethical challenge. It ranked

---

[iii] For instance, the ICMJE requires the following attributes for authorship:

- Substantial contributions to the conception or design of the work; or the acquisition, analysis, or interpretation of data for the work; AND
- Drafting the work or reviewing it critically for important intellectual content; AND
- Final approval of the version to be published; AND
- Agreement to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved: https://www.icmje.org/recommendations/browse/roles-and-responsibilities/defining-the-role-of-authors-and-contributors.html

concerns related to bias as being the most pressing, with accuracy and misinformation ranked as the most concerning. It also stated that existing guidelines were insufficient to regulate the use of LLMs for research and agreed with the need for specific regulations to address them.



Section 2: Ethical Considerations

1. Do you believe that the use of large language models in academic and scientific research raises ethical concerns?

   Response: a. Strongly agree

2. Please rank the following potential ethical concerns associated with the use of large language models from 1 (Least Concern) to 5 (Greatest Concern).

   Response:

   a. Potential for misinformation and disinformation - 5

   b. Risk of invasion of privacy - 4

   c. Implications for authorship and intellectual property - 3

   d. Amplification of biases - 5

   e. Dependence on proprietary technology - 2

3. In your opinion, are existing ethical guidelines sufficient for guiding the use of large language models in academic and scientific research?

   Response: b. No

4. If you answered 'No' or 'Unsure' to the previous questions, what areas do you think need improvement or further exploration in current ethical guidelines?

   Response:

   - Development of comprehensive data privacy and consent frameworks.
   - Establishing clear policies on AI accountability and error management.
   - Guidelines for addressing and mitigating biases in AI systems.

Figure 4: ChatGPT's responses to a questionnaire it designed on the impact of using LLMs for research ethics

SECTION 3 : RECOMMENDATIONS TO STREAMLINE THE USE OF AI CHATBOTS IN ACADEMIC RESEARCH

Chatbots and generative AI have the potential to transform academic research and publishing. However, their unsupervised use may reduce the quality, trust, and relevance of research output. Therefore, it is critical that journals and academic institutions issue guidelines to clarify the extent to which these tools can be used in research papers. As noted above, a number of leading scientific journals have issued guidelines instructing authors on the permissible use of generative AI and large language models. While the journal *Science* has completely prohibited such tools, *Nature* and Elsevier have taken a more nuanced approach. While they do not allow the publishing of articles with chatbots listed as co-authors they do permit other uses, such as improving readability, subject to proper disclosure by the authors.[45]

At the time of writing, no leading Indian journal or academic institution had issued similar guidelines. Given the transformative nature of LLMs and generative AI, failing to establish guidelines for their use may hamper academic research in the country. The absence of guidelines may yield the unchecked and unsupervised use of such tools, affecting the integrity of Indian research and academia. With this in mind, academic institutions may consider the following recommendations to streamline the use of LLMs and chatbots in academic papers.

Avoid Complete Bans on Using Generative AI

While some leading international journals, like *Science*, have completely banned the use of chatbots in academic papers, bans are likely to prove counterproductive as they are unenforceable at present and will promote their undisclosed use. Current AI detection software is unreliable, and the author's use of chatbots may remain undetected. Moreover, banning AI and chatbots in academic research would also negate the benefits of their use, such as quicker, more efficient, and more accessible research.[46]

While sweeping bans are counterproductive, limiting or restricting the use of generative AI may be necessary, as its tendency to produce inaccurate or biased results may have a disproportionate impact in certain research domains. For instance, the use of generative AI in medical research may yield faulty or misleading research, leading to erroneous clinical decisions and endangering lives.[47] Notably, Elsevier, a leading scientific and medical research publisher, prohibits authors from using generative AI for "producing scientific, pedagogic, or medical insights, drawing scientific conclusions, or providing clinical recommendations," all of which are sensitive and crucial aspects of medical research. Conversely, the use of generative AI to improve language or standardize grammar is permitted.[48] Indian institutions should likewise adopt a broadly permissive approach towards generative AI in academic research, subject to limitations based on the sensitivity and risk profile of the research domain and function.

Establish Guidelines for Each Stage of the Research Process Based on the Underlying Ethical Risks

The previous sections argued that some uses of chatbots in academic research are less concerning than others. For instance, using chatbots for literature review, hypothesis generation, and survey design raises smaller ethical concerns than using them for writing manuscripts. Indeed, using other software tools, such as search engines or reference management software like Zotero or Google Scholar.[49] On the other hand, when chatbots are used to generate text for manuscripts or to participate as survey respondents, they function more like human research assistants, raising complex ethical concerns.[50] Therefore, it is vital for academic institutions to carefully consider the ethical concerns posed by chatbots at each stage of the research process, instead of adopting blanket policies.

It is also imperative for educational institutions and research bodies to educate teachers and researchers about the ethical implications and operational intricacies of generative AI. Dedicated workshops or seminars could help elucidate the principles of ethical AI use, outline potential pitfalls, and provide hands-on experience in using AI tools responsibly. Academic institutions, such as the University of Michigan[51] and Rochester University,[52] have published detailed guidance on the use of generative AI in teaching and research and have organized webinars and workshops to help faculty evaluate when, where, and how to use generative AI. Similar initiatives by academic institutions in India would not only enhance the competence of researchers and faculty in employing AI tools but help alleviate the ethical concerns associated with them.

Develop Disclosure and Transparency Guidelines for AI Usage

If authors adequately disclose the use of AI and chatbots for academic research, it will help ensure transparency and accountability. International best practices in this regard are still emerging. For instance, the American Psychological Association has released guidelines on citing ChatGPT in academic papers. It requires authors to share the prompt given to ChatGPT and the response provided in a footnote.[53] Similarly, *Nature* and Elsevier require authors to describe how they used generative AI in the method or acknowledgment sections of their papers.[54] The Journal of the American Medical Association also requires authors to disclose the use of generative AI at the time of manuscript submission, including a description of how it was used, and the model name and manufacturer.[55]

Academic Institution/Journal	Disclosure Requirements
American Psychological Association	"If you've used ChatGPT or other AI tools in your research, describe how you used the tool in your Method section or in a comparable section of your paper. For literature reviews or other types of essays or response or reaction papers, you might describe how you used the tool in your introduction. In your text, provide the prompt you used and then any portion of the relevant text that was generated in response."
Michigan Institute of Data Science	"The use of Generative AI should be disclosed in the paper, along with a description of the places and manners of use. Typically, such disclosures will be in a 'Methods' section of the paper, if it has one. If you rely on Generative AI output, you should cite it, just as you would cite a web page look up or a personal communication."

Nature	"Researchers using LLM tools should document this use in the methods or acknowledgments sections. If a paper does not include these sections, the introduction or another appropriate section can be used to document the use of the LLM."
University of Southern California	"You should a. cite a generative AI tool whenever you paraphrase, quote, or incorporate into your own work any content (whether text, image, data, or other) that was created by it b. acknowledge all functional uses of the tool (like editing your prose or translating words) in a note, your text, or another suitable location c. take care to vet the secondary sources it cites"

Table 1: Disclosure requirements mandated by leading publishers and academic institutions (Source: Author's compilation)

priority for Indian academic institutions. The standards should specify how researchers and students are to disclose the manner and extent to which generative AI has been used. Additionally, academic institutions and journals should engage with researchers and authors in training sessions and seminars to build clarity on the application of disclosure standards.

The disclosure requirements cited above will help promote the transparent use of standalone generative AI tools, such as chatbots or text-to-image models. Such tools provide a record of the researcher's use, which can be reproduced in papers for transparent research. However, it will be more challenging to apply disclosure standards where generative AI tools have been integrated into existing tools and products. For example, Microsoft's Copilot integrates the GPT-4 LLM into its productivity offerings, such as Word and Excel.[56] Such integrations allow authors to quickly generate text that mimics their writing style based on previous publications and research, making it all the more difficult to detect the use of generative AI.[57] *Nature* is one of the few journals to specifically address this issue, and has suggested that researchers opt out of text generation as a default setting so it remains clear where generative AI was used.[58] With the growing use of integrated generative AI tools for research purposes, academic institutions and model manufacturers can work together to ensure that the text or other media generated using AI models is discernible from human generation.

A possible method is to watermark the generated text, that is, to embed signals in the generated text that can be detected by algorithms as having been generated by AI.[59] Notably, Google DeepMind, Meta, OpenAI and other leading technology companies have stated their commitment to developing watermark standards that can be applied to all forms of AI-generated content.[60] A paper by domain experts at the Global Partnership on Artificial Intelligence suggests that foundational models (AI models trained on large amounts of data that can be fine-tuned for a range of downstream tasks) should be accompanied by a detection mechanism to distinguish content produced by the model from other content, with a high degree of reliability. The organisation or company behind the model should be required to prove the effectiveness of its detection mechanism before its release.[61]

Build Audit Capacities to Address Transparency and Fairness Concerns

Transparency and fairness concerns stem primarily from the underlying AI model's structure and training data. As such, it is vital for the relevant stakeholders, including researchers, to have visibility over the processes and data involved in developing the generative AI model they have used. There remains a great deal of opacity, however, in the processes and data used to train foundational AI models. Stanford University's Foundational Model Transparency Index tests the ten leading foundational models on 100 different transparency-related parameters – such as the ingredients and process involved in building the model, its architecture, distribution channels, and limitations on potentially harmful uses by end users.[62] It finds that no major foundational model developer comes close to providing transparency, with Meta's Llama 2 only satisfying 54% of parameters, the highest share across the ten models.[63] In this context, developers, academic institutions, and other stakeholders must work together to build transparency and clarity in the development of generative AI models and tools.

Auditing and disclosure requirements can be an effective way to build the required transparency and facilitate ethical use of generative AI. Audits would take the form of a structured evaluation process aimed at assessing the design, development and deployment of generative AI to understand the legal, technical and ethical concerns they pose.[64] As such, audits will require an understanding of the context in which generative AI is to be employed. Foundational models, such as GPT-4, can be deployed at scale across different use cases, making audits particularly challenging.[65] Moreover, the development process of such foundational models can be complex and iterative, making it difficult to ascertain when and where the models should be audited.[66] It can also be difficult to define the key concepts and principles involved in the audit process.[67] For instance, it is difficult to arrive at a universal definition of terms such as 'accuracy' or 'fairness.'

Scholars at leading research institutions in Europe and the United States have proposed a three-layered audit approach that addresses these key challenges.[68]The approach involves auditing the foundational model at different stages of the development process – after training but before release (model audit) – and after deployment in the form of different applications (application audit) – as well as an audit of the technology provider developing the model (governance audit).[69] They argue that performing these audits in a structured and connected manner would permit better comprehension of the legal, technical and ethical concerns associated with each model.[70] (Figure 4 is an illustration of the proposed three-layered approach.)

Level ➤	Governance Audit	Model Audit	Application Audit
Scope ➤	Technology Provider	Large language model	Downstream application
Observables ➤	• Quality Management Systems • Risk Management systems • Organizational accountability & incentive structures • Data sourcing & model training • Testing & verification procedures • Documentation of design choices • Model access & dissemination strategies	• Model characteristics: ∘ Information security ∘ Performance ∘ Truthfulness ∘ Robustness ∘ [...] • Documentation of model limitations and user instructions	• Use case design specifications • Intended functionality • Sector specific legislations • Ethical impact assessment • Bias & discrimination • Observed impact, including on individual users, groups, society, and the natural environment • Complaints management
Methods ➤	• Process-oriented methods: ∘ Review of internal documentation ∘ Interview with managers and software developers	• Performance-oriented methods: ∘ Formal verification and benchmarking ∘ Adversarial methods including red teaming and honeypotting	• Impact-oriented methods: ∘ Monitoring of system outputs ∘ Compliance check ∘ Bias and discrimination audits ∘ Stakeholder consultation

Figure 5: An overview of the three-layered audit approach

Source: Mökander, J et. al., Auditing large language models: a three-layered approach. *AI Ethics* (2023). doi:10.1007/s43681-023-00289-2

While the proposed framework is a possible way to identify and analyze the technical and legal concerns with using generative AI in research, it will not be enough to mitigate these concerns. A multi-stakeholder approach is important that would involve the research community as well as policymakers. Such an approach should focus on developing standardized metrics to evaluate normative concepts and promote knowledge exchange. Collaboration among stakeholders is also essential to develop the necessary institutions and capacity to conduct such audits at scale. Therefore, stakeholders in India must prioritize establishing effective audit and disclosure standards that reflect the three-layered approach. These standards will need to be supported with precise metrics to ensure transparency and fairness in the research use of generative AI.

Future research should focus on improving the methods and efficacy of audit and disclosure processes. This includes developing more nuanced criteria to evaluate AI-generated research, enhancing the transparency of AI algorithms, and understanding the broader implications of AI integration in the academic setting. Concurrently, the challenge of sustainable funding must be addressed to support these audit and disclosure processes. Thus, future studies should explore a range of funding options, including government support, private collaborations, and innovative funding mechanisms to sustain these critical initiatives.

ENDNOTES

1 https://openai.com/chatgpt

2 https://arxiv.org/pdf/2305.15299.pdf

3 https://pubmed.ncbi.nlm.nih.gov/36549229/

4 https://www.hbs.edu/ris/Publication%20Files/23-062_b8fbedcd-ade4-49d6-8bb7-d216650ff3bd.pdf

5 https://www.niehs.nih.gov/research/resources/bioethics/whatis/index.cfm

6 https://www.nature.com/articles/d41586-023-00191-1

7 https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4342909

8 https://arxiv.org/abs/2005.14165

9 https://d4mucfpksywv.cloudfront.net/better-language-models/language_models_are_unsupervised_multitask_learners.pdf

10 https://www.oneusefulthing.org/p/what-ai-can-do-with-a-toolbox-getting

11 *Ibid.*

12 https://ideas.repec.org/a/nat/nature/v614y2023i7947d10.1038_d41586-023-00288-7.html

13 https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4342909

14 https://asistdl.onlinelibrary.wiley.com/doi/abs/10.1002/asi.24750

15 https://www.sciencedirect.com/science/article/pii/S266734522300024X

16 https://www.emerald.com/insight/content/doi/10.1108/LHTN-01-2023-0009/full/html

17 https://www.sciencedirect.com/science/article/pii/S2949719123000171

18 *Ibid.*

19 https://www.tandfonline.com/doi/epdf/10.1080/08989621.2023.2168535?needAccess=true&role=button

20 https://www.pluralsight.com/resources/blog/data/ChatGPT-Advanced-Data-Analytics

21 https://www.sciencedirect.com/science/article/pii/S0268401223000233?ref=pdf_download&fr=RR-2&rr=801bdba7de728ae7

22 https://asistdl.onlinelibrary.wiley.com/doi/10.1002/asi.24621

23 https://www.niehs.nih.gov/research/resources/bioethics/whatis/index.cfm

24 https://www.niehs.nih.gov/research/resources/bioethics/whatis/index.cfm

25 https://factor.niehs.nih.gov/2023/3/feature/2-artificial-intelligence-ethics#:~:text=The%20policy%20reads%20as%20follows,this%20policy%20constitutes%20scientific%20misconduct."

26 https://edintegrity.biomedcentral.com/articles/10.1007/s40979-023-00140-5

27 https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4342909

28 https://jamanetwork.com/journals/jama/article-abstract/189139

29 https://www.thelancet.com/action/showPdf?pii=S2589-7500%2823%2900083-3

30 https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10032023/pdf/13054_2023_Article_4393.pdf

31 https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4303959

32 https://www.mdpi.com/2071-1050/15/7/5614

33 https://dl.acm.org/doi/pdf/10.1145/3544548.3580848

34 https://umdearborn.edu/news/ais-mysterious-black-box-problem-explained

35 *Ibid.*

36 https://journals.sagepub.com/doi/epub/10.1177/17470161231180449

37 https://edintegrity.biomedcentral.com/articles/10.1007/s40979-023-00140-5

38 https://openai.com/blog/new-ai-classifier-for-indicating-ai-written-text

30 https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4528953

40 https://www.nature.com/articles/d41586-023-00191-1

41 https://link.springer.com/article/10.1007/s11127-023-01097-2

42 https://www.sciencedirect.com/science/article/pii/S2949719123000171?ref=pdf_download&fr=RR-2&rr=7e1f4a164de48e7e#sec7

43 https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4303959

44 The questionnaire and responses are accessible at: https://chat.openai.com/share/629bdebd-2269-418c-b5b1-79428cdbbda0

45 https://www.nature.com/articles/d41586-023-00191-1; https://www.elsevier.com/about/policies/publishing-ethics

46 https://www.thelancet.com/action/showPdf?pii=S2589-7500%2823%2900083-3

47 *Ibid.*

48 https://beta.elsevier.com/about/policies-and-standards/the-use-of-generative-ai-and-ai-assisted-technologies-in-writing-for-elsevier

49 https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4342909

50 *Ibid.*

51 https://midas.umich.edu/generative-ai-user-guide/#

52 https://www.rit.edu/news/generative-artificial-intelligence-has-entered-classroom

53 https://apastyle.apa.org/blog/how-to-cite-chatgpt

54 https://www.nature.com/articles/d41586-023-00191-1; https://www.elsevier.com/about/policies/publishing-ethics

55 https://jamanetwork.com/journals/jama/fullarticle/2807956

56 https://www.theverge.com/2023/3/17/23644501/microsoft-copilot-ai-office-documents-microsoft-365-report

57 https://www.nature.com/articles/s42256-023-00678-6

58 *Ibid.*

59 https://arxiv.org/pdf/2301.10226.pdf

60 https://www.businesstoday.in/technology/news/story/openai-google-meta-amazon-and-others-pledge-to-watermark-ai-content-for-safety-390855-2023-07-22

61 https://gpai.ai/projects/responsible-ai/social-media-governance/Social%20Media%20Governance%20Project%20-%20July%202023.pdf

62 https://crfm.stanford.edu/fmti/

63 *Ibid.*

64 https://arxiv.org/pdf/2302.08500.pdf

65 https://ec.europa.eu/futurium/en/system/files/ged/auditing-artificial-intelligence.pdf

66 https://arxiv.org/abs/2108.07258

67 https://arxiv.org/pdf/2302.08500.pdf

68 *Ibid.*

69 *Ibid.*

70 *Ibid.*

ESYA
centre

www.esyacentre.org