ACE INSTITUTE for ADVANCING COMPUTING EDUCATION

GRAISER Guidelines for the Responsible Use of AI in STEM Education Research

Research Report · May 2026

Guidelines for the Responsible Use of AI in STEM Education Research

Protecting Our Participant Communities and Research Integrity

PRINCIPAL INVESTIGATOR
Dr. Monica McGill

CO-PRINCIPAL INVESTIGATOR
Dr. Julie M. Smith

FUNDING
NSF Award No. 2519885

Acknowledgements & Contributors

Acknowledgement

This material is based upon work supported by the U.S. National Science Foundation Mid-scale Research Infrastructure Program under Award No. 2519885. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

Our team also thanks the American Institutes for Research® (AIR®) for donating office space for holding the workshop at their Chicago, IL, offices.

AI Use Disclosure

The guidelines and this report were jointly developed by workshop participants, reviewers, and the project team. The project team only used AI to find relevant background literature, which was human-reviewed before inclusion or exclusion.

Suggested Citation

Smith, J.M. and McGill, M.M. 2026. *Guidelines for the Responsible Use of AI in STEM Education Research: Protecting Our Participant Communities and Research Integrity*. Institute for Advancing Computing Education. DOI:

Acknowledgements & Contributors

STEERING COMMITTEE

Dr. Justin Reich
Associate Professor of Comparative Media Studies, MIT

Dr. Megan Stubbs-Richardson
Associate Research Professor at the Social Science Research Center, Mississippi State University

Dr. Aman Yadav
Lappan-Phillips Professor of Computing Education, Michigan State University

WORKSHOP PARTICIPANTS

Dr. Elizabeth M. Adams
Chief Engagement Officer, Minnesota Responsible AI Institute

Gadhaun Aslam
PhD Candidate & Graduate Assistant, University of Florida

Dr. Stephanie Baker
Senior Evaluation Researcher, Texas Advanced Computing Center at The University of Texas at Austin

Kristan Beck
Director of Computer Science, Chicago Public Schools

Dr. Oladayo Bello
Assistant Professor, New Mexico State University

Dr. Lisa Bosman
Associate Professor, Purdue University

Dr. Tugba Boz
Assistant Professor of Instructional Design and Technology, University of North Dakota

Tara Chattoraj
Researcher, American Institutes for Research

Ali Crawford
Senior Research Analyst, Center for Security and Emerging Technology

Dr. Adrienne Decker
Associate Professor, University at Buffalo

William “Ike” Eisenhauer
MEng, Assistant Professor (Adjunct), Mechanical and Materials Engineering, Portland State University

Rachel Garrett, PhD.
Managing Researcher, American Institutes for Research

Dr. Sharin Jacob
Researcher in Computational Thinking & Learning Pathways, Digital Promise

Sofía De Jesús
Computer Science Strategy Specialist, The University of Texas at Austin

Grace Li
PhD Student in Computer Science, University of Chicago

Dr. Szymon Machajewsk,
Associate Director, Academic Technology and Learning Innovation, University of Illinois Chicago

Watsachol Narongsaksakul
Graduate Research Assistant, College of Education, Illinois State University

Diana Dajung Oh
PhD, Researcher, American Institutes for Research

Dr. Nirav S. Patel
Asst. Professor, University of Colorado, Boulder

Rebecca J. Peterson
PhD, Associate Director, Stark Science Learning Center, University of Utah

Jeremy Riel
Assistant Professor, University of Illinois Chicago

Acknowledgements & Contributors

WORKSHOP PARTICIPANTS

Yunus Doğan Telliel
Assistant Professor of Anthropology and Director of Great Problems Seminars, Worcester Polytechnic Institute

Prof. Krista K. Thomason
Associate Professor and Chair of Philosophy, Swarthmore College

Eben B. Witherspoon
PhD; Senior Researcher; American Institutes for Research

Dr. Michole Washington
NASA SCoPE Postdoctoral Research Scholar, School of Earth and Space Exploration at Arizona State University

Lois Wong
AI Librarian, University of Chicago

REVIEWERS

Dr. Kyriaki Chatzikyriakidou
Faculty, American Farm School - Perrotis College

Dr. Maria Ebling
Assistant Professor, United States Military Academy

Maia Elkana
Evaluation Director - Institute for School Partnership, Washington University in St. Louis

Rudy Escobar
STEM and Computer Science Coordinator, Stanislaus County Office of Education/ CSforCA Co-Chair

Dr. Mic Fenech
Assistant Professor, Gardner-Webb University

Dr. Krishna Chaitanya Rao Kathala
Program Director, The Ohio State University

Dr. Hyejin Park
Assistant Professor, Drake University

Dr. Katerina Sergi
Assistant Research Professor, Social Science Research Center, Mississippi State University

Dr. Micah Swartz
Assistant Professor, Troy University

Dr. Monica Torres
Associate Teaching Professor, Rutgers University

Dr. Jamie Huber Ward
Associate Director of Higher Education, NCWIT

Rebecca Zarch
SageFox Consulting Group, Executive Director

Table of Contents

Executive Summary

The recent advent of generative AI tools has impacted many fields, including STEM education research.

These tools can be applied to virtually every aspect of the research process, from writing research questions to finding relevant literature to analyzing data. They hold promise for efficiency and scaling, but they also present some significant threats to the practice of research itself. For example, many of these tools lack the transparency that is a cornerstone of the research process. As a result, there is a pressing need for guidelines that promote the responsible use of AI in STEM education research to ensure that research teams can harness benefits without exacerbating risks. To respond to this need, the project team gathered experts in STEM education research and in AI to articulate guidelines; several rounds of expert review served to refine the guidelines.

These guidelines suggest concerns to be aware of and steps to take as STEM education researchers consider using AI in their research workflows in order to ensure that their AI use comports with best practices for conducting responsible research.

Thorough human critique of AI input and output is the most important guideline (see the graphic on the next page). This AI-in-the-loop (née human-in-the-loop) approach ensures that education researchers' expertise addresses the fact that AI is an assistive tool for humans. Then, four other guidelines are prioritized: (1) determine holistically if AI is the best option, (2) ensure that data is protected, (3) document AI use, and (4) disclose AI use. These four guidelines serve to recognize and mitigate the biggest risks of AI use and to set the foundation for transparent use. The complete set of guidelines are divided into two categories: preparing to use AI and using AI by research phase; these categories are, in turn, divided into subcategories.

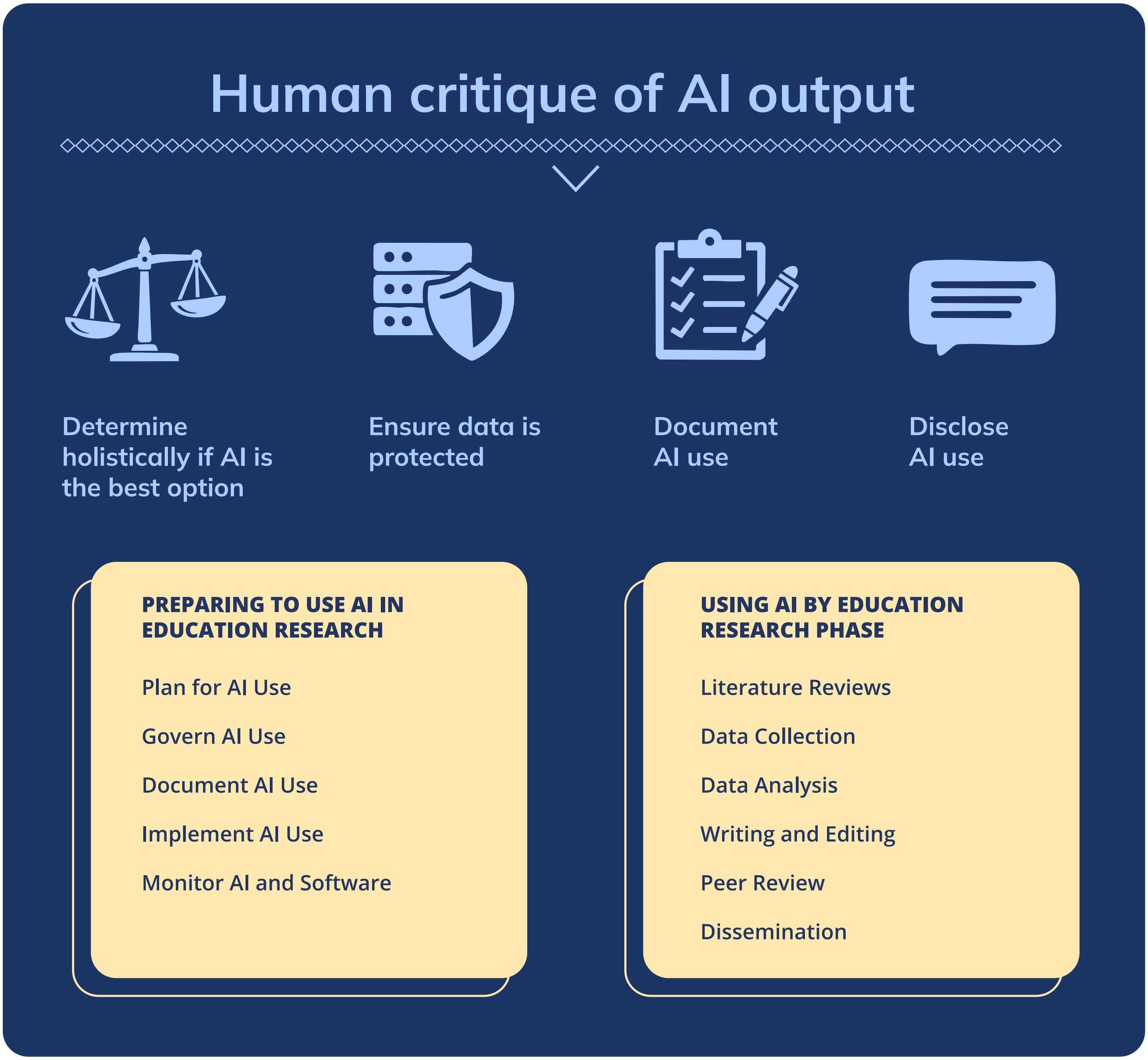


FIGURE 1: A summary of the guidelines for responsible AI use.

Project Website

Please visit the project's website at https://iacomputinged.org/graiser for more information and materials.

Chapter 1.

Introduction

Education research has seen tremendous changes over the last few decades, driven by expanded and refined methods as well as a broader understanding of what quality education research is and who it includes. Tools that support education research have also significantly impacted how education researchers conduct studies. For example, many qualitative researchers have moved from coding paper transcripts with colored highlighters to using sophisticated software. Now, AI functionality has been added to some of this software, with the promise of saving time and improving the analysis.

The evolution of AI-infused tools has brought a mix of optimism and concern, including for the accuracy of the results. Unsurprisingly, this mix applies to using AI in STEM education research (STEM-ER), particularly given the rapid proliferation of new AI tools. For example, ChatGPT, a generative AI tool, has been the most quickly adopted app in history, garnering over 100 million users just two months after its initial release (Hu, 2023). This stunning growth attests to the versatility of AI and the many possible use cases for tools that can generate responses based on natural language prompts. Not surprisingly, ChatGPT and other AI tools are becoming increasingly common in nearly all domains, including in various aspects of the STEM-ER process (Lee et al., 2025) such as conducting literature reviews and analyzing data.

AI has the potential to be useful across many phases of education research (Johri et al., 2023). For example, literature reviews are becoming increasingly labor intensive as knowledge advances, and AI-based tools might be used to assist with every step, from formulating the problem, to searching relevant databases,

STEM-ER and Researchers

These guidelines focus on the needs of *researchers* whose field of study is STEM (science, technology, engineering, or math) education research, abbreviated throughout as STEM-ER.

Throughout this report, the term *researchers* will refer to STEM education researchers (1) unless otherwise noted and (2) in instances where the term describes the results of studies about the behavior of researchers when using AI (in which case the researchers are normally from a variety of disciplines).

to screening articles, and to extracting and interpreting data (Wagner et al., 2022). Data analysis – whether that data is quantitative or qualitative, textual, audio, or visual – can be performed by various AI tools, where the ability to analyze very large data sets and to work with unstructured data can be harnessed (Hashimoto et al., 2024). Indeed, one strength of AI tools in STEM-ER is their capability to identify patterns and trends that may not be evident to human observers (Mulally, 2024).

However, using AI tools in STEM-ER is not without disadvantages, including the potential for inaccurate output, the extension of historical biases, and a negative environmental impact (Siau & Wang, 2020). Recent applications of AI tools to education research have found that using the tools may reduce the quality of a literature review (Ngwenyama & Rowe, 2024) and reduce transparency in the research process (Bolaños et al., 2024). If inaccurate AI output were to become common in STEM-ER, the results could include less effective instructional practices and policies (Roe, 2025), ultimately harming students and society. Plagiarism and data privacy are also concerns with using these tools in research contexts (Pack & Maloney, 2023; Santiago Jr. et al., 2023).

Researchers with at least a baseline familiarity with issues related to responsible AI use can make more sound decisions about whether, when, and how to use AI tools (Łodzikowski et al., 2024). But due to the tools' novelty, there is minimal guidance concerning their responsible use in the context of STEM-ER. Further, there is no consensus as to what types of AI usage are acceptable by researchers (Kwon, 2025).

Generative AI

These guidelines are focused on generative AI, although some of the guidelines will be relevant to other forms of AI as well. The term *generative AI* describes computing tools that, based upon the data they have been trained upon, can generate text, images, and/or audio (Feuerriegel et al., 2024). In this report, generative AI is referred to as simply "AI" for parsimony.

In November 2025, the Institute for Advancing Computing Education (IACE) convened experts in STEM-ER and in AI at a workshop to co-create guidelines for the responsible use of AI (see the appendix for more details about this process). Building upon previous work related to AI ethics and research ethics (Bos, 2020; Fox, 2022; Huang et al., 2022; Jobin et al., 2019; Johri et al., 2023; Katz et al., 2023), the guidelines are a response to growing calls to articulate what responsible AI use is (Borenstein & Howard, 2021). Work related to responsible AI use for education researchers has been conducted at a high level by the Community for Advancing Discovery Research in Education (CADRE) (Barnes et al., 2024) and broadly practice-focused by the American Institutes for Research® (AIR®) (American Institutes for Research, 2026). The current project complements these prior works by **providing specific and actionable guidance for STEM education researchers who are considering using AI in their research workflows.**

The Scope of the Guidelines

These guidelines address the responsible application of AI within STEM-ER.

Key use cases covered include education research processes such as:

- Identifying relevant literature for reviews.
- Generating and utilizing synthetic (i.e., AI-generated) data.
- Analyzing research data.
- Drafting or revising research studies.

Topics outside the scope of these guidelines include:

- AI education, such as teaching learners to use or develop AI in computer science programs.
- AI literacy, such as training learners on the use of AI for tasks (e.g., protein folding).
- Research on AI tools, such as investigating the effectiveness of AI applications (e.g., an AI tutor), but see the Researchers in Fields Supporting STEM-ER section in Recommendations for Adjacent Communities.
- Classroom AI use, such as AI usage by students or teachers for assignments (e.g., lab reports) or grading.
- Specific tool recommendations.

These guidelines focus on the *responsible* use of AI, not on AI use in general. Therefore, our team emphasizes preventing harm by raising awareness of concerns and detailing strategies to avoid negative outcomes. This emphasis does not imply that only negative outcomes come from AI use. Much as guidelines for keeping yourself safe while driving a car might seem to focus on the negative (e.g., the dangers of texting while driving) and not the positive (efficient transportation), our team has attempted to highlight the most significant cautions related to responsible AI use.

Finally, our team notes that this document is a product of a moment in time: late 2025 and early 2026. Recent months have seen widening adoption of agentic AI, and it seems reasonable to anticipate that advances in AI tools and their applications will continue apace. We expect – and hope – that future research will address the many unanswered questions about AI tools, their impact on the research process, and their impact on researchers.

Responsible AI and AI Ethics

The terms "responsible AI" and "AI ethics" are not defined or used consistently in public discourse, and some writers use the terms interchangeably. In this report, our team uses the term "responsible AI," which focuses on issues that are more immediately relevant to the user (e.g., governance policies). Responsible AI is distinct from AI ethics, which more broadly concerns issues that impact all members of society (e.g., protection of human rights) (Shah-Dand, 2025). However, responsible AI use often involves weighing ethical issues, such as the environmental impact of AI tools.

Chapter 2.

Critical Areas of Concern

Our team identified critical areas of concern related to AI use from a review of relevant literature and from ideas generated and refined at the workshop. Their order should not imply priority.

Critical Area 1: Accuracy and bias

AI can often offer impressive specificity and coherence, but it also sometimes struggles to produce accurate output (Barile et al., 2024; Borji, 2023; Hicks et al., 2024). For example, AI is known to generate citations for papers that do not exist or misrepresent the contents of a paper; this propensity may be a feature of the architecture of these models that cannot be eliminated (Xu et al., 2024). Several studies have found that AI can generate biased output related to demographic factors, such as gender, race, or disability (Smith, 2024; Zack et al., 2023; M. Zhou et al., 2024).

Critical Area 2: Data privacy

Interacting with most AI tools may necessitate sharing user data with the tool. That data, in turn, might be used to train future iterations of the AI tool (Wach et al., 2023), or it may be subject to a data breach. Users may or may not have control over this process, and data privacy concerns will likely have implications for IRBs. An additional concern is data re-identification (i.e., where powerful AI tools can combine disparate data sets and/or make inferences from limited data in order to identify an individual). Re-identification of data is an increasing concern in the age of AI, and this risk is higher for certain race/ethnicities and genders (Xia et al., 2023). In some cases, only a few pieces of data are sufficient for re-

Differences between AI Tools

Some material in this report is applicable only to certain AI tools and implementations. For example, data privacy can be a concern because many AI tools share data externally. However, some AI tools are run locally, meaning that all data processing occurs on local computers, and data may be limited to those who have access to the AI tool. With locally run tools, the risks to data privacy may be much lower. Similarly, the environmental impact of training and using tools will vary based on a range of factors. Thus, not all material in this report is equally applicable to all AI tools.

Researchers should be aware of the features of the AI tools that they use so that they can determine whether the data privacy, environmental, and other risks are acceptable for their use case.

identification (Rocher et al., 2021), and the re-identification risk continues to grow with the improvement of AI tools (Kancherla, 2020).

Critical Area 3: Lack of settled law and practice

Due to the novelty of AI tools, there is not yet an established body of law around issues related to their use. For example, as of November 2025, OpenAI is under court order to deliver millions of transcripts of ChatGPT discussions to the *New York Times* to assess potential copyright infringement (Brittain, 2025b). Similarly, in September 2025, Anthropic settled a lawsuit related to their use of others' intellectual property without permission as training data for their AI tool (Brittain, 2025a), but other similar cases have not resulted in compensation for use of intellectual property (Tobin, 2025). The legal landscape around AI use is not settled, and this creates some uncertainty for AI tool users. The decision of whether and how to use AI tools in STEM-ER may require research teams to consider if their use may open up any legal liability or other consequences, such as the disclosure of AI chat logs or participant data to a third party.

Critical Area 4: Environmental and human impacts

Generative AI usually requires substantive computational resources (Chen et al., 2025), although technological advances may reduce energy requirements (Fernandez et al., 2025; Rajput et al., 2025). These costs present a concern in light of the various societal costs related to high energy consumption (Berthelot et al., 2024). The rapid expansion of data centers to serve AI development and use can impact society – including students and their families – in many ways, including by increasing electricity bills and pollution. Further, the environmental impact of AI appears to be borne disproportionately by communities that are already experiencing worse environmental conditions (Marrinan et al., 2025).

In terms of human impacts, the most common concern is the potential for job loss (Wach et al., 2023), but there are additional concerns related to, for example, the mental health toll on low-paid workers in developing countries who screen AI input and output for objectionable content (Perrigo, 2023). Other wide-scale impacts of AI include greater ease in creating mis- and dis-information (Bontcheva et al., 2024). However, it is difficult to predict the extent to which various negative effects of AI will materialize (Zarifhonarvar, 2024): the lack of research on the impacts of AI and the potential for technological or policy changes to alter outcomes combine to make it nearly impossible to predict longer-term impacts of AI use. Thus, researchers face a challenge in aligning their AI use with their goals and values, not only due to a multifaceted cost-benefit analysis but also to the lack of information for conducting the analysis itself.

Critical Area 5: Use of intellectual property

Many AI tools rely on vast amounts of textual data to train their models. This training data often includes materials protected by copyright (Cooper et al., 2025; Lucchi, 2024) or gathered in violation of prohibitions on scraping data (Amarikwa, 2023). Whether this use of copyrighted material constitutes a violation of copyright is currently the subject of lawsuits (Grynbaum & Mac, 2023) as well as differing views regarding its appropriateness (Klosek & Blumenthal, 2024; Samuelson, 2023).

Critical Area 6: The digital divide and related systemic factors

The costs of training AI and the cost of accessing the newest and most powerful models can be considerable. Additionally, language barriers, cultural context, and infrastructure costs may contribute to a digital divide that impacts the design, deployment, and use of AI tools. As a result, the creation and use of the best AI models and tools may be limited to a small group (Sathish et al., 2024). Concerns about a digital divide pre-date generative AI, but lack of access to modern AI tools – especially those with greater accuracy or privacy protections – may exacerbate the digital divide (Twinomurinzi & Gumbo, 2025), both in general as well as between researchers. Relatedly, systemic issues might encourage inappropriate AI use. For example, some researchers might choose to use AI to peer review research, despite the fact that some publication venues do not permit it. The systemic issue underlying the pressure to use AI likely stems from extreme burdens on researchers' time (Paris et al., 2025) – which might be mitigated in a variety of ways that do not involve AI use (e.g., limiting the number of papers that can be submitted to a conference; requiring authors to review other papers).

Critical Area 7: Transparency and replicability

The processes by which most AI tools determine their output is not transparent to the user (Hashimoto et al., 2024), and even if the user asks, they may get an inaccurate explanation (Turpin et al., 2023). This issue can compound other concerns. For example, it is often not possible for a user to determine whether and how there is bias in AI output due to the lack of transparency. This lack of transparency may lead to problems with the replicability of scientific research (Yan et al., 2024).

Critical Area 8: Impacts on researchers' knowledge and thinking

AI use may present several challenges to researchers' knowledge and thinking.
Related to the lack of transparency, the use of AI tools can lead to the illusion of understanding where deep understanding is lacking (Messeri & Crockett, 2024), leading to an intellectual debt (i.e., a situation where there is greater knowledge of *what* works than of *why* it works), which tends to stymie future scientific progress. By analogy, it is sometimes the case that medical researchers determine that a medicine is effective without knowing how it works; this knowledge gap may make it impossible to know when use of the drug is contraindicated, constituting an intellectual debt. Zittrain outlines three concerns for AI-related intellectual debt (Zittrain, 2022). First, it is difficult to determine how the tool will respond to novel circumstances, such as data from a student population that differs in some way from its training data. Second, AI-generated data will be used to train new AI tools, compounding intellectual debts. Finally – and most germane to STEM-ER – intellectual debt implies an emphasis on empirical data to the exclusion of theory or conceptual understanding of phenomena, and this lack of theoretical development may substantially slow the progress of research.
More frequent AI tool use has been shown to be negatively correlated with critical thinking skills, leading to the conclusion that there is a "need for strategies that balance the benefits of AI integration with the development of independent analytical skills, particularly in educational and organisational settings" (Gerlich, 2025, p. 12). Similarly, some research has found that AI use can lead to decreases in cognitive performance (i.e., deskilling) (Budzyń et al., 2025; Ehsan et al., 2026), which is problematic given that researchers will need to vet the output of those tools, and the ability to do so well may be reduced by deskilling from using the tool.

There may also be differences in knowledge acquisition when AI tools are used, with some evidence showing that learning about a topic from AI output results in less depth of knowledge than learning about the topic via a traditional internet search (Melumad & Yun, 2025). As a result, researchers might consider not just the impact of AI tool use on their present workflow but also the broader impacts – such as shallower learning – of AI tool use.

Intellectual conformity is an additional risk: when AI is used, several factors may contribute to the homogenization of ideas, which may be detrimental to the advancement of research (Messeri & Crockett, 2024) and to free inquiry in general.

The statistical nature of generative AI implies that the ideas most common in its training data are most likely to be reflected in its output, sidelining less common ideas in the process – a phenomenon that may become even more prevalent as AI-generated content is used to train new AI models – resulting in the cyclical reinforcement of dominant perspectives.

Further, there is evidence that some AI tools produce output that does not reflect the full variety of their training data (F. Wu et al., 2024), such as outputting a positive book review when the training data contains mixed reviews. Messeri and Crockett use the analogy of monocultures (Messeri & Crockett, 2024) to describe the risk that AI use by researchers poses: much as growing only one crop puts the food supply at risk (e.g., to disease), AI use may create a scientific monoculture, where the kinds of questions, data, and analyses best suited to AI tools come to dominate the discourse. Some research shows that AI tools are less likely to output sound but uncommon methodologies, effectively narrowing the field of scientific endeavor (Wright et al., 2026).

Critical Question

Will my use of AI contribute to the creation of a norm where other teams will feel compelled to use AI? For example, if AI permits faster data analysis and therefore quicker publication of results, other teams may feel pressure to use AI tools, or if publications are created with AI, peer reviewers may feel pressured to use AI to keep up with the rate of publication. If the norm in our field changes, what will be gained and what will be lost?

Critical Area 9: Impact on productivity

There is no consensus on the impact of AI tools on productivity, with some studies showing improvements (Dell'Acqua et al., 2025) and others no effect (Humlum & Vestergaard, 2025) or even reductions. For example, Becker et al. found harms to productivity as well as misperceptions: software developers anticipated that using AI would reduce the time needed to complete a task by 20%, but it instead *increased* the time needed by about 20% (Becker et al., 2025). Hao et al. identified productivity benefits to AI use by researchers – but at the expense of a narrowing of focus of research topics (Q. Hao et al., 2025). Further, best practice for responsible AI use involves critiquing AI output (see below) – a process that may not save time when compared to generating the output without AI.

Table of Contents

Critical Area Interactions

Critical concerns can interact in ways that amplify each other. For example, accuracy and bias issues may be exacerbated by a digital divide that minimizes the participation of some in the development, deployment, and use of AI tools. And the digital divide may expand if problems with accuracy and bias lead some to avoid AI tools.

Critical Question

How will I determine if the perspectives of the AI tool's creators are influencing the tool's output?

As a result of these critical concerns, responsible AI use avoids defaulting to AI use. Rather, researchers will be prepared to justify their rationale for choosing to use AI for any given task, much as they would be prepared to justify their choice of research design or of a specific statistical approach for data analysis.

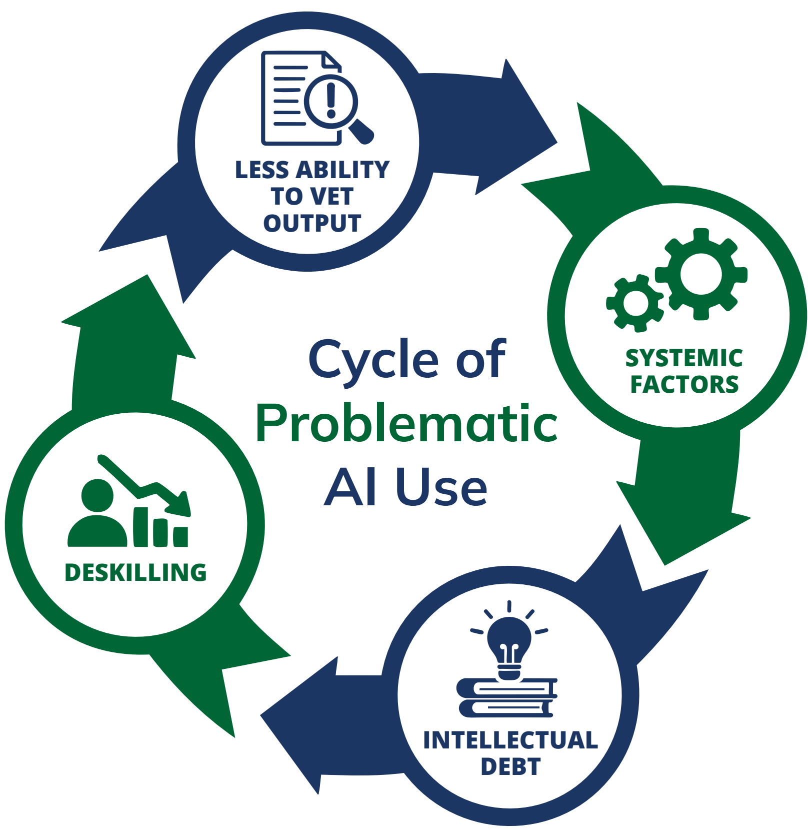


FIGURE 2: Cycle diagram showing how problematic AI use can reinforce itself through four connected issues.

Responsible use of AI in STEM-ER will first consider the questions

- *What is the argument for using AI in this research?*
- *Will AI use contribute something that can't be gained without it?*
- *What risks does AI use pose, and are the benefits worth the risks?*

The decision to use AI involves consideration not just of the immediate use case but of the broader impacts of widespread AI use on STEM-ER – as a field, on the research process, and on individuals who interact with the research in any way (e.g., researchers, study participants, education leaders).

AI as a Tool for Adversarial Review

In evaluation research, a critical friend is one who "asks provocative questions, provides data to be examined through another lens, and offers critique of a person's work as a friend" (Costa & Kallick, 1993, p. 50). AI tools may be able to serve as critical friends, although one might object to the anthropomorphizing involved in describing AI as a "friend" and frame it instead as an adversarial reviewer (Woodruff et al., 2026). However, many AI tools are not designed to be critical or adversarial but instead display signs of sycophancy in their output (Fanous et al., 2025), which may lead researchers to evaluate their own ideas too positively (Roe, 2025). Thus, using AI adversarially will require researchers to carefully prompt for critical feedback. Example use cases of AI for adversarial review include:

- **once a researcher has drafted research questions, they might prompt an AI tool to critically evaluate the questions for unintended bias in their framing (while keeping in mind that the tool itself may have biases).**
- **a research team might upload the research plan section from a newly-awarded grant, asking the AI tool to identify possible threats to successful implementation of the study.**
- **researchers could ask an AI tool to identify aspects of new survey questions that might be unclear, ambiguous, or unfamiliar to the target audience.**

Chapter 3.

Preparing to Use AI in STEM Education Research

3.1 The Key to Responsible AI Use: Researcher Critique

There is one overriding principle of using AI in STEM-ER: **responsible AI use requires that a researcher with relevant knowledge and skills review AI output before using it.** Education researchers have the responsibility to vet all aspects of their work before distribution, which requires the researcher to have adequate knowledge and experience in research methods, domain knowledge, and other areas. This responsibility does not change when AI tools are used. If the researcher (or research team) does not have the ability to vet the AI output, either the AI tool should not be used, or a researcher with the appropriate competency should be consulted.

AI Training Data Is Pretty WEIRD

AI training data is largely from western, educated, industrialized, rich, and democratic (known as WEIRD) nations (K. Zhou et al., 2025). This is not surprising because people from WEIRD nations are more likely to contribute to online data sources (e.g., Reddit posts, social media, academic research, news articles) and to participate in research studies. For example, participants in learning analytics research studies are more likely to be from WEIRD countries (Baek & Doleck, 2024). The topics investigated in these studies differed as well, with research in non-WEIRD settings more focused on the social nature of learning and research in WEIRD settings more focused on self-regulated learning and feedback (Baek & Doleck, 2024).

WEIRD participants are also disproportionately represented in research related to sex/gender (Klein et al., 2022), psychology (Nielsen et al., 2017), and cognitive science (Blasi et al., 2022). There is evidence of differences between individuals from WEIRD and not-WEIRD countries in terms of their motivation (Medvedev et al., 2024), personality (Schmitt et al., 2007), and moral foundations (Atari et al., 2023). Further, even within WEIRD countries, minoritized communities may not be well represented in the training data. Thus, the disproportionate presence of training data about WEIRD people is likely to constitute a limitation on the accuracy of AI output for other cultural contexts.

3.1.1 Why Human Critique is Crucial

Biased training data may lead to bias in AI-supported education research. The output of generative AI tools depends on their training data. If the training data is biased or otherwise incomplete, then the output will reflect these shortcomings. For example, if there are retracted papers in the training data set, then the output could reflect the content of those papers. Similarly, if an AI model is not adequately trained on specific uses of language, dialect, or tasks that researchers are asking it to analyze, then it may produce inaccurate results. Human bias can be exacerbated by the use of biased AI tools, leading to a cycle of bias amplification between humans and AI (Glickman & Sharot, 2025) – a cycle that may be particularly pernicious given the potential for newer models to be trained on AI-generated data.

AI model architecture shapes the output. AI tools are limited by their design. For example, text output from LLMs involves statistical predictions of what words are most likely to occur next given a specific input and an LLM's training data. An LLM, by nature of its design, has no mechanism to determine the accuracy of its output text. Some assume that future technological advances will eliminate the problems of verifiability, transparency, and reliability of LLMs. That may or may not be the case: some AI researchers hold that LLMs will always struggle with errors of fact (Banerjee et al., 2025), including fabrications.

AI Fabrications

It is common to refer to AI 'hallucinations,' but this term is not technically accurate (as hallucinations are a sensory perception that occurs without a stimulus) and may be stigmatizing (Østergaard & Nielbo, 2023). Throughout this report, our team refers instead to AI *fabrications*.

Lack of transparency in AI tools reduces education research replicability and transparency. Most AI tools are not transparent about what data sources are used in their training and how their model uses and processes data to generate output. Additionally, the design of many AI models and tools is such that the same prompt will not necessarily yield the same output. Therefore, using these tools can present a challenge to the transparency and replicability that are central aspects of the research process. Many AI tools have an element of randomness, which may interfere with transparency and replicability.

Critical Question

Some scholars have argued that the norm of transparency in scientific research requires that any AI tool used must create reproducible output (Guest, Suarez, Müller, et al., 2025). **Can the responsible use of AI include using tools that create output that is not necessarily reproducible?**

3.1.2 Critiquing AI Output

One challenge of critiquing AI output is that the researcher must have sufficient knowledge, skills, and self-efficacy to perform the critique, but these attributes tend to decline as task complexity increases; the end result may be over-confidence in AI output (Hümmer, 2025). Additionally, when choosing whether to use AI tools within the research process or not, researchers should consider whether the time required to vet AI output will exceed the time required to generate the content without using AI (Roe, 2025).

LLMs, with and without RAG

Retrieval-augmented generation (RAG) approaches can improve LLM performance by incorporating additional information (e.g., a webpage, a database). LLMs with and without RAG have different advantages and disadvantages. For example, LLM output accuracy will be limited by the cut-off date of its training data (e.g., when describing recent events), and LLM errors may be reduced by RAG tools (e.g., if the tool verifies that a citation is to research that exists before outputting it). However, some RAG tools may have additional data privacy risks (Jiang et al., 2025; Liu et al., 2025), and they may still have error rates of up to one-third (Magesh et al., 2025; K. Wu et al., 2025). Familiarity with these features of LLMs with and without RAGs is important to understanding the limitations and capabilities of the AI tools that incorporate them.

The adoption of several mindsets related to AI will promote thorough vetting of AI output. First, ***the overall approach to critiquing output should be critical, skeptical, and thorough* –** not pro forma. For example, if a researcher uses an AI tool to generate a list of topics to include in a literature review, the expert review should consider the following:

- *What topics are missing from this list?*
- *What topics should be removed from this list?*
- *Should any of the topics be combined? Should any topic be separated into two or more topics?*
- *Should the topics be reframed?*
- *What is the most logical order to address these topics in the review?*
- *Are there other researchers (i.e., with more or different knowledge and skills) that should help me evaluate this list?*

In other words, critical review and evaluation by human reviewers should occur throughout the workflow.

Do not assume that AI output will be consistent. Output from repeating the same prompt may not be the same due to the probabilistic nature of most AI tools, though it may be possible to mitigate this problem somewhat. Output may also be inconsistent in other ways, including logical inconsistencies. For example, one study found overt rejection of racial stereotypes coupled with extreme covert stereotypes (based on racially-related dialect usage) in AI output (Hofmann et al., 2024). So lack of bias in one type of output would not be indicative of lack of bias in another type of output from the same tool.

Also, ***be skeptical about an AI tool's explanation of reasoning.*** "Chain of thought" is a prompting technique that has been shown to improve LLM accuracy (Wei et al., 2022). In this approach, the user first demonstrates the steps taken to solve an example problem, then asks the AI to solve a similar problem. However, given the nature of how LLMs operate, asking an AI tool to reverse engineer the steps that it used to solve a problem does *not* mean that the chain of thought will accurately represent the process by which the tool determined the output (Turpin et al., 2023) and so should not be relied upon as part of the researcher's evaluation of AI output. Similarly, prompting an AI tool to explain its output may not result in accurate explanations (C. Agarwal et al., 2024).

Relatedly, ***be skeptical about the content of sources referenced in AI output.*** In a study of seven AI tools given tasks related to medical knowledge, Wu et al. found that 50% to 90% of the AI output was *not* supported by the sources cited in the output (K. Wu et al., 2025). ***Consider that newer AI models or tools may not have better output.*** Interestingly, one study of the ability of LLMs to summarize scientific research found that older models tended to outperform newer ones (Peters & Chin-Yee, 2025).

Researchers can also adopt specific practices for critiquing output. First, it is helpful to ***center human thought by generating ideas before consulting AI output.*** Some research shows improved results for idea generation and user self-efficacy when an AI tool is used *after* human ideation instead of before (Qin et al., 2025). Further, as with any tool, the use of AI will impact the user. Although research on these impacts is still in the early stages, there is already some evidence that current AI usage is inversely correlated with critical thinking (Gerlich, 2025).

Similarly, as a result of the limitations of AI tools, a researcher's view of the possible scope of a topic may be restricted to the most common topic(s) in the training data. Therefore, human ideation (e.g., generating possible research questions, interpreting research data) should precede examining AI output. Next, ***document all aspects of the AI usage as well as of the human review.*** Comprehensive notes that document the AI prompts and human review of all output will ensure that the process is verifiable, defendable, and transparent.

Research requires careful use of terminology, especially technical terms. An AI tool may or may not produce output that reflects appropriate use of specialized terminology. ***Review word choices that are important in an educational research context*** (e.g., a 'significant' impact, language describing correlation versus causation, deficit versus asset framing) to mitigate this concern.

Another helpful approach is to ***use prompts that solicit multiple outputs*** instead of just one. Compare these two AI prompts: (1) "What are six different ways to describe self-efficacy in a consent form?" and (2) "Describe self-efficacy for a consent form." With the first prompt, the researcher is required to exercise their discretion as opposed to engaging in cognitive offloading. With the second prompt, the researcher's role is reduced to accepting or rejecting the output, and their role in making choices between possible outputs is not as clear. Therefore, it is often preferable for the researcher to review and select from multiple outputs. Similarly, ***consider using more than one AI tool*** where it doesn't increase risks (e.g., to participant privacy). For example, if using AI to provide synthetic data (i.e., AI generated data) in response to an open-ended survey question, use multiple AI tools to gain a better understanding of the landscape of possible responses.

An essential part of critiquing AI output is to ***check for inaccuracies and biases.*** Many research studies have found evidence of various forms of bias in AI outputs, including race, gender, socioeconomic status, and disability as well as various cognitive biases (Gadiraju et al., 2023; Hofmann et al., 2024; Omar et al., 2025; Salinas et al., 2025, 2025; Smith, 2024). The presence of these biases is perhaps not surprising, given that AI tools generate output based on their training data. ***Checking for evidence gaps is also necessary***, since AI output is limited by its training data. Absence of evidence should not be interpreted as non-existence. For example, if an AI tool's training data did *not* include paywalled research studies, a researcher prompting the tool to identify gaps in the literature may generate output identifying gaps that do not actually exist.

It is important to ***establish attribution of any AI-generated ideas***. Most AI tools do not provide the source of ideas in their output, contravening academic conventions. In many cases, researchers will need to investigate whether an idea is a novel creation of the AI tool or whether previous researchers should be credited with the idea (Lemley & Ouellette, 2025) in order to conform to academic norms and to ensure proper attribution. Because it is likely that retracted papers, out-of-date findings, AI generated content, and other low-quality information is part of an AI tool's training data, it is important to source ideas before using them.

Another step in critiquing output is to ***determine whether any intellectual debt accrued is acceptable***. Intellectual debt stemming from AI use can be problematic (see above), but sometimes such debts are appropriate or even helpful (Zittrain, 2022) because they may permit research to advance more quickly (as presented in the sidebar Two Cases of Intellectual Debt).

Prepare t***o defend the content of any output that will be used.*** For example, if an AI tool recommends a specific statistical data analysis approach, the researcher may be asked about it during a conference presentation and should be well-equipped to justify the approach.

Finally, when critiquing AI output, ***do not assume that suboptimal AI output is the fault of the user***. Despite recent advances, AI tools have several important limitations, and researchers (particularly newer researchers and those not in leadership positions) should not be criticized for documenting or reporting suboptimal outcomes. Rather, all reports of unhelpful AI outcomes should be carefully evaluated to determine to what extent the AI tool and the researcher bears responsibility so appropriate steps can be taken to improve future performance. A climate that stigmatizes reports of suboptimal outcomes will not encourage accurate reporting and may hamper future research efforts.

It can be useful for education researchers to work closely with AI researchers and computer scientists when possible to anticipate any structural and methodological shortcomings of a study that uses AI. However, many AI tools – including those based on LLMs – are by definition probabilistic. AI is but one of many tools in a researcher's toolbox: it can indeed provide some evidence on phenomena, but there are limitations on what questions AI methods can reliably answer. Just like most questions in education research, investigating phenomena requires triangulation and multiple methodologies to gain a more complete picture.

Two Cases of Intellectual Debt

Consider these two hypothetical cases of accruing intellectual debt:

A researcher uses an AI tool to quickly generate numerous responses (i.e., synthetic data) to a new survey instrument designed to assess a student's intention to pursue a career in biology, using those responses to pilot and improve the instrument before testing it with students. In this case, because the AI output doesn't constitute the final assessment on the instrument's reliability or validity, the researcher's inability to know precisely how the AI tool generated its responses is less important.

A researcher uses an AI tool to analyze video recordings of the implementation of a new biology curriculum in elementary school classrooms to determine which activities were most engaging to students; the tool's output is a student engagement score assigned to each activity, with no indication of how the score was determined. This output would involve substantial knowledge gaps (e.g., were facial expressions or eye tracking used to determine engagement? How does the scoring formula treat outlying data?), and these gaps will limit the utility and generalizability of the findings. For example, it will be unclear how the score would vary if the classroom were composed of students with different cultural expectations for eye contact or for expressing strong emotions in school. In this case, the researcher might conclude that the intellectual debt from this tool use is too high to meet their goals.

Guidelines for Critiquing AI Output

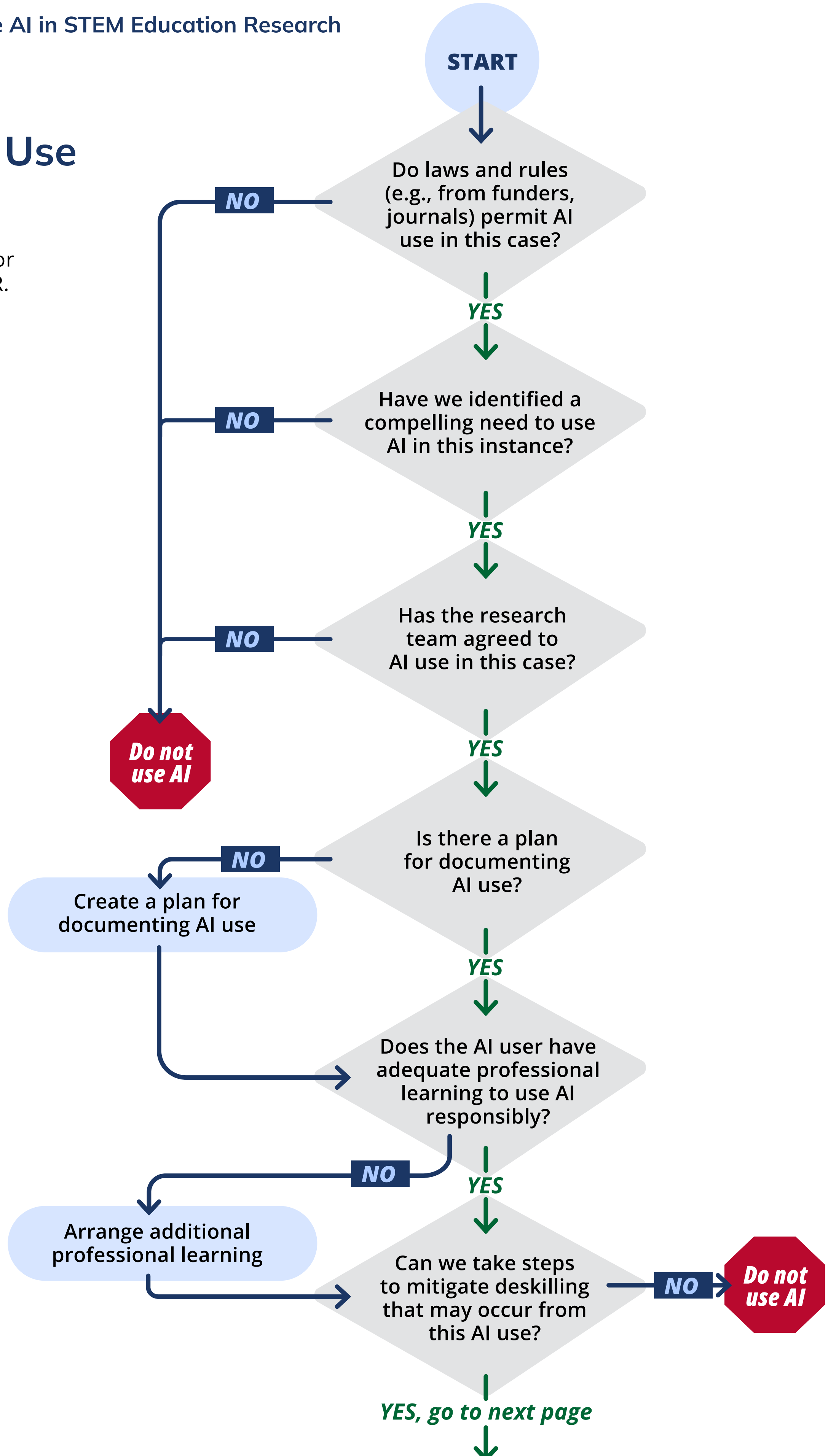
- [ ] Cultivate a mindset for critiquing AI output, including:
    - Take an approach to critiquing output that is critical, skeptical, and thorough – not pro forma.
    - Do not assume that AI output will be consistent.
    - Be skeptical about an AI tool's explanation of reasoning.
    - Be skeptical about the content of sources referenced in AI output. Check the sources carefully – be sure they exist and that they say what the AI claims that they say.
    - Consider that newer AI models or tools may not have better output.
- [ ] Center human thought by generating ideas before consulting AI output.
- [ ] Document all aspects of the AI usage as well as of the human review.
- [ ] Check for evidence gaps. If an AI tool claims, for example, that something doesn't exist, it may be the case that it doesn't exist in its training data.
- [ ] Establish attribution of any AI-generated ideas, with careful attention for ideas from retracted papers or disreputable or outdated sources.
- [ ] Determine whether any intellectual debt accrued from the output is acceptable.
- [ ] Prepare to defend the content of any AI output.
- [ ] Do not assume that suboptimal AI output is the fault of the user.
- [ ] Review word choices that are important in an educational research context.
- [ ] Use prompts that result in multiple outputs where possible.
- [ ] Consider using multiple AI tools when this won't increase risks (e.g., to participant privacy).
- [ ] Check for inaccuracies and biases.

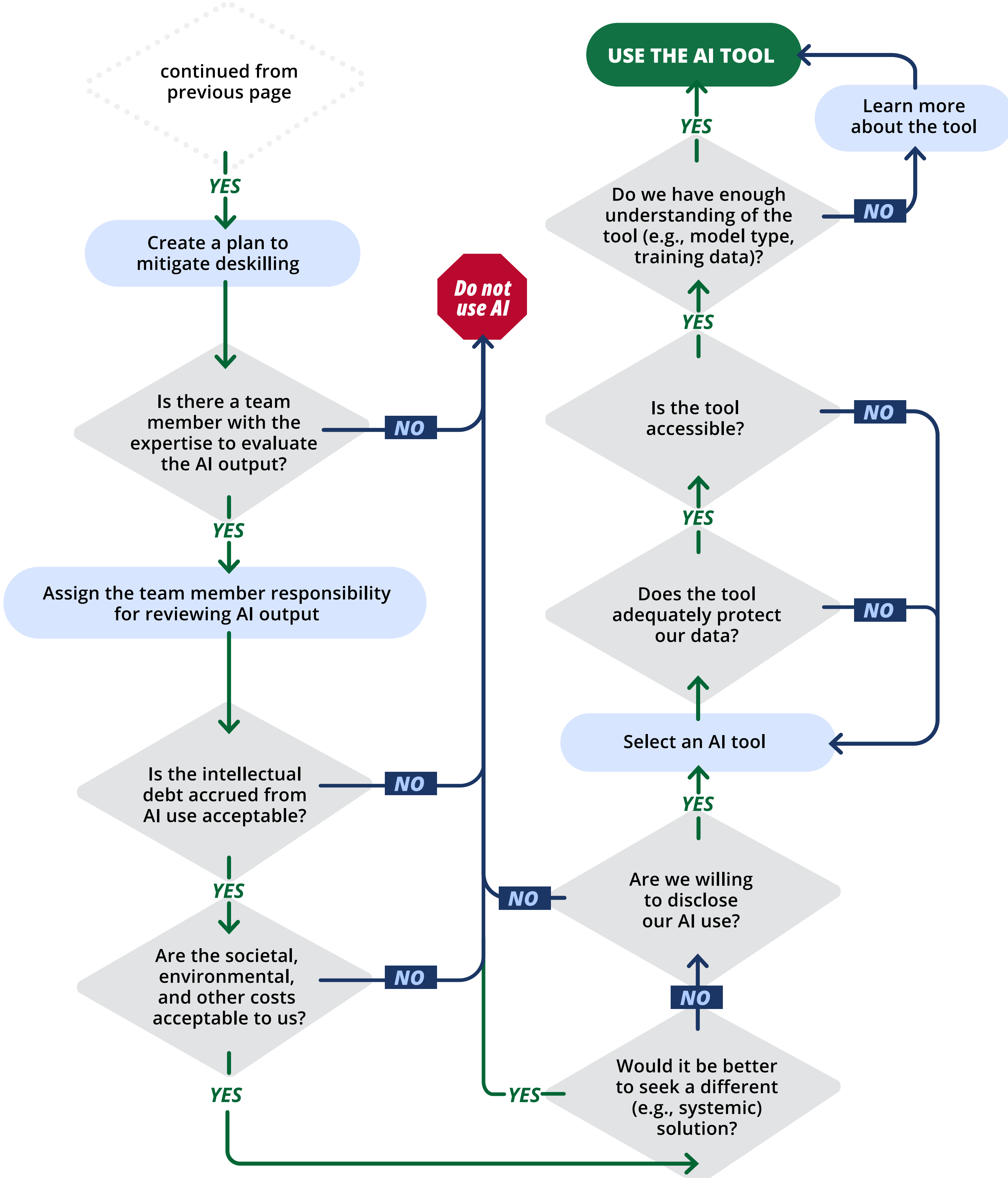
Critical Question

Responsible use of AI tools requires that a researcher with appropriate knowledge and skills evaluate AI output before using it. **How will new researchers develop those knowledge and skills if AI tools are completing the tasks that develop these skills, such as summarizing research literature?**

Flowchart for Responsible AI Use

FIGURE 3: The flowchart summarizes the guidelines presented in this document for responsible AI use in STEM-ER.




continued from previous page
YES
Create a plan to mitigate deskilling
Is there a team member with the expertise to evaluate the AI output?
NO
Do not use AI
YES
Assign the team member responsibility for reviewing AI output
Is the intellectual debt accrued from AI use acceptable?
NO
YES
Are the societal, environmental, and other costs acceptable to us?
NO
YES
Would it be better to seek a different (e.g., systemic) solution?
YES
NO
Are we willing to disclose our AI use?
NO
YES
Select an AI tool
Does the tool adequately protect our data?
NO
YES
Is the tool accessible?
NO
YES
Do we have enough understanding of the tool (e.g., model type, training data)?
NO
Learn more about the tool
YES
USE THE AI TOOL

3.2 Planning for AI Use

Many researchers have discovered the benefits of carefully planning research studies, often based on experience with projects that lacked such planning. Similarly, careful preparatory work before using AI tools can help avoid many challenges arising from ad hoc tool use.

First, there are several concerns that should be considered and discussed as planning begins. An overarching principle of these discussions is to ***ensure that potential risks and harms of AI use are taken seriously, and be open to choosing to forgo using AI.*** Guest et al. warn of the practice of "critical washing," where discussions of AI-related harms are articulated and then quickly dismissed "as if merely listing the harms to people and planet is enough to absolve us of responsib[ility and] dissolve the damage" (Guest, Suarez, & van Rooij, 2025, pp. 6–7). In contrast, over 400 qualitative researchers have decided that using AI for conducting types of research that require "a subjective, positioned, and reflexive researcher," is "not methodologically congruent" (Jowsey et al., 2025, p. 1), in addition to concerns over human and environmental impacts. If your research team identifies substantive harms or risks, be open to the conclusion that AI should not be used in that case.

Responsible use of AI is built upon knowledge. It is helpful to ***understand the architecture of AI tools*** – at least in the most general terms. (See the sidebar "LLMs, with and without RAG.") Much as researchers are expected to conceptually understand the statistics that they use to ensure that they are using them appropriately, researchers should also conceptually understand the AI tools that they choose to use. This understanding is important for selecting appropriate tools and for clearly conveying information about AI use (e.g., as part of informed consent, in research publications) as well as ensuring success and safety when using AI.

Research teams also can engage in ***discussing the implications of using a tool that is not entirely accurate.*** A calculator or a spreadsheet might yield inaccurate results based on faulty data or user error, but most AI tools introduce an additional type of error stemming from the probabilistic nature of their output.

It is also important to ***explore how using AI impacts team members***. For example, some research shows that users of AI tools are perceived as less competent or less motivated (Reif et al., 2025) – and these perceptions may have an oversized impact on some research team members such as graduate students. Additionally, it is important to ***consider that researchers whose first language is not English may have different perspectives on AI***. On the one hand, AI has the potential to democratize the publication process; a study of acknowledgements of ChatGPT in academic publications found that roughly two-thirds of those acknowledgements were from countries with a primary language other than English (Kousha, 2024). At the same time, there is some evidence that some AI tools produce lower quality output for users with less English proficiency, less formal education, or who were born outside of the U.S. (Poole-Dayan et al., 2025).

It is also important to ***mitigate security concerns*** stemming from sharing data with AI tools as well as the distinct security issues that can arise with AI use (Mohawesh et al., 2025). Research on generative AI-related security concerns is still in its infancy, but there are already several new attack vectors. For example, through prompt injection attacks, an AI tool could be manipulated to disclose unauthorized data (Rababah et al., 2024), potentially exposing confidential participant information. Another threat is model inversion: an attacker reverse engineers confidential data that is part of the model's training data (Teo et al., 2024). Data poisoning presents an additional risk, involving an attacker modifying data that is later used for training, reducing the accuracy of a model's output (Teo et al., 2024). To provide some mitigation against data privacy concerns, ***where possible, use local AI tools*** (i.e., tools that run on local hardware and that limit sharing of data to those who may have access to the tool).

Security concerns are particularly acute for agentic AI tools (i.e., AI tools that can act across platforms and with greater autonomy). For example, a researcher might use agentic AI to maintain their schedule across multiple platforms, with the agent reading and replying to emails, adding events to their calendar, and booking meeting rooms on campus. But the access (e.g., to email and university log-in credentials) required to enable agentic AI creates vulnerabilities (Datta et al., 2025). In this hypothetical case, the researcher might not consider the implications for the security of research participants' data (e.g., transcripts of interviews with study participants) because they are using the agentic AI for their personal productivity, but it is entirely possible that the access granted to the AI could be abused in order to access sensitive participant data. Therefore, it is important to ***be particularly careful with agentic AI, and avoid using it in systems with sensitive data***. Consult with your institution's IT department to find out how your use of agentic AI tools might present new risks.

A data flow diagram may be helpful to support discussions of how project data will interact with AI tools. A data flow diagram visualizes project data at each stage from collection to analysis to destruction, allowing team members to identify and mitigate risks at each step (Li & Chen, 2009).

Finally, ***consider how others will react to the decision to use AI***. The perceptions of research study participants are particularly important, particularly when they are choosing to participate and are concerned about their privacy. These perceptions may be influenced by public discourse around AI; Mahmoud et al. found that the main themes of such discourse were concerns about job loss, possible improvements in solving problems, limits in some types of intelligence (e.g., creativity), increased misinformation, existential risks, and privacy risks (Mahmoud et al., 2025).

Additionally, communities have attitudes toward AI shaped by their previous experiences with use of their data and their role in research (see sidebar CARE: Indigenous Principles for Data Governance.) These views may shape willingness to participate in research projects. Also consider how to respond to a graduate student or a study participant who joins the project at a later date and who objects to the use of AI. Finally, consider how the AI use might shape perceptions of the research study itself; ask yourself, Will its use undercut the transparency of the methodology and lead the work to be evaluated as lower quality than it might otherwise be?

CARE: Indigenous Principles for Data Governance

A group of organizations and individuals, the International Indigenous Data Sovereignty Interest Group, articulated principles to guide the use of data (Carroll et al., 2020) Although not specifically focused on AI, the principles suggest questions to explore when making decisions about data use involving AI:

- ***Collective Benefit:*** **Who benefits from the data collection, use, and reuse? For example: will the data help the community from which it was collected to make data-driven decisions?**
- ***Authority to Control:*** **Who will control future uses of the data? For example: Who will decide whether and how this dataset is used in a project in ten years' time?**
- ***Responsibility:*** **Who will take responsibility to ensure that the data use is part of a respectful relationship between the source of the data and the STEM-ERs? For example: will research findings be shared in the primary language of a study's participants?**
- ***Ethics:*** **What benefits and harms might stem from this data, and who will decide which benefits and harms to consider? For example: how will evidence of environmental harm be weighed against benefits of improved productivity?**

Guidelines for Planning for AI Use

- [ ] Ensure that potential risks and harms of AI use are taken seriously, and be open to choosing to forgo using AI.
- [ ] Understand (in general terms) the architecture of any AI tools used.
- [ ] Where possible, use local AI tools.
- [ ] Discuss the implications of using a tool that is not entirely accurate.
- [ ] Explore how AI impacts team members (e.g., perceptions of their competence).
- [ ] Consider that researchers whose first language is not English may have different perspectives on AI.
- [ ] Understand and plan to mitigate security concerns rising from AI use.
- [ ] Be particularly careful with agentic AI, and avoid using it in systems with sensitive data.
- [ ] Consider how others will react to the decision to use AI, including potential study participants and graduate students or others who join the project team at a later date.

Critical Question

Do the anticipated outcomes of using AI justify the resources required to use it responsibly (e.g., time vetting AI output)?

3.3 Governance Issues

Good research practice requires adherence to relevant laws and rules (McGill et al., 2023), and this principle applies to researchers using AI tools. Governance issues related to AI can fall into three categories: external restrictions, communicating use externally, and internal agreements about AI use. Preparing in advance for all three can help avoid situations where AI use does not align with internal or external restrictions.

Various groups – from governments to funders to school districts to publishers – have laws, policies, or norms related to AI use, and the responsible use of AI involves compliance with these external requirements. Also, some laws that pre-date the rise of generative AI (e.g., Family Educational Rights and Privacy Act (FERPA), Children's Online Privacy Protection Act (COPPA)) apply to data used by AI tools. Be prepared to ***observe any restrictions on AI use from organizations, funders, governments, and other relevant parties***. For example, it is important to ***understand the copyright implications*** of AI-generated content: AI output (i.e., in response to a prompt) without human modification is generally not protected by copyright (United States Copyright Office, 2025). It is best practice to ***consult the AI use policies for publication and presentation venues*** (e.g., conferences, journals) where findings may be shared to ensure that the use aligns with the venue's requirements. For international contexts, be aware of location-based AI-related copyright laws. Because requirements change frequently, it is also advisable to revisit the venue's AI use policies periodically, especially during longer projects.

Determine when and how AI use will be disclosed to others, such as funders or in publications. However, be aware that it is not always evident to a user when and how software – such as a platform for searching literature or managing citations – is using AI. As with other aspects of the research process, responsible AI use requires transparency. Researchers should aim to be transparent and thorough in their disclosure of how AI was used in the research process.

In terms of internal governance, it is important to ***reach agreement with all team members before a research study commences on whether, when, how, and what AI tools will be used*** for research tasks. Do not presume agreement on what constitutes responsible AI use: there is wide divergence in opinion as to what types of AI use are or are not acceptable in the research process (Kwon, 2025). ***Consider implementing a signed agreement for all team members clarifying that they will not use AI unless it is in accordance with the team's agreed-upon usage terms.*** It is possible that a team member might use AI (intentionally or otherwise) outside of the agreed terms. ***Consider whether and how to determine if a research team member uses AI in ways that violate the team's agreement and how to respond to such use.*** This determination of use is especially important if it presents threats to participants' data privacy or is in violation of laws and rules. Be aware that AI detectors are not always accurate (Erol et al., 2025), especially for users whose first language is not English (Liang et al., 2023). Explore whether frequent instances of AI use that violate norms are indicators of a systemic issue, such as unreasonable productivity demands.

Guidelines for Governing AI Use

- [ ] Observe any restrictions on AI use from organizations, funders, governments, and other relevant parties.
- [ ] Understand the copyright and other legal implications of AI-generated content.
- [ ] Consult the AI use policies for publication and presentation venues.
- [ ] Determine when and how AI use will be disclosed, aiming for transparency.
- [ ] Reach agreement with all team members before a research study commences on whether, when, how, and what AI tools will be used for research tasks.
- [ ] Consider implementing a signed agreement for all team members clarifying that they will not use AI unless it is in accordance with the team's agreed-upon usage terms.
- [ ] Consider whether and how to determine if a research team member uses AI in ways that violate the team's agreement and how to respond to such use.

3.4 Documenting AI Use

Sound research practice requires maintaining appropriate documentation (Antunes & Hill, 2024) throughout the research process. This principle applies to AI use as well: a key tenet of responsible AI use is to ***maintain a written record of how it is used***. This record will ensure that researchers working 'downstream' of previous AI use understand whether and how AI use may have impacted their work. Further, research suggests that AI users may misjudge the extent to which they or their AI tool contributed to a work product (Skulmowski, 2024); a written record may mitigate this issue and help research teams gain more clarity about how they are using AI. Additionally, this record will be useful if the team decides to submit to a different venue than originally planned (e.g., due to manuscript rejection): they will need to ensure that their AI use comports with the policies of the new venue.

The record of AI use should include which tool (and version, if available) was used, what it was used for, how the use interacted with the researcher's contribution to the task, and notes evaluating its use (concept credit: Stephanie Brown and Erik Rawls, Florida State University). Figure 4 is an example of an AI use record. In addition to this general record, document with precision how AI was used as part of the record of information for the study's methodology (e.g., what prompts were used, what output was generated). It will often be appropriate to include this information in the methodology section of research papers.

AI Tool	Task	Researcher	Researcher Contribution	Evaluation Notes	Link to Detailed Information
Asta (no version listed)	Finding papers for lit review	Laila	Vetted results by reading abstracts at the publishers' websites; used Google Scholar to find additional papers.	Asta identified several very helpful papers that I didn't find in Google Scholar, but some it marked as "perfectly relevant" were not.	<Link>
Gemini (version 2.5)	Summarizing responses to open-ended questions	Armando	Read raw responses first, noted themes, then read the AI summary, checking for additional themes.	Summaries were not helpful; they were too general and vague.	<Link>

FIGURE 4: Example of an AI Use Record.

Guideline for Documenting AI Use

- [ ] Maintain a written and detailed record of all AI use throughout the research process. Detail when and how it was used, as well as AI tools and versions used.

3.5 Implementation Considerations

If AI is to be implemented, there are several important guidelines to consider. First, ***implement training around responsible AI use,*** particularly data privacy issues. For example, the CITI program offers an "Essentials of Responsible AI" course for researchers (CITI Program, n.d.). Due to the rapidly evolving nature of AI tools, an appropriate cadence for updating the training materials and for researchers to complete 'refresher' instruction should be established. Plans to train new research team members as they onboard should also be established.

It is important to ***ensure that the AI tools selected support the accessibility needs of all research team members***. One study of AI tools used by computer science researchers with differing abilities (e.g., those who are blind or hard-of-hearing) uncovered many barriers to accessibility and accuracy (Glazko et al., 2023). Particularly concerning were requests of AI tools to make materials accessible that resulted in output that promised but did not deliver accessibility. Thus, do not assume that AI tools can reliably generate accessible output when prompted to do so.

What Training about AI Do STEM Education Researchers Need?

Given the novelty of AI tools, there is not yet consensus regarding what researchers need to know in order to use them responsibly. However, Hümmer has articulated a framework for responsible AI use (Hümmer, 2025) that may serve as starting points for training:

- Prompting: learning to effectively decompose their problem into smaller problems in order to craft more effective prompts.
- Pre-determining success criteria: determining the elements of an acceptable answer *before* using an AI tool. For example, what atypical cases does code for performing data analysis need to be able to handle accurately before the code is put into use?
- Validation: comparing the AI output to accepted sources (e.g., books, expert peers) to verify the output.
- Documentation: recording AI use to promote review, auditability, transparency, and (to the extent possible) replicability.

It is also important to ***select AI tools that support responsible use*** whenever possible. These features include transparency (e.g., of their training data, of their model), explainability (i.e., accurate, user-friendly explanations of how the output was generated from the data), ease for human review and modification of AI output, strong data privacy policies, and documentation supporting their security and compliance policies. Be aware that AI tool providers have a financial incentive to use lower-cost approaches, and these approaches may lack the features listed above. These limitations may be masked by other appealing features (e.g., more positive commentary on the researcher's ideas, a nicer interface).

Critical Question

How will I evaluate the quality of an AI tool? For example, how will I be able to tell if an AI tool uses a lower quality LLM?

Where possible, researchers should ***consider performing accuracy and bias auditing*** for their AI tools in their research context. For example, researchers might take a dataset that was previously analyzed by human researchers, repeat the analysis with an AI tool, and compare the two analyses for accuracy and bias. These findings will help inform researchers' future decisions about AI tool use and will likely fill an important gap given the lack of research data on AI tools.

Guidelines for Implementing AI Use

- [ ] Implement training around responsible AI use, with an appropriate cycle of updated learning.
- [ ] Ensure that the AI tools selected support the accessibility needs of all research team members.
- [ ] Select AI tools that support responsible use by offering transparency, explainability, ease of human review and output modification, excellent data privacy policies, and a reputation for security and compliance.
- [ ] Consider performing accuracy and bias auditing.

3.6 Monitoring AI Use

Even software that may not be thought of as being AI-powered requires monitoring. ***Be aware that many traditional software applications (e.g., Google Sheets and Microsoft Excel) have recently added AI functionality.*** Often, this functionality is added to software without advance notice to users. Thus, AI use policies (e.g., data protection) may be needed even if researchers are not using AI features. Since some software enables AI features by default (e.g., AI-generated transcripts or summaries of virtual meetings), ***turn off default AI tools*** – which may, for example, capture discussions about study participants – unless they are necessary. Similarly, ***prepare to respond if AI features are added to software or if the AI model used changes***. AI functionality may be added to tools, sometimes without informing users beforehand or permitting the user to opt-out. For example, Google Forms began automatically showing AI summaries of free responses to form questions before showing the raw responses (Uzondu, 2025). Researchers should consider in advance how they will respond if a tool begins adding AI output. Quick responses during the middle of a research project may be needed, particularly if updates to the tool counter researchers' established AI use policies (i.e., analyzing raw data oneself before using an AI tool to analyze it).

It is also helpful to ***monitor research on AI tools***, particularly those used in STEM-ER or education research more generally. This research area is still emerging, but there is already a growing body of studies exploring best practices, use cases, tool performance, security vulnerabilities, and so forth. Be familiar with the literature on human-AI collaboration, particularly studies focused on researchers, in order to be aware of best practices and concerns with AI use in research tasks. Additionally, monitor research on topics such as the transparency of the models upon which AI tools are built (Bommasani et al., 2023) and the wider impacts of AI use (e.g., environmental impact). Although time-consuming, awareness of sound research will promote the ability to make responsible decisions about AI use.

Update lab/team/organization AI use policies as tools and their uses change. The AI landscape is rapidly evolving, and establishing a cadence for revisiting AI use policies is important; track issues to address (e.g., new uses of an extant tool) in future policy revisions. AI use logs can provide helpful information for updating AI use policies. Some AI tools may be underpriced to develop their user base; plan how to respond if the pricing structure changes during a project.

Guidelines for Monitoring AI and Other Software

- [ ] Be aware that many traditional software applications (e.g., Google Sheets and Microsoft Excel) have recently added AI functionality, and so AI use policies may be needed.
- [ ] Turn off default AI tools (e.g., meeting summaries) unless they are needed.
- [ ] Prepare to respond if AI features are added to software or if the AI model used changes.
- [ ] Monitor research on AI tools.
- [ ] Update lab/team/organization AI use policies as tools – and their uses – change.

Chapter 4.

Guidelines for Specific Phases of the Research Process

In this section, the guidelines are geared to AI use cases in distinct phases of the research process, such as collecting data or disseminating results. More general aspects of research, like ideation and research question development, are covered in Section 3.1.2.

Critical Question

How will I determine if an AI tool we use to identify relevant research studies prioritizes certain research (e.g., scholars, organizations) **methodologies over others?**

4.1 Reviewing Literature

Literature reviews may be conducted in the context of preparing a grant proposal, writing the background section of a research study, or as a systematic or structured literature review. Regardless, if AI tools are used to review literature, certain practices will promote responsible AI use.

Most of the guidelines in this section are predicated on the idea that AI output is imperfect (Ngwenyama & Rowe, 2024). For example, it is important to ***be aware that some AI tools have a cut-off date for their training data***, which will prevent more recent research from appearing in their output (Roe, 2025). Particularly with emerging topics, researchers should ensure that any AI tools used have appropriate coverage. Also, ***be aware of research showing that LLMs tend to overgeneralize*** when summarizing research literature (Peters & Chin-Yee, 2025) – with worse performance for newer AI models and, surprisingly, if the LLM was specifically asked for an accurate summary. As discussed previously, AI tools may highlight only the most popular (rather than the most sound or those that encompass a wide range of participants) research, may have output that includes various biases, or may be inaccurate.

Next, determine whether an AI tool has a mechanism for excluding or removing retracted papers or whether these materials might be present in the training data and therefore influence the output. Larger or general-purpose AI tools are unlikely to (or may be incapable of) removing problematic material from their training data or data set.

There are instances of fictitious citations and similar fabrications in AI output (Butler, 2025; Lemley & Ouellette, 2025; Merken, 2025), so verify from the publisher's website that an authentic paper exists for each AI-provided citation, that each element of the citation is accurate, and that the paper contains the information claimed by the AI. It is important to check the publisher's website and not rely on secondary sources; it is possible that other research aggregators (e.g., university libraries, online databases) may include fictitious, AI-generated citations.

For a systematic literature review, include the prompt used, the AI tool, its version, and the date to promote replication and transparency. (But note that it is often the case that providing the same prompt to an AI tool will not result in the same output.)

Finally, ensure that the use of AI tools does not impede researchers' knowledge development, such as their understanding of the literature. Conducting literature reviews, although time-consuming, is important for developing content knowledge,

Literature Review AI Policy Options

Options for a research group's AI use policy for literature reviews include:

- **All AI use is prohibited.**
- **AI may be used to find relevant papers.**
- **AI may be used to quickly identify relevant details in a paper (e.g., What instrument was used to measure students' astronomy content knowledge?).**
- **AI may be used to identify gaps in the research literature.**
- **AI may be used to summarize papers.**

particularly for new researchers. That understanding forms the core of researchers' ability to vet AI output, so neglecting its development makes future responsible AI use less likely and may even result in impeding the advancement of STEM-ER. It may therefore be preferable to conduct literature reviews without AI tools, even at the cost of potential efficiency gains.

Guidelines for AI Use in Literature Reviews

- [ ] Be aware that some AI tools have a cut-off date for their training data.
- [ ] Be aware of research showing that LLMs tend to overgeneralize.
- [ ] Determine whether an AI tool has a mechanism for excluding or removing retracted papers.
- [ ] Verify from the publisher's website that an authentic paper exists for each AI-provided citation, that each element of the citation is accurate, and that the paper contains the information claimed by the AI.
- [ ] For a systematic literature review, include the prompts, tools, and versions used.
- [ ] Ensure that the use of AI tools does not impede researchers' knowledge development.

4.2 Collecting Data

Responsible AI use is critical in the data collection phase of a research study, including the ***disclosure of planned AI use in informed consent and assent.*** This disclosure is especially important for collecting and analyzing participant data. Both students (Johri et al., 2024) and adults (Brachman et al., 2025) can have misconceptions or lack understanding about AI, which can negate truly *informed* consent. Ensure that the nature of the AI tool and its use is explained so that participants understand it enough to offer genuinely informed consent/assent. Informed consent/assent includes, but is not limited to, many aspects related to AI:

Critical Question

How will I ensure that participant data shared with an AI tool is adequately protected?

- Defining AI in terms understandable to the participants or participants' parent/guardian
- Explaining what AI will be used, either by naming the tool or by describing it (e.g., whether the data input will be used to train future models)
- Explaining how AI will be used (e.g., to identify themes in interview transcripts)
- Explaining how the researchers and the AI will interact (e.g., the researchers will identify themes, then have the AI identify themes, and then compare the results)
- Specifying what participant data will be used with AI
- Clarifying if and how personally identifiable information will be shared with AI
- Stating the potential risks of AI use (e.g., data re-identification in the future, societal risks)
- Stating the potential benefits of AI use (e.g., allowing more research participants and therefore including more viewpoints and experiences)

Understand the data use policy of any AI tools used, particularly if the tool will interact with participants' data. Important questions to ask include:

- Will the data be used for training future AI models?
- How likely is the data to be re-identified in the future?
- What is the tool's privacy policy?
- Does the tool align with principles for responsible data use (e.g., the CARE Principles)?

Do not input personally identifiable information into an AI tool if there is a risk of data exposure. ***Be aware that the expanding capabilities of AI tools mean that what constitutes 'personally identifiable information' is changing.*** For example, one study found that LLMs can often infer personal information that is not overtly present in a text, even when steps are taken to remove information that, although not identifying, can be used to make inferences (Staab et al., 2023).

Although researchers might choose to use AI generated-data (i.e., synthetic data) in specific use cases (e.g., to test instruments), researchers should ***avoid using synthetic data as the sole data source***. Relying entirely on synthetic data in a research study risks promulgating results that have not been vetted against ground truth (i.e., non-synthetic data). Synthetic data might be biased, incomplete, inaccurate, or lack sufficient variety (Birhane, 2025; S. Hao et al., 2024). Kieser et al. examined the use of synthetic data in physics education research and found that it did not always accurately mirror student performance or the conditions of the prompt used (Kieser et al., 2023).

It is also important to carefully track the sources of data in any situation where synthetic data (e.g., from a pilot test of a new instrument) is mixed with data from human study participants. If synthetic data is used, it may be helpful to compare results based on synthetic data to results based on human data (e.g., performance in an educational game) to gain a better understanding of how well the synthetic data mirrors human data.

Be aware that some individuals use AI tools to complete online forms and surveys (Rilla et al., 2025; Westwood, 2025); develop plans to ensure that data is collected from humans.

Data Collection AI Policy Options

Options for a research group's AI use policy for data collection include:

- **All AI use is prohibited.**
- **AI may be used to draft data collection instruments (e.g., surveys).**
- **AI may be used to create supplemental materials for data collection (e.g., images that accompany a survey).**
- **AI may be used to create digital tools used for data collection (e.g., a webpage or a form).**
- **AI may be used to pilot test data collection instruments by generating synthetic data.**
- **AI may be used to gather feedback on data collection instruments.**
- **AI may be used for transcription (e.g., of audio recordings of interviews).**
- **AI may be used for translation (i.e., when a study participant uses a language not spoken by a research team member).**

Guidelines for AI Use for Data Collection

- [ ] Disclose planned AI use in informed consent and assent.
- [ ] Understand the data use policy of any AI tools used.
- [ ] Be aware that the expanding capabilities of AI tools mean that what constitutes 'personally identifiable information' is changing.
- [ ] Avoid using synthetic data as the sole data source.
- [ ] Be aware that some individuals use AI tools to complete online forms and surveys.

4.3 Analyzing Data

Responsible research – with or without AI use – requires protecting sensitive data. When AI tools are used, it is essential to ***de-identify data as completely as possible before sharing it with an AI tool***. Be aware of the potential for future data re-identification as technologies improve, and remove all possible identifiers – not just the data traditionally considered to be personally identifiable information (e.g., names). AI systems often excel at data triangulation, or the ability to merge disparate datasets (i.e., the Mosaic Effect). For example, unspecific age indicators in an interview transcript (e.g., “I used to watch *Friends* when I was in high school”) might be combined with other information to identify a research participant in the future (Staab et al., 2023). Whenever this information is not germane to the research study, remove it before analysis.

Critical Question

If I use AI for data analysis, how will I ensure that my own data analysis skills do not atrophy?

Even when not using AI, it is helpful to ***avoid practices that may contribute to data re-identification***. For example, a photo of student participants in a classroom posted to social media or included in a research publication – even with students’ faces blurred – may enable precise identification (e.g., school name) based on the background of the picture. And, combined with other datasets with additional information, this identification may make it possible to identify participants by name. Whenever possible, avoid situations where datasets might be combined by an AI tool’s owner, which would increase the odds of data re-identification.

Responsible data analysis with AI requires that researchers avoid practices that can encode bias in their results. ***Consider differences between the population/setting of the AI's training data and the study participants*** – if there are important differences, then AI data analysis may lead to inaccurate output. Also, ***be aware of evidence that AI tools can show bias***, including racial bias. For example, based on dialect indicators alone (i.e., absent other indicators of race), there is evidence of AI stereotypes of speakers of African American English that "are more negative than any human stereotypes about African Americans ever experimentally recorded" (Hofmann et al., 2024, p. 147). Ensure that any AI-based analysis is free from such bias.

Finally, ***ensure that statistical analysis with AI follows sound education research practices***. The ease of performing data analyses with AI may increase the risk of *p*-hacking, or statistical practices designed to increase the chances of identifying significant research results where none may exist (Stefan & Schönbrodt, 2023). Pre-registration may be a helpful mitigation strategy. Note that it is inadvisable to use LLM-based tools for any statistical analysis due to their model architecture; it may be more appropriate to use other forms of AI (e.g., machine learning algorithms such as random forest or support vector machines) for some quantitative data analysis.

Data Analysis AI Policy Options

- **All AI use is prohibited.**
- **A local AI tool (i.e., that does not transmit data to any other entity, tool, or location) may be used to analyze data.**
- **An external AI tool may be used.**
- **AI may be used to gather insights into data (e.g., identify patterns).**
- **AI may be used to visualize data (e.g., create charts).**

Vet results generated from code written by AI. If using AI to generate code to conduct statistical analyses (e.g., in R or Stata), ensure the results produced make sense given your data (e.g., no correlations are above 1). It is also advisable to double-check any data manipulations performed by AI-generated code, such as recoding values, transforming datasets from long to short, or handling duplicate responses.

Guidelines for AI Use in Data Analysis

- [ ] De-identify data as completely as possible before sharing it with an AI tool.
- [ ] Avoid practices that may contribute to data re-identification.
- [ ] Consider differences between the population/setting of the AI's training data and the study participants.
- [ ] Be aware of evidence that AI tools can show bias.
- [ ] Ensure that statistical analysis with AI follows sound education research practices.
- [ ] Vet results generated from code written by AI.

4.4 Reporting and Sharing Results

4.4.1 Writing and Editing

Before AI is used for any writing or editing tasks, researchers should ***be aware that some research shows negative outcomes related to AI use for writing tasks***. For example, using an AI tool to assist with writing can influence the writer's opinions (Jakesch et al., 2023). Plus, productivity gains varied based on country of origin, and writers tended to adopt Western writing norms (D. Agarwal et al., 2025). Although users completing tasks generated more ideas (and more elaborate ideas) when using AI, the ideas were less diverse, and the user felt less responsibility for them (Anderson et al., 2024).

Critical Question

How will I preserve human creativity and autonomy while using AI tools for writing or editing?

If AI tools are used, ***record and store all AI-related work***, including human-authored drafts, AI prompts, AI output, and how the draft was changed based on the AI output. And, ***attribute AI-generated content appropriately*** when writing research papers, which may include describing how the AI was used (e.g., outlining, drafting, editing). Some style guides, including APA, have developed guidelines for citing generative AI (McAdoo et al., 2025).

Citing Generative AI

APA Style offered the following template for citing AI output (McAdoo et al., 2025):

Reference List
AI Company Name. (year, month day). Title of chat in italics [Description, such as Generative AI chat]. Tool Name/Model. URL of the chat

Parenthetical Citation
(AI Company Name, year)

Narrative Citation
AI Company Name (year)

Writing and Editing AI Policy Options

- All AI use is prohibited.
- AI may be used for grammar and spell checking.
- AI may be used to generate paragraphs from a bullet list of items generated by the researcher.
- AI may be used to get feedback on a draft.
- AI may be used to identify discrepancies between a draft and a publication venue's requirements (e.g., formatting issues).

Finally, ***do not credit an idea to AI unless one has verified that there is no other (human) source for the idea*** (Lemley & Ouellette, 2025). See Section 3.1.2 for more information.

Guidelines for AI Use for Writing and Editing

- ☐ Be aware that some research shows negative outcomes related to AI use for writing tasks.
- ☐ Record and store all AI-related work.
- ☐ Attribute AI-generated content appropriately.
- ☐ Do not credit an idea to AI unless one has verified that there is no other (human) source for the idea.

4.4.2 Peer Reviewing

AI use presents several challenges to the peer review process. Navigating them responsibly requires researchers who are conducting a peer review to first ***check the conference or journal's AI use policies to ensure that AI use comports with their policy***. Because papers that are peer reviewed are the intellectual property of the authors, ***do not upload others' work to an AI system*** without prior approval. It is good practice to ***disclose whether, how, and what AI was used in the peer review process*** in the review itself. Finally, ***if you suspect that an article that you are reviewing was written by AI, do not assume that your suspicion is accurate***. Some research shows that human peer reviewers cannot themselves distinguish AI-generated from human written text (Hadan et al., 2024), and AI detectors are not completely accurate, especially for users whose first language is not English (Erol et al., 2025; Liang et al., 2023).

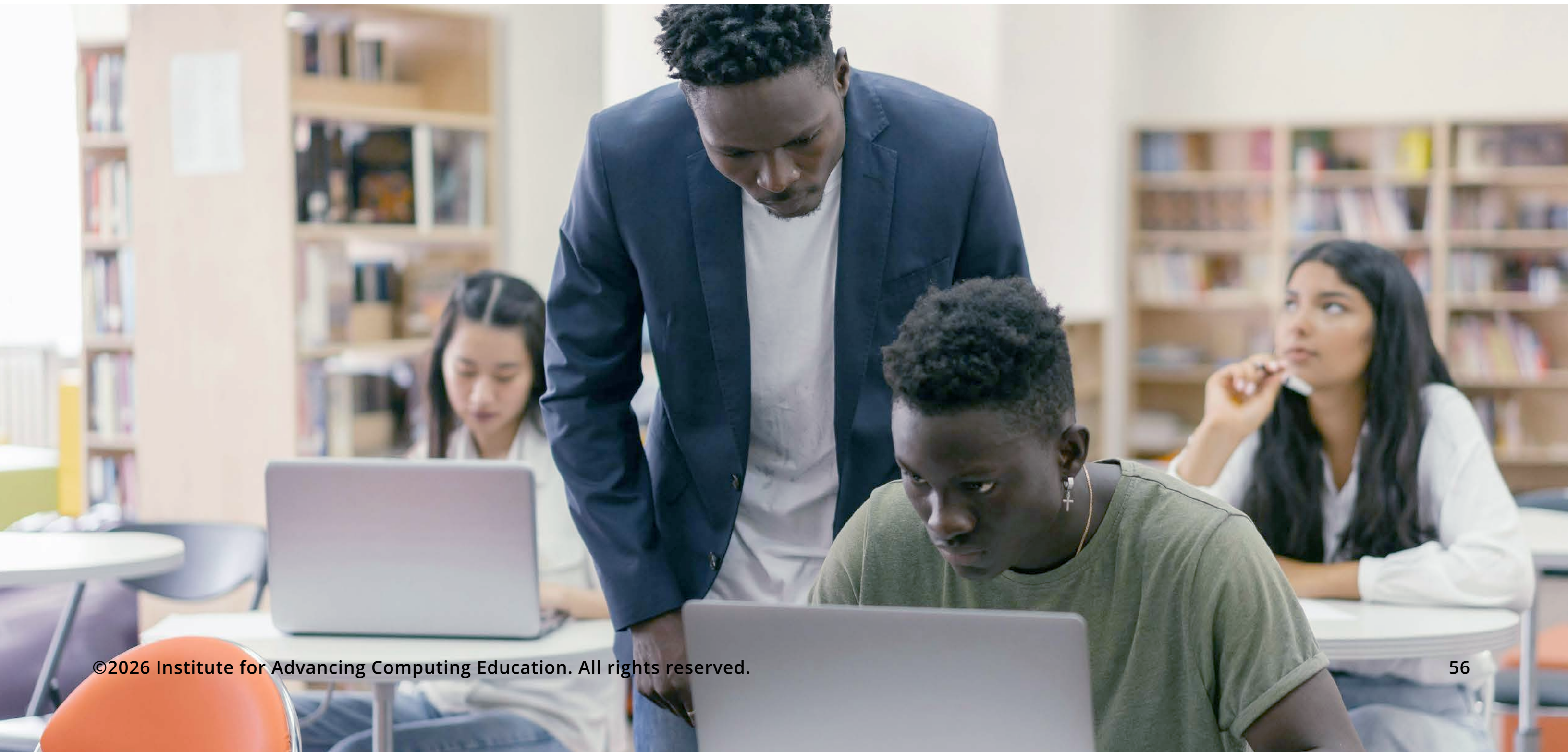
> Critical Question
>
> **If I am not comfortable disclosing my AI use, is that an indication that the use is not aligned with my values and goals?**

Peer Review AI Policy Options

- **All AI use is prohibited.**
- **AI may be used for grammar and spell checking.**
- **AI may be used to generate paragraphs from a bullet list of items generated by the reviewer.**
- **After the reviewer has written their review, they may use an AI tool to generate additional items to consider including in the review.**
- **After the reviewer has written their review, they may use an AI tool to get feedback on the review's clarity.**

Guidelines for AI Use in the Peer Review Process

- [ ] Check the conference or journal's AI use policies to ensure that AI use comports with their policy.
- [ ] Do not upload others' work to an AI system.
- [ ] Disclose whether, how, and what AI was used in the peer review process.
- [ ] If you suspect that an article that you are reviewing was written by AI, do not assume that your suspicion is accurate.



4.4.3 Sharing Research

Using AI in the research dissemination process presents both challenges and opportunities to researchers. To use AI responsibly, first ***understand the AI policies of potential publication venues*** (e.g., journals) to ensure that their AI policy aligns with the preferences of the researchers. These policies include both (1) whether and how the venue permits authors to use AI (e.g., to perform data analysis, draft paper sections, polish their writing) and (2) whether and how the venue will use AI on authors' submissions (e.g, Will submissions or publications be used to train new AI models? If not, is there assurance that that policy won't change in the future?). Although venues for informal dissemination (e.g., for blog posts or infographics) are unlikely to have AI policies, be aware that some readers will have a negative impression of AI use (Zhang & Gosline, 2023) due to a preference for human-generated content.

Next, ***disclose AI use in presentations and publications***. Consider using a modified version of the CRediT author statement (Allen et al., 2019) to disclose AI use in publications (see Figure 5). The modified version of the statement includes tasks performed by AI and, for each such task, a researcher would have a corresponding task of supervising the AI. (Note that supervision implies that the researcher has the expertise to evaluate the AI output.) Or, describe AI use in a separate statement (see Figure 6). Where fully describing AI use might constitute a too-lengthy addition to a paper, it may be appropriate to put a brief disclosure in the paper and the full disclosure in supplementary materials, or to make it available upon request.

Traditional CRediT author statement
Jonathan Curtis: funding acquisition, conceptualization, supervision. **Zainab Jackson:** data collection, formal analysis, data curation. **Gemma Morales:** writing, editing.

Modified author statement
Jonathan Curtis: funding acquisition, conceptualization, supervision. **Zainab Jackson:** data collection, formal analysis, <u>supervision of formal analysis</u>, data curation. **Gemma Morales:** writing, <u>supervision of editing</u>. <u>**AI tools:** formal analysis, editing</u>.

FIGURE 5: Traditional CRediT author statement and a modified version to include AI use (with additions underlined for contrast).

"Use of AI assistance: An AI assistant (ChatGPT/GPT-5 Thinking) was used for language editing, outline refinement, and stylistic suggestions. All conceptual content, framework design, interpretations, and final decisions are the author's own. The AI system did not have authorship or decision-making roles and is not listed as an author."

FIGURE 6: Use of AI statement (from Hümmer, 2025).

Open science practices have called for researchers to share datasets publicly. However, re-identification of data has become more powerful as AI becomes more sophisticated. To avoid the potential for data re-identification, instead of sharing datasets publicly, ***adopt practices to keep abreast of data re-identification advances in order to protect participant data.*** One such way is to have interested researchers ask for data (if this sharing has been approved by your IRB). Keep in mind that it is necessary to understand how data triangulation works, the latest progress that has been made in re-identification processes, and how these combined may change how you clean and de-identify data before sharing.

Guidelines for AI Use in Sharing Research

- [ ] Understand the AI policies of potential publication venues.
- [ ] Disclose AI use in presentations and publications.
- [ ] Adopt practices to keep abreast of data re-identification advances in order to protect participant data (e.g., requiring data requests instead of sharing datasets publicly).

Critical Question

How will I determine whether AI-generated ideas for dissemination have evidence of bias or unfairness (e.g., only suggest sharing results with certain types of community groups)?

Sharing Research AI Policy Options

- All AI use is prohibited.
- AI may be used to provide feedback on academic writing.
- AI may be used for spelling/grammar correction.
- AI may be used to format materials (e.g., according to a journal's guidelines).
- AI may be used to create slide decks for academic presentations.
- AI may be used to design or format slide decks for academic presentations.
- AI may be used to suggest venues for popular dissemination (e.g., community groups).
- AI may be used to create materials (e.g., infographics, podcasts) for popular dissemination.

Chapter 5.

Recommendations for Adjacent Communities

These guidelines are focused on researchers who are conducting STEM-ER, but some of the concerns related to AI use can be mitigated by other parties. This section offers recommendations for other groups who might enable responsible AI use.

5.1 Professional Organizations

Professional organizations like the Association for Computing Machinery (ACM or Institute of Electrical and Electronics Engineers (IEEE) have large platforms for promoting responsible AI use among their members.

- Articulate guidelines for responsible AI use. Doing so will reduce the need for each researcher or research team to develop their own, a process that requires time and expertise beyond what they may have.

- Promote professional learning about responsible AI use.

5.2 Institutions

Postsecondary institutions, research institutions, and other organizations that conduct STEM-ER all bear responsibility in protecting participant data.

- Articulate AI use policies, especially where institution-wide consistency will be a benefit.

- Promote policies that permit a choice of whether to use AI. For example, if a project’s AI use plans are included when recruiting undergraduate researchers, they will be able to select a project that aligns with their preferences.

- Avoid policies and practices that directly or indirectly penalize researchers who choose to use (or not use) AI. These policies may discourage researchers from disclosing if and how they use AI.

5.3 AI Model and Other Software Developers

Developers of AI models and other software developers have a responsibility to education researchers in protecting participants and in supporting education research.

- Provide advance notice to users when AI is to be added to a tool, and provide users with an easy way to opt out of its use. When AI output appears automatically, it removes the autonomy of users to decide when and how they will use AI output.

- Develop AI tools optimized for education researchers. These needs include non-negotiables, such as protecting participants' data, as well as features that would assist in research workflows.

- Design tools that provide transparency and replicability of research. Since transparency is a central value of the research process, AI tools for researchers should be as transparent as possible (e.g., by describing their training data). Repeating the same prompt may not lead to the same results which undermines the ability to replicate research.

- Create models and tools that handle retracted papers appropriately (i.e., remove them from training data).

- Develop tools that explicitly counter the tendency to provide the statistically most likely result. An AI tool designed to assist with literature reviews might instead highlight research written in languages other than English (dependent on the researcher's needs) or involving participants who are different from the majority of research subjects (e.g., students who attend community college instead of large research institutions).

- Develop AI tools that foreground responsible use. For example, some AI training data is scraped from the Internet in violation of websites' terms of service, despite research showing that compliance does not degrade model performance (Fan et al., 2025) if followed by specialized training.

- Develop tools for open science platforms to ensure that the addition of a new dataset or technological improvements do not enable data re-identification.

5.4 Researchers in Fields Supporting STEM-ER

Gaining a better understanding of the affordances and limitations of AI use in STEM-ER will require collaboration among researchers of various specialties, including AI, education, and human-computer interaction, among others.

- Create benchmarks for STEM-ER-related tasks. For example, it would be helpful if researchers could compare the performance of various AI tools on the task of developing a list of knowledge components based on a set of learning standards. Benchmarks should include issues related to bias.

- Consider using alternative approaches to AI besides the dominant paradigms and products, as well as non-AI approaches to researchers' needs (e.g., data analysis tools that do not use AI). Alternatives may include an approach to AI based on Indigenous knowledge (Lewis et al., 2025); non-LLM-based AI, including neuro-symbolic AI, embodied AI, and quantum AI (Jensen et al., 2025); or interpretable AI, meaning that their output can be traced to their input (Guide Labs, 2026/2026).

- Assess AI models and tools for their performance on STEM-ER tasks, as there is very little empirical data in this area. For example, researchers might use AI tools to generate synthetic data for quickly pilot-testing new evaluation instruments. More research is needed on whether this process is effective (e.g., With AI pilot testing, do instruments perform better with research subjects? Do researchers save time and other resources? Are there any unanticipated outcomes?).

- Develop a checklist for AI tool creators to enable easy evaluation of new AI tools. This is currently a challenge given the rapid evolution of these tools. For example, items on the checklist might include, "provide evidence that a malicious user cannot access the data entered into the tool by another user" or "provide evidence that the tool's output does not preference students from different socioeconomic backgrounds."

- Develop user-friendly tools that assist in determining the environmental impact of using various AI tools to enable researchers to gauge the potential impact of their choices. For example, knowing the resource cost per query would assist researchers in making informed decisions about whether and how to use a tool.

5.5 Funders

Funding organizations are often in a position to drive responsible AI use in ways that other organizations are unable to.

- Support research into the use of AI tools by STEM education researchers that focuses on issues related to responsible use and best practices for use.

- Because AI functionality is embedded in so much software, it is often impossible to avoid interacting with AI, even when a research team determines that this is not optimal. Funding the development of AI-free ("off the grid") software would enable researchers to have a choice about whether and how to use tools that include AI.

- Provide funding for AI tools that better protect privacy, have evidence of avoiding bias, and have other indicators of responsible development.

- Avoid policies that inadvertently pressure researchers to use AI. For example, lengthy reporting requirements may nudge resource-limited researchers into using AI to write reports.

Chapter 6.

Conclusion

One of the greatest challenges in seeking to use AI responsibly is the lack of information: When and how is deskilling likely to occur, and what are the best ways to prevent it? What is the actual environmental impact of a given AI tool? STEM education researchers simply do not have the necessary information to answer these and similar questions with certainty, and that makes the complicated calculus of weighing the various personal, professional, societal, and planetary impacts of AI all the more challenging. On the other hand, rapidly advancing tools and research can also lead to information overload (e.g., about AI's environmental impact) with rapidly-changing findings that can be difficult to digest and to monitor.

And yet decisions about AI use must still be made. Our team acknowledges that the research presented in this report is preliminary, and we expect future studies to modify the findings shared here. These guidelines are hopefully a helpful starting point for thinking about the issues related to responsible AI use in STEM-ER. Our team encourages the STEM-ER community to promote the use of responsible AI by others whenever and however possible; the actions of researchers will determine norms around AI use.

Finally, our team encourages researchers to harness the capabilities of AI to further STEM-ER and to do so in ways that align with the field's goals and values while centering humans, protecting participants, and sustaining ethical and sound research practices.

References

Agarwal, C., Tanneru, S. H., & Lakkaraju, H. (2024). *Faithfulness vs. plausibility: On the (un)reliability of explanations from large language models* (arXiv:2402.04614; Version 3). arXiv. https://doi.org/10.48550/arXiv.2402.04614

Agarwal, D., Naaman, M., & Vashistha, A. (2025). AI suggestions homogenize writing toward Western styles and diminish cultural nuances. Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems, 1–21. https://doi.org/10.1145/3706598.3713564

Allen, L., O'Connell, A., & Kiermer, V. (2019). How can we ensure visibility and diversity in research contributions? How the Contributor Role Taxonomy (CRediT) is helping the shift from authorship to contributorship. *Learned Publishing, 32*(1), 71–74. https://doi.org/10.1002/leap.1210

Amarikwa, M. (2023). Internet openness at risk: Generative AI's impact on data scraping. *Rich. JL & Tech*., 30, 533.

American Institutes for Research. (2026). *Principles for the Use of Artificial Intelligence in Education Research: Draft for Public Comment*. American Institutes for Research. https://www.air.org/sites/default/files/2026-02/AI-Principles-for-Education-Research-Draft-for-Public-Comment-Feb-2026.pdf

Anderson, B. R., Shah, J. H., & Kreminski, M. (2024). Homogenization effects of large language models on human creative ideation. *Proceedings of the 16th Conference on Creativity & Cognition*, 413–425. https://doi.org/10.1145/3635636.3656204

Antunes, B., & Hill, D. R. C. (2024). Reproducibility, replicability and repeatability: A survey of reproducible research with a focus on high performance computing. *Computer Science Review, 53*, 100655. https://doi.org/10.1016/j.cosrev.2024.100655

Atari, M., Haidt, J., Graham, J., Koleva, S., Stevens, S. T., & Dehghani, M. (2023). Morality beyond the WEIRD: How the nomological network of morality varies across cultures. *Journal of Personality and Social Psychology, 125*(5), 1157–1188. https://doi.org/10.1037/pspp0000470

Baek, C., & Doleck, T. (2024). Learning analytics: A comparison of Western, educated, industrialized, rich, and democratic (WEIRD) and non-WEIRD research. *Knowledge Management & E-Learning, 16*(2), 217–236.

Banerjee, S., Agarwal, A., & Singla, S. (2025). LLMs will always hallucinate, and we need to live with this. In K. Arai (Ed.), I*ntelligent Systems and Applications* (pp. 624–648). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-99965-9_39

Barile, J., Margolis, A., Cason, G., Kim, R., Kalash, S., Tchaconas, A., & Milanaik, R. (2024). Diagnostic accuracy of a large language model in pediatric case studies. *JAMA Pediatrics*. https://doi.org/10.1001/jamapediatrics.2023.5750

Barnes, T., Danish, J., Finkelstein, S., Molvig, O., Burriss, S., Humburg, M., Reichert, H., & Limke, A. (2024). *Toward Ethical and Just AI in Education Research*. Community for Advancing Discovery Research in Education (CADRE). Education Development Center, Inc. https://cadrek12.org/sites/default/files/2024-06/CADRE-Brief-AI-Ethics-2024.pdf

Becker, J., Rush, N., Barnes, E., & Rein, D. (2025). *Measuring the Impact of Early-2025 AI on Experienced Open-Source Developer Productivity* (arXiv:2507.09089). arXiv. https://doi.org/10.48550/arXiv.2507.09089

Berthelot, A., Caron, E., Jay, M., & Lefèvre, L. (2024). Estimating the environmental impact of Generative-AI services using an LCA-based methodology. *Procedia CIRP, 31st CIRP Conference on Life Cycle Engineering, 122*, 707–712. https://doi.org/10.1016/j.procir.2024.01.098

Birhane, A. (2025). Cheap science, real harm: The cost of replacing human participation with synthetic data. *Synthetic Data Workshop at the Aarhus 2025 Conference*. https://synthetic-data-workshop.github.io/papers/13.pdf

References

Blasi, D. E., Henrich, J., Adamou, E., Kemmerer, D., & Majid, A. (2022). Over-reliance on English hinders cognitive science. *Trends in Cognitive Sciences, 26*(12), 1153–1170. https://doi.org/10.1016/j.tics.2022.09.015

Bolaños, F., Salatino, A., Osborne, F., & Motta, E. (2024). Artificial intelligence for literature reviews: Opportunities and challenges. *Artificial Intelligence Review, 57*. https://doi.org/10.1007/s10462-024-10902-3

Bommasani, R., Klyman, K., Longpre, S., Kapoor, S., Maslej, N., Xiong, B., Zhang, D., & Liang, P. (2023). *The Foundation Model Transparency Index* (arXiv:2310.12941). arXiv. https://doi.org/10.48550/arXiv.2310.12941

Bontcheva, K., Papadopoulous, S., Tsalakanidou, F., Gallotti, R., Dutkiewicz, L., Krack, N., Teyssou, D., Severio Nucci, F., Spangenberg, J., & Srba, I. (2024). *Generative AI and disinformation: Recent advances, challenges, and opportunities.* https://lirias.kuleuven.be/retrieve/758830

Borenstein, J., & Howard, A. (2021). Emerging challenges in AI and the need for *AI ethics education. AI and Ethics, 1*(1), 61–65. https://doi.org/10.1007/s43681-020-00002-7

Borji, A. (2023). *A categorical archive of ChatGPT failures* (arXiv:2302.03494). arXiv. https://doi.org/10.48550/arXiv.2302.03494

Bos, J. (2020). *Research Ethics for Students in the Social Sciences*. Springer Nature. https://doi.org/10.1007/978-3-030-48415-6

Brachman, M., Kunde, S., Miller, S., Fucs, A., Dempsey, S., Jabbour, J., & Geyer, W. (2025). Building appropriate mental models: What users know and want to know about an agentic AI chatbot. *Proceedings of the 30th International Conference on Intelligent User Interfaces, IUI '25*, 247–264. https://doi.org/10.1145/3708359.3712071

Brittain, B. (2025a, September 25). US judge preliminarily approves $1.5 billion Anthropic copyright settlement. *Reuters.* https://www.reuters.com/sustainability/boards-policy-regulation/us-judge-approves-15-billion-anthropic-copyright-settlement-with-authors-2025-09-25/

Brittain, B. (2025b, November 12). OpenAI fights order to turn over millions of ChatGPT conversations |*Reuters*. Reuters. https://www.reuters.com/business/media-telecom/openai-fights-order-turn-over-millions-chatgpt-conversations-2025-11-12/

Budzyń, K., Romańczyk, M., Kitala, D., Kołodziej, P., Bugajski, M., Adami, H. O., Blom, J., Buszkiewicz, M., Halvorsen, N., Hassan, C., Romańczyk, T., Holme, Ø., Jarus, K., Fielding, S., Kunar, M., Pellise, M., Pilonis, N., Kamiński, M. F., Kalager, M., ... Mori, Y. (2025). Endoscopist deskilling risk after exposure to artificial intelligence in colonoscopy: A multicentre, observational study. *The Lancet Gastroenterology & Hepatology, 10*(10), 896–903. https://doi.org/10.1016/S2468-1253(25)00133-5

Butler, P. (2025, September 12). N.L.'s 10-year education action plan cites sources that don't exist. *CBC News.* https://www.cbc.ca/news/canada/newfoundland-labrador/education-accord-nl-sources-dont-exist-1.7631364

Carroll, S. R., Garba, I., Figueroa-Rodríguez, O. L., Holbrook, J., Lovett, R., Materechera, S., Parsons, M., Raseroka, K., Rodriguez-Lonebear, D., Rowe, R., Sara, R., Walker, J. D., Anderson, J., & Hudson, M. (2020). The CARE principles for indigenous data governance. *Data Science Journal, 19*. https://doi.org/10.5334/dsj-2020-043

Chen, X., Wang, X., Colacelli, A., Lee, M., & Xie, L. (2025). *Electricity Demand and Grid Impacts of AI Data Centers: Challenges and Prospects* (arXiv:2509.07218). arXiv. https://doi.org/10.48550/arXiv.2509.07218

CITI Program. (n.d.). *Essentials of Responsible AI*. https://About.Citiprogram.Org/. Retrieved March 17, 2026, from https://about.citiprogram.org/course/essentials-of-responsible-ai/

Cooper, A. F., Gokaslan, A., Ahmed, A., Cyphert, A. B., Sa, C. D., Lemley, M. A., Ho, D. E., & Liang, P. (2025). *Extracting memorized pieces of (copyrighted) books from open-weight language models* (arXiv:2505.12546). arXiv. https://doi.org/10.48550/arXiv.2505.12546

References

Costa, A. L., & Kallick, B. (1993). Through the lens of a critical friend. *Educational Leadership*, 51, 49–49.

Datta, S., Nahin, S. K., Chhabra, A., & Mohapatra, P. (2025). *Agentic AI Security: Threats, Defenses, Evaluation, and Open Challenges* (arXiv:2510.23883). arXiv. https://doi.org/10.48550/arXiv.2510.23883

Dell'Acqua, F., Ayoubi, C., Lifshitz-Assaf, H., Sadun, R., Mollick, E. R., Mollick, L., Han, Y., Goldman, J., Nair, H., Taub, S., & Lakhani, K. R. (2025). *The cybernetic teammate: A field experiment on generative AI reshaping teamwork and expertise* (SSRN Scholarly Paper No. 5188231). Social Science Research Network. https://doi.org/10.2139/ssrn.5188231

Ehsan, U., Passi, S., Saha, K., McNutt, T., Riedl, M. O., & Alcorn, S. (2026). *From Future of Work to Future of Workers: Addressing Asymptomatic AI Harms for Dignified Human-AI Interaction*. https://doi.org/10.1145/3772318.3791081

Erol, G., Ergen, A., Gülşen Erol, B., Kaya Ergen, Ş., Bora, T. S., Çölgeçen, A. D., Araz, B., Şahin, C., Bostancı, G., Kılıç, İ., Macit, Z. B., Sevgi, U. T., & Güngör, A. (2025). Can we trust academic AI detective? Accuracy and limitations of AI-output detectors. *Acta Neurochirurgica, 167*(1), 214. https://doi.org/10.1007/s00701-025-06622-4

Fan, D., Sabolčec, V., Ansaripour, M., Tarun, A. K., Jaggi, M., Bosselut, A., & Schlag, I. (2025). *Can Performant LLMs Be Ethical? Quantifying the Impact of Web Crawling Opt-Outs* (arXiv:2504.06219). arXiv. https://doi.org/10.48550/arXiv.2504.06219

Fanous, A., Goldberg, J., Agarwal, A., Lin, J., Zhou, A., Xu, S., Bikia, V., Daneshjou, R., & Koyejo, S. (2025). Syceval: Evaluating LLM sycophancy. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, 8*(1), 893–900. https://ojs.aaai.org/index.php/AIES/article/view/36598

Fernandez, J., Na, C., Tiwari, V., Bisk, Y., Luccioni, S., & Strubell, E. (2025). *Energy Considerations of Large Language Model Inference and Efficiency Optimizations* (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2504.17674

Feuerriegel, S., Hartmann, J., Janiesch, C., & Zschech, P. (2024). Generative AI. *Business & Information Systems Engineering, 66*(1), 111–126. https://doi.org/10.1007/s12599-023-00834-7

Fox, A. (2022). Educational research and AIED: Identifying ethical challenges. In *The Ethics of Artificial Intelligence in Education* (pp. 47–73). Routledge. https://www.taylorfrancis.com/chapters/edit/10.4324/9780429329067-4/educational-research-aied-alison-fox

Gadiraju, V., Kane, S., Dev, S., Taylor, A., Wang, D., Denton, R., & Brewer, R. (2023). "I wouldn't say offensive but...": Disability-centered perspectives on large language models. *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency, FAccT '23*, 205–216. https://doi.org/10.1145/3593013.3593989

Gerlich, M. (2025). AI tools in society: Impacts on cognitive offloading and the future of critical thinking. *Societies, 15*(1), Article 1. https://doi.org/10.3390/soc15010006

Glazko, K. S., Yamagami, M., Desai, A., Mack, K. A., Potluri, V., Xu, X., & Mankoff, J. (2023). An autoethnographic case study of generative artificial intelligence's utility for accessibility. *Proceedings of the 25th International ACM SIGACCESS Conference on Computers and Accessibility, ASSETS '23*, 1–8. https://doi.org/10.1145/3597638.3614548

Glickman, M., & Sharot, T. (2025). How human–AI feedback loops alter human perceptual, emotional and social judgements. *Nature Human Behaviour, 9*(2), 345–359. https://doi.org/10.1038/s41562-024-02077-2

Grynbaum, M. M., & Mac, R. (2023, December 27). The Times sues OpenAI and Microsoft over A.I. use of copyrighted work. *The New York Times*. https://www.nytimes.com/2023/12/27/business/media/new-york-times-open-ai-microsoft-lawsuit.html

References

Guest, O., Suarez, M., Müller, B., van Meerkerk, E., Oude Groote Beverborg, A., de Haan, R., Reyes Elizondo, A., Blokpoel, M., Scharfenberg, N., Kleinherenbrink, A., Camerino, I., Woensdregt, M., Monett, D., Brown, J., Avraamidou, L., Alenda Demoutiez, J., Hermans, F., & van Rooij, I. (2025). *Against the uncritical adoption of "AI" technologies in academia.* Zenodo. https://doi.org/10.5281/zenodo.17065099

Guest, O., Suarez, M., & van Rooij, I. (2025). *Towards critical artificial intelligence literacies.* Zenodo. https://doi.org/10.5281/zenodo.17786243

Guide Labs. (2026). *Guidelabs/steerling* [Python]. https://github.com/guidelabs/steerling (Original work published 2026)

Hadan, H., Wang, D. M., Mogavi, R. H., Tu, J., Zhang-Kennedy, L., & Nacke, L. E. (2024). The great AI witch hunt: Reviewers' perception and (mis)conception of generative AI in research writing. *Computers in Human Behavior: Artificial Humans, 2*(2), 100095. https://doi.org/10.1016/j.chbah.2024.100095

Hao, Q., Xu, F., Li, Y., & Evans, J. (2025). *Artificial intelligence tools expand scientists' impact but contract science's focus* (arXiv:2412.07727). arXiv. https://doi.org/10.48550/arXiv.2412.07727

Hao, S., Han, W., Jiang, T., Li, Y., Wu, H., Zhong, C., Zhou, Z., & Tang, H. (2024). *Synthetic data in AI: Challenges, applications, and ethical implications* (arXiv:2401.01629). arXiv. https://doi.org/10.48550/arXiv.2401.01629

Hashimoto, D. A., Varas, J., & Schwartz, T. A. (2024). Practical guide to machine learning and artificial intelligence in surgical education research. *JAMA Surgery, 159*(4), 455–456. https://doi.org/10.1001/jamasurg.2023.6687

Hicks, M. T., Humphries, J., & Slater, J. (2024). ChatGPT is bullshit. *Ethics and Information Technology, 26*(2), 38. https://doi.org/10.1007/s10676-024-09775-5

Hofmann, V., Kalluri, P. R., Jurafsky, D., & King, S. (2024). AI generates covertly racist decisions about people based on their dialect. *Nature, 633*(8028), 147–154. https://doi.org/10.1038/s41586-024-07856-5

Hu, K. (2023, February 2). *ChatGPT sets record for fastest-growing user base—Analyst note* | Reuters. https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/

Huang, C., Zhang, Z., Mao, B., & Yao, X. (2022). An overview of artificial intelligence ethics. *IEEE Transactions on Artificial Intelligence, 4*(4), 799–819.

Humlum, A., & Vestergaard, E. (2025). *Large language models, small labor market effects* (Working Paper No. 33777). National Bureau of Economic Research. https://doi.org/10.3386/w33777

Hümmer, M. (2025). *On the influence of artificial intelligence on human problem-solving: Empirical insights for wave 1 from a multinational longitudinal pilot study* (75h3t_v1). SocArXiv. https://doi.org/10.31235/osf.io/75h3t_v1

Jakesch, M., Bhat, A., Buschek, D., Zalmanson, L., & Naaman, M. (2023). Co-writing with opinionated language models affects users' views. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, CHI '23*, 1–15. https://doi.org/10.1145/3544548.3581196

Jensen, D., Danks, E., Elbaum, S., Regli, W., Turk, M., Wierman, A., Yanco, H., Maher, M., & Griffin, H. (2025). *Envisioning possible futures for AI research*. Computing Research Association. https://cra.org/ccc/wp-content/uploads/sites/2/2025/07/Envisioning_Possible_Futures_for_AI_ Research.pdf

Jiang, C., Pan, X., Hong, G., Bao, C., Chen, Y., & Yang, M. (2025). *Feedback-guided extraction of knowledge base from retrieval-augmented LLM applications* (arXiv:2411.14110). arXiv. https://doi.org/10.48550/arXiv.2411.14110

References

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence,* 1(9), 389–399.

Johri, A., Hingle, A., & Schleiss, J. (2024). Misconceptions, pragmatism, and value tensions: Evaluating students' understanding and perception of generative AI for education. *2024 IEEE Frontiers in Education Conference (FIE)*, 1–9. https://doi.org/10.1109/FIE61694.2024.10893017

Johri, A., Katz, A. S., Qadir, J., & Hingle, A. (2023). Generative artificial intelligence and engineering education. *Journal of Engineering Education, 112*(3), 572–577. https://doi.org/10.1002/jee.20537

Jowsey, T., Braun, V., Clarke, V., Lupton, D., & Fine, M. (2025). We reject the use of generative artificial intelligence for reflexive qualitative research. *Qualitative Inquiry*, 10778004251401851. https://doi.org/10.1177/10778004251401851

Kancherla, J. (2020). Re-identification of health data through machine learning. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3794927

Katz, A., Shakir, U., & Chambers, B. (2023). *The utility of large language models and generative AI for education research* (arXiv:2305.18125). arXiv. https://doi.org/10.48550/arXiv.2305.18125

Kieser, F., Wulff, P., Kuhn, J., & Küchemann, S. (2023). Educational data augmentation in physics education research using ChatGPT. *Physical Review Physics Education Research, 19*(2), 020150. https://doi.org/10.1103/PhysRevPhysEducRes.19.020150

Klein, V., Savaş, Ö., & Conley, T. D. (2022). How WEIRD and androcentric is sex research? Global inequities in study populations. *The Journal of Sex Research, 59*(7), 810–817. https://doi.org/10.1080/00224499.2021.1918050

Klosek, K., & Blumenthal, M. (2024, January 23). Training generative AI models on copyrighted works is fair use. *Association of Research Libraries*. https://www.arl.org/blog/training-generative-ai-models-on-copyrighted-works-is-fair-use/

Kousha, K. (2024). How is ChatGPT acknowledged in academic publications? *Scientometrics, 129*(12), 7959–7969. https://doi.org/10.1007/s11192-024-05193-y

Kwon, D. (2025). Is it OK for AI to write science papers? Nature survey shows researchers are split. *Nature, 641*(8063), 574–578. https://doi.org/10.1038/d41586-025-01463-8

Lee, G., Yun, M., Zhai, X., & Crippen, K. (2025). Artificial intelligence in science education research: Current states and challenges. *Journal of Science Education and Technology*. https://doi.org/10.1007/s10956-025-10239-8

Lemley, M. A., & Ouellette, L. L. (2025). *Plagiarism, copyright, and AI* (SSRN Scholarly Paper No. 5399463). Social Science Research Network. https://doi.org/10.2139/ssrn.5399463

Lewis, J. E., Whaanga, H., & Yolgörmez, C. (2025). Abundant intelligences: Placing AI within Indigenous knowledge frameworks. *AI & SOCIETY, 40*(4), 2141–2157. https://doi.org/10.1007/s00146-024-02099-4

Li, Q., & Chen, Y.-L. (2009). Data flow diagram. In Q. Li & Y.-L. Chen (Eds.), *Modeling and Analysis of Enterprise and Information Systems: From Requirements to Realization* (pp. 85–97). Springer. https://doi.org/10.1007/978-3-540-89556-5_4

Liang, W., Yuksekgonul, M., Mao, Y., Wu, E., & Zou, J. (2023). GPT detectors are biased against non-native English writers. *Patterns, 4*(7), 100779. https://doi.org/10.1016/j.patter.2023.100779

Liu, J., Zhang, J., & Wang, S. (2025). *Exposing privacy risks in graph retrieval-augmented generation* (arXiv:2508.17222). arXiv. https://doi.org/10.48550/arXiv.2508.17222

Łodzikowski, K., Foltz, P. W., & Behrens, J. T. (2024). Generative AI and its educational implications. In D. Kourkoulou, A.-O. (Olnancy) Tzirides, B. Cope, & M. Kalantzis (Eds.), *Trust and Inclusion in AI-Mediated Education: Where Human Learning Meets Learning Machines* (pp. 35–57). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-64487-0_2

References

Lucchi, N. (2024). ChatGPT: A case study on copyright challenges for generative artificial intelligence systems. *European Journal of Risk Regulation, 15*(3), 602–624. https://doi.org/10.1017/err.2023.59

Magesh, V., Surani, F., Dahl, M., Suzgun, M., Manning, C. D., & Ho, D. E. (2025). Hallucination-free? Assessing the reliability of leading AI legal research tools. *Journal of Empirical Legal Studies, 22*(2), 216–242. https://doi.org/10.1111/jels.12413

Mahmoud, A. B., Kumar, V., & Spyropoulou, S. (2025). Identifying the public's beliefs about generative artificial intelligence: A big data approach. *IEEE Transactions on Engineering Management*. https://ieeexplore.ieee.org/abstract/document/10854904/

Marrinan, C., Hinton, L., Goins, R., & Koshy, S. (2025). *The unequal burden of data centers: An examination of the environmental and public health impacts on communities in California*. https://kaporfoundation.org/wp-content/uploads/2025/12/The-Unequal-Burden-of-Data-Centers-2025.pdf

McAdoo, T., Denneny, S., & Comments, C. L. 16 min read 9. (2025, September 9). *Citing generative AI in APA Style: Part 1—Reference formats*. https://Apastyle.Apa.Org. https://apastyle.apa.org/blog/cite-generative-ai-references

McGill, M. M., Heckman, S., Chytas, C., Liut, M., Kazakova, V., Sanusi, I. T., Shah, S. M., & Szabo, C. (2023). Conducting Sound, Equity-Enabling Computing Education Research. *Proceedings of the 2023 Working Group Reports on Innovation and Technology in Computer Science Education, ITiCSE-WGR '23*, 30–56. https://doi.org/10.1145/3623762.3633495

Medvedev, D., Davenport, D., Talhelm, T., & Li, Y. (2024). The motivating effect of monetary over psychological incentives is stronger in WEIRD cultures. *Nature Human Behaviour, 8*(3), 456–470. https://doi.org/10.1038/s41562-023-01769-5

Melumad, S., & Yun, J. H. (2025). Experimental evidence of the effects of large language models versus web search on depth of learning. *PNAS Nexus, 4*(10), pgaf316. https://doi.org/10.1093/pnasnexus/pgaf316

Merken, S. (2025, October 23). Two federal judges say use of AI led to errors in US court rulings. *Reuters*. https://www.reuters.com/sustainability/society-equity/two-federal-judges-say-use-ai-led-errors-us-court-rulings-2025-10-23/

Messeri, L., & Crockett, M. J. (2024). Artificial intelligence and illusions of understanding in scientific research. *Nature, 627*(8002), 49–58. https://doi.org/10.1038/s41586-024-07146-0

Mohawesh, R., Ottom, M. A., & Salameh, H. B. (2025). A data-driven risk assessment of cybersecurity challenges posed by generative AI. *Decision Analytics Journal, 15*, 100580. https://doi.org/10.1016/j.dajour.2025.100580

Mulally, T. (2024). An experiential journey: A year of a professor using AI in the classroom and research. *International Journal of Studies in Education and Science, 5*(3), Article 3. https://doi.org/10.46328/ijses.98

Ngwenyama, O., & Rowe, F. (2024). Should we collaborate with AI to conduct literature reviews? Changing epistemic values in a flattening world. *Journal of the Association for Information Systems, 25*(1), 122–136. https://doi.org/10.17705/1jais.00869

Nielsen, M., Haun, D., Kärtner, J., & Legare, C. H. (2017). The persistent sampling bias in developmental psychology: A call to action. *Journal of Experimental Child Psychology, 162,* 31–38. https://doi.org/10.1016/j.jecp.2017.04.017

Omar, M., Soffer, S., Agbareia, R., Bragazzi, N. L., Apakama, D. U., Horowitz, C. R., Charney, A. W., Freeman, R., Kummer, B., Glicksberg, B. S., Nadkarni, G. N., & Klang, E. (2025). Sociodemographic biases in medical decision making by large language models. *Nature Medicine, 31*(6), 1873–1881. https://doi.org/10.1038/s41591-025-03626-6

Østergaard, S. D., & Nielbo, K. L. (2023). False responses from artificial intelligence models are not hallucinations. *Schizophrenia Bulletin, 49*(5), 1105–1107. https://doi.org/10.1093/schbul/sbad068

Pack, A., & Maloney, J. (2023). Using generative artificial intelligence for language education research: Insights from using OpenAI's ChatGPT. *TESOL Quarterly, 57*(4), 1571–1582. https://doi.org/10.1002/tesq.3253

References

Paris, B., Conti-Cook, C., Greene, D., Jones, K. M., Justie, B., Kirschenbaum, M., Kresge, L., May, E., Nguyen, A., & Reynolds, R. (2025). Artificial intelligence and academic professions. *Available at SSRN 5376147*. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5376147

Perrigo, B. (2023, January 18). *OpenAI used Kenyan workers on less than $2 per hour*. Time. https://time.com/6247678/openai-chatgpt-kenya-workers/

Peters, U., & Chin-Yee, B. (2025). Generalization bias in large language model summarization of scientific research. *Royal Society Open Science*, 12(4), 241776. https://doi.org/10.1098/rsos.241776

Poole-Dayan, E., Roy, D., & Kabbara, J. (2025). *LLM Targeted Underperformance Disproportionately Impacts Vulnerable Users* (arXiv:2406.17737). arXiv. https://doi.org/10.48550/arXiv.2406.17737

Qin, P., Yang, C.-L., Li, J., Wen, J., & Lee, Y.-C. (2025). Timing matters: How using LLMs at different timings influences writers' perceptions and ideation outcomes in AI-assisted ideation. *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems, CHI '25*, 1–16. https://doi.org/10.1145/3706598.3713146

Rababah, B., Shang, Wu, Kwiatkowski, M., Leung, C., & Akcora, C. G. (2024). *SoK: Prompt hacking of large language models* (arXiv:2410.13901). arXiv. https://doi.org/10.48550/arXiv.2410.13901

Rajput, S., Saad, M., & Sharma, T. (2025). *Tu(r)ning AI Green: Exploring Energy Efficiency Cascading with Orthogonal Optimizations* (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2506.18289

Reif, J. A., Larrick, R. P., & Soll, J. B. (2025). Evidence of a social evaluation penalty for using AI. *Proceedings of the National Academy of Sciences, 122*(19), e2426766122. (world). https://doi.org/10.1073/pnas.2426766122

Rilla, R., Werner, T., Yakura, H., Rahwan, I., & Nussberger, A.-M. (2025). *Recognising, anticipating, and mitigating LLM pollution of online behavioural research* (arXiv:2508.01390). arXiv. https://doi.org/10.48550/arXiv.2508.01390

Rocher, L., Muthu, M. S., & de Montjoye, Y.-A. (2021). The observatory of anonymity: An interactive tool to understand re-identification risks in 89 countries. *Companion Proceedings of the Web Conference 2021, WWW '21*, 687–689. https://doi.org/10.1145/3442442.3458606

Roe, J. (2025). *How to use generative AI in educational research*. Cambridge University Press.

Salinas, A., Haim, A., & Nyarko, J. (2025). *What's in a name? Auditing large language models for race and gender bias* (arXiv:2402.14875). arXiv. https://doi.org/10.48550/arXiv.2402.14875

Samuelson, P. (2023). Generative AI meets copyright. *Science, 381*(6654), 158–161. https://doi.org/10.1126/science.adi0656

Santiago Jr, C., Embang, S., Acanto, R., Ambojia, K. W., Aperocho, M. D., Balilo Jr, B., Cahapin, E., Conlu, M., Lausa, S., Malabag, B., Paderes, J., & Romasanta, J. (2023). *Utilization of writing assistance tools in research in selected higher learning institutions in the Philippines: A text mining analysis. 22*, 259–284.

Sathish, V., Lin, H., Kamath, A. K., & Nyayachavadi, A. (2024). *LLeMpower: Understanding disparities in the control and access of large language models* (arXiv:2404.09356). arXiv. https://doi.org/10.48550/arXiv.2404.09356

Schmitt, D. P., Allik, J., McCrae, R. R., & Benet-Martínez, V. (2007). The geographic distribution of big five personality traits: Patterns and profiles of human self-description across 56 nations. *Journal of Cross-Cultural Psychology, 38*(2), 173–212. https://doi.org/10.1177/0022022106297299

Shah-Dand, M. (2025). *Responsible AI vs. AI ethics—What's the difference?* Women in AI Ethics. https://womeninaiethics.org/wp-content/uploads/2025/05/Responsible-AI-vs-AI-Ethics.pdf

References

Siau, K., & Wang, W. (2020). Artificial intelligence (AI) ethics: Ethics of AI and ethical AI. *Journal of Database Management (JDM), 31*(2), 74–87.

Skulmowski, A. (2024). Placebo or assistant? Generative AI between externalization and anthropomorphization. *Educational Psychology Review, 36*(2), 58. https://doi.org/10.1007/s10648-024-09894-x

Smith, J. M. (2024). "I'm sorry, but I can't assist": Bias in generative AI. *Proceedings of the 2024 RESPECT Annual Conference, RESPECT 2024*, 75–80. https://doi.org/10.1145/3653666.3656065

Staab, R., Vero, M., Balunović, M., & Vechev, M. (2023). *Beyond memorization: Violating privacy via inference with large language models* (arXiv:2310.07298; Version 1). arXiv. https://doi.org/10.48550/arXiv.2310.07298

Stefan, A. M., & Schönbrodt, F. D. (2023). Big little lies: A compendium and simulation of *p*-hacking strategies. *Royal Society Open Science, 10*(2), 220346. https://doi.org/10.1098/rsos.220346

Teo, Z. L., Quek, C. W. N., Wong, J. L. Y., & Ting, D. S. W. (2024). Cybersecurity in the generative artificial intelligence era. *Asia-Pacific Journal of Ophthalmology, 13*(4), 100091. https://doi.org/10.1016/j.apjo.2024.100091

Tobin, S. (2025, November 4). *Getty Images largely loses landmark UK lawsuit over AI image generator | Reuters*. https://www.reuters.com/sustainability/boards-policy-regulation/getty-images-largely-loses-landmark-uk-lawsuit-over-ai-image-generator-2025-11-04/

Turpin, M., Michael, J., Perez, E., & Bowman, S. R. (2023). *Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting* (arXiv:2305.04388). arXiv. https://doi.org/10.48550/arXiv.2305.04388

Twinomurinzi, H., & Gumbo, S. (2025). Generative AI: Concerns, usage, challenges, opportunities and sentiments. *South African Computer Journal, 37*(1). https://doi.org/10.18489/sacjv37i1/17890

United States Copyright Office. (2025). *Copyright and artificial intelligence. Part 2: Copyrightability*. https://www.copyright.gov/ai/Copyright-and-Artificial-Intelligence-Part-2-Copyrightability-Report.pdf

Uzondu, D. (2025, November 7). *Google improves AI summaries of forms responses*. Neowin. https://www.neowin.net/news/google-improves-ai-summaries-of-forms-responses/

Wach, K., Duong, C. D., Ejdys, J., Kazlauskaitė, R., Korzynski, P., Mazurek, G., Paliszkiewicz, J., & Ziemba, E. (2023). The dark side of generative artificial intelligence: A critical analysis of controversies and risks of ChatGPT. *Entrepreneurial Business and Economics Review, 11*(2), 7–24.

Wagner, G., Lukyanenko, R., & Paré, G. (2022). Artificial intelligence and the conduct of literature reviews. *Journal of Information Technology, 37*(2), 209–226. https://doi.org/10.1177/02683962211048201

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E. H., Le, Q. V., & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Proceedings of the 36th International Conference on Neural Information Processing Systems, NIPS '22*, 24824–24837.

Westwood, S. J. (2025). The potential existential threat of large language models to online survey research. *Proceedings of the National Academy of Sciences, 122*(47), e2518075122. https://doi.org/10.1073/pnas.2518075122

Woodruff, D. P., Cohen-Addad, V., Jain, L., Mao, J., Zuo, S., Bateni, M., Branzei, S., Brenner, M. P., Chen, L., Feng, Y., Fortnow, L., Fu, G., Guan, Z., Hadizadeh, Z., Hajiaghayi, M. T., JafariRaviz, M., Javanmard, A., S, K. C., Kawarabayashi, K., … Mirrokni, V. (2026). Accelerating Scientific Research with Gemini: Case Studies and Common Techniques (arXiv:2602.03837; Version 2). arXiv. https://doi.org/10.48550/arXiv.2602.03837

References

Wright, D., Masud, S., Moore, J., Yadav, S., Antoniak, M., Christensen, P. E., Park, C. Y., & Augenstein, I. (2026). *Epistemic Diversity and Knowledge Collapse in Large Language Models* (arXiv:2510.04226). arXiv. https://doi.org/10.48550/arXiv.2510.04226

Wu, F., Black, E., & Chandrasekaran, V. (2024). *Generative monoculture in large language models* (arXiv:2407.02209). arXiv. https://doi.org/10.48550/arXiv.2407.02209

Wu, K., Wu, E., Wei, K., Zhang, A., Casasola, A., Nguyen, T., Riantawan, S., Shi, P., Ho, D., & Zou, J. (2025). An automated framework for assessing how well LLMs cite relevant medical references. *Nature Communications, 16*(1), 3615. https://doi.org/10.1038/s41467-025-58551-6

Xia, W., Basford, M., Carroll, R., Clayton, E. W., Harris, P., Kantacioglu, M., Liu, Y., Nyemba, S., Vorobeychik, Y., Wan, Z., & Malin, B. A. (2023). Managing re-identification risks while providing access to the All of Us research program. *Journal of the American Medical Informatics Association, 30*(5), 907–914. https://doi.org/10.1093/jamia/ocad021

Xu, Z., Jain, S., & Kankanhalli, M. (2024). *Hallucination is inevitable: An innate limitation of large language models* (arXiv:2401.11817). arXiv. https://doi.org/10.48550/arXiv.2401.11817

Yan, L., Sha, L., Zhao, L., Li, Y., Martinez-Maldonado, R., Chen, G., Li, X., Jin, Y., & Gašević, D. (2024). Practical and ethical challenges of large language models in education: A systematic scoping review. *British Journal of Educational Technology, 55*(1), 90–112. https://doi.org/10.1111/bjet.13370

Zack, T., Lehman, E., Suzgun, M., Rodriguez, J. A., Celi, L. A., Gichoya, J., Jurafsky, D., Szolovits, P., Bates, D. W., Abdulnour, R.-E. E., Butte, A. J., & Alsentzer, E. (2023). *Coding inequity: Assessing GPT-4's potential for perpetuating racial and gender biases in healthcare* (p.2023.07.13.23292577). medRxiv. https://doi.org/10.1101/2023.07.13.23292577

Zarifhonarvar, A. (2024). Economics of ChatGPT: A labor market view on the occupational impact of artificial intelligence. *Journal of Electronic Business & Digital Economics, 3*(2), 100–116.

Zhang, Y., & Gosline, R. (2023). Human favoritism, not AI aversion: People's perceptions (and bias) toward generative AI, human experts, and human–GAI collaboration in persuasive conten generation. *Judgment and Decision Making, 18, e*41. https://doi.org/10.1017/jdm.2023.37

Zhou, K., Constantinides, M., & Quercia, D. (2025). *Should LLMs be WEIRD? Exploring WEIRDness and human rights in large language models*. https://doi.org/10.48550/ARXIV.2508.19269

Zhou, M., Abhishek, V., Derdenger, T., Kim, J., & Srinivasan, K. (2024). *Bias in generative AI* (arXiv:2403.02726). arXiv. https://doi.org/10.48550/arXiv.2403.02726

Zittrain, J. (2022). Intellectual debt: With great power comes great ignorance. In O. Mueller, P. Kellmeyer, S. Voeneky, & W. Burgard (Eds.), *The Cambridge Handbook of Responsible Artificial Intelligence: Interdisciplinary Perspectives* (pp. 176–184). Cambridge University Press. https://doi.org/10.1017/9781009207898.014

Appendix

Guidelines Development Process

In late summer 2025, our project team sent invitations to participate to a variety of venues (e.g., relevant AERA special interest groups, Society of Women Engineers, relevant ASEE divisions, Society for the Advancement of Biology Education Research, the special interest group for computer science education of the Association for Computing Machinery). Over 120 individuals completed the application form, which requested information about the applicant's institution and job title, their focus area within STEM (e.g., biology, computer science), years of experience (i.e., current career, conducting STEM-ER, building AI tools, using AI tools), type of research conducted (e.g., qualitative), knowledge of topic areas related to responsible AI use (e.g., data privacy issues), and degree of support for and skepticism about AI.

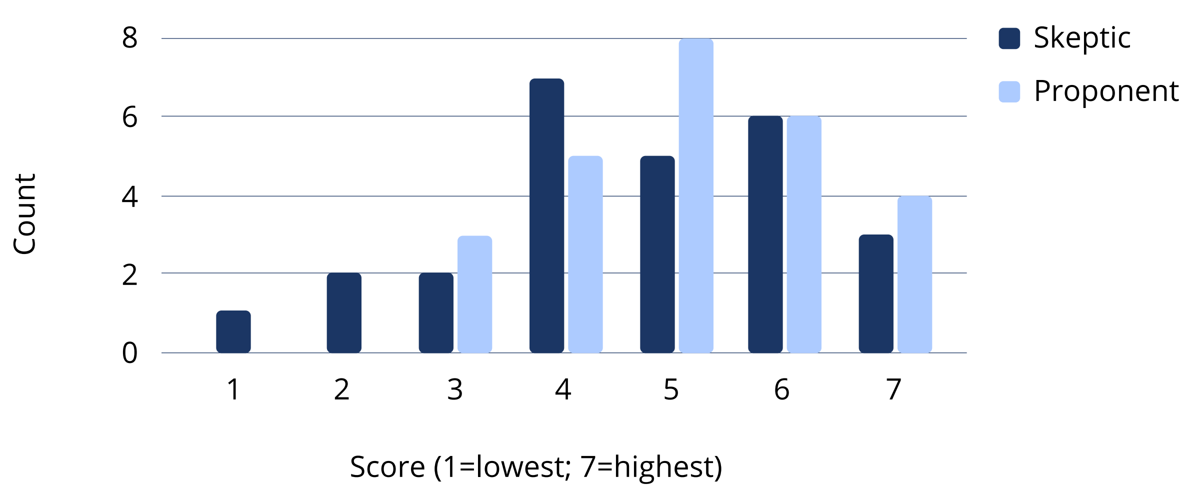


FIGURE 7: AI Proponent and AI Skeptic Self-Assessment Scores of workshop participants.

Based on this information and in order to ensure a broad level of expertise in the group, 27 participants and the project team participated in a two-day in-person workshop held in Chicago in November 2025. Applicants who were not selected (or who were not available) for the workshop, or who expressed interest in the project in other settings (e.g., a conference talk) were invited to provide feedback on draft versions of this report.

At the workshop, participants engaged in a variety of collaborative activities to generate and then refine the guidelines. For example, participants began by exploring “blue sky” (i.e., idealistic) and “grey sky” (problematic) outcomes of AI use in STEM-ER for each of the steps of the research process and topics related to responsible use. Several guest speakers highlighted specific issues, such as whether it is possible to future-proof guidelines given the rapid evolution of AI.

After the workshop, the project team synthesized the draft guidelines, other artifacts, and related research to generate a first draft of this report. The steering committee provided feedback on that draft, and an updated draft was shared with workshop participants for their feedback. Once updated, the draft was then shared with other reviewers for their feedback.

Throughout the project, the steering committee provided guidance and feedback.

Research Report · May 2026

Guidelines for the Responsible Use of AI in STEM Education Research

Protecting Our Participant Communities and Research Integrity

PRINCIPAL INVESTIGATOR
Dr. Monica McGill

CO-PRINCIPAL INVESTIGATOR
Dr. Julie M. Smith

FUNDING
NSF Award No. 2519885