Innovations in Pedagogy and Technology
https://journals.niepublish.com/ipt

ARTICLE

Educators' Discussion with a Generative Artificial Intelligence Trainer

Amanda Gantt Sawyer [1*] , ***Devin Maxwell Christian*** [2] , ***Pierre Sutherland*** [3]

[1] *Middle, Secondary, and Mathematics Education Department, College of Education, James Madison University, Harrisonburg, VA 22802, USA*

[2] *Scale AI, San Francisco, CA 94103, USA*

[3] *Sparx Learning, Bristol BS20 6RE, UK*

ABSTRACT

Educators have investigated the impact of generative Artificial Intelligence (genAI) tools, such as ChatGPT and Large Language Models (LLMs), on curriculum development in K-12 schools. Still, little is known about the development of these tools and their impact on the future of education. Therefore, a teacher educator conducted a duoethnography to investigate these problems with a genAI trainer, specifically a Reinforcement Learning from Human Feedback (RLHF) Administrator, to answer the following questions. How does an RLHF administrator train genAI tools? What should mathematics educators know based on this training? What impact could this new technology have on the future of education? The data indicate that errors and error-finding are central parts of training LLMs. Since humans train the genAI, it inherits their biases. We learned that genAI was created to be people pleasing which can override its ability to find truth. We learn how it supports error analysis in Large Language Models, and how the public should view these tools. Finally, we describe what RLHF administrators view as the role of genAI in education, what the public should be aware of, and the future of the genAI market. The data highlights the need for teachers to consider the impact of profit motivation, regulation, and economic implications on genAI tools.

Keywords: Generative Artificial Intelligence; Reinforcement Learning from Human Feedback; Duoethnography

*CORRESPONDING AUTHOR:

Amanda Gantt Sawyer, Middle, Secondary, and Mathematics Education Department, College of Education, James Madison University, Harrisonburg, VA 22802, USA; Email: sawyerag@jmu.edu

ARTICLE INFO

Received: 10 March 2026 | Revised: 15 June 2026 | Accepted: 22 June 2026 | Published Online: 2 July 2026
DOI: https://doi.org/10.63385/ipt.v2i3.101084

CITATION

Sawyer, A.G., Christian, D.M., Sutherland, P., 2026. Educators' Discussion with a Generative Artificial Intelligence Trainer. Innovations in Pedagogy and Technology. 2(3): 42–53. DOI: https://doi.org/10.63385/ipt.v2i3.101084

1. Introduction

Technology is changing rapidly, and the newest curriculum developers on the market are generative artificial intelligence (genAI) tools[1]. Teacher Educators (TEs) have studied these tools and found that pre-service teachers are often overconfident in the abilities of generative AI (genAI) chatbots[2]. Other researchers have found that genAI chatbots can be mathematically incorrect[3], produce biased answers[4], and generate inappropriate responses for students[5]. However, most research on genAI suggests that training methods are not well understood and that there is limited information available on its creation[6]. Therefore, as TEs, we discussed this with a genAI trainer, who holds the formal job title of Reinforcement Learning from Human Feedback (RLHF) Administrator. This investigation can serve as an informational document on genAI, the role of the RLHF Administrator, the impact of genAI on education, and its future implications for education.

Due to their popularity, many researchers have begun investigating genAI[1, 7, 8]. The concept has existed since the 1950s, when Turing coined the term "Artificial Intelligence"[9]. However, it is the modern use of artificial intelligence in the genAI realm that has benefited significantly from public attention and use. We define genAI as a subset of AI that creates new content based on the data it was given or trained on[8]. GenAI utilizes Large Language Models (LLMs), computer models trained on vast amounts of data, enabling them to understand and generate natural-language responses for a wide range of tasks[10].

In November 2022, OpenAI introduced its tool ChatGPT, attracting millions of new users interested in its capabilities and utilizing it in new areas, such as education. The US Department of Education acknowledged the widespread adoption of genAI chatbots[11]. It emphasized that further research is needed to determine their effectiveness and whether they should be used in the classroom. The National Council of Teachers of Mathematics[12] also recognized its usefulness but cautioned about potential harm in educational applications. This has led to numerous research studies examining how teachers can utilize these tools[7], the errors these tools can create[3], and the knowledge and skills teachers need to use them effectively[2]. However, a common theme emerged across all those research projects: we know very little about how genAI tools are created. Therefore, a mathematics teacher educator conducted a duoethnography with an RLHF administrator to investigate these problems. In this paper, we examined the training methods individuals use to develop these tools and how teachers can use this knowledge to support their students effectively. Our research questions were:

1. How does an RLHF administrator train genAI tools?
2. What should mathematics educators know based on this training?
3. What impact could this new technology have on the future of education?

2. Literature

Although the use of genAI in education classrooms is a relatively new field, researchers and computer scientists have provided extensive knowledge about these tools, cautioning others on their capabilities and limitations[13, 14]. The first caution from the literature is that although genAI tools can produce new information, the information they generate can be biased[4]. Specifically, this can be harmful to minority populations, with one researcher identifying that when the tool is asked for details about felons, it tends to create names of individuals who are Latino or African American[13]. Other researchers recognize that the information can lean in a specific political direction; thus, the answers may be biased[15].

The second area of caution concerns the factual nature of responses constructed by these genAI tools. It is known that they can hallucinate[16] and create inaccurate information or faulty information that has nothing to do with what exists[2]. For example, when asked to construct a list of children's books on telling time, the program created nonexistent books[17]. We also know it will confidently solve mathematical problems, but it will also be completely inaccurate[3]. Therefore, the researchers have identified that we should not trust any information generated by genAI tools as facts without verifying their validity.

Another area of caution regards its impact on the environment. We know that these genAI tools require enormous resources, including infrastructure, urban production, energy consumption, and freshwater[14]. This should concern all users of genAI, as their environmental impact will significantly affect the future. Specifically, generating images with

genAI tools can consume more energy than simply asking a question, cautioning users to consider environmental impact. The next area of caution is privacy. When individuals sign up to use genAI tools, they must agree to terms stating that the data may be used by the genAI tool to construct future answers and determine its responses to the user[18]. Therefore, when individuals' images, likenesses, or intellectual property are input into these genAI tools, they can be subject to these terms, essentially exploiting that IP for model-building purposes. Due to these policies, many universities and states prohibit the use of certain genAI tools on their public networks because the materials are sensitive[18]. For example, DeepSeek has been identified as a tool that should not be downloaded or used by individuals who are using government property[19]. There are numerous privacy concerns, and a significant amount of data is being provided willingly to these genAI tools, raising concerns about the implications for the future of privacy.

And finally, considerable research has examined who is responsible for the use of genAI tools in education[2, 20]. The question becomes the following: is it the responsibility of the creators of the tools, the teachers who implement them, or the students who use them to determine whether the information is accurate, unbiased, and appropriate? There is speculation about what to do in this regard. Still, ultimately, if our students are minors, the teacher must understand the issues associated with these technologies and have the right to determine whether it is appropriate to use them with their students. Because of these concerns, teachers first need to decide whether the activity they are trying to design for their mathematics classroom is worth the time required to verify the information's validity, the risk of exposing private information, and even the environmental impact[20].

Duoethnography Literature

Duoethnography is a dialogic methodology in which researchers investigate issues by drawing on their differing life experiences[21]. This methodology was developed in 2003 and introduced to the mathematics education community in Rapke's[22] article, "*Duoethnography: A New Research Methodology for Mathematics Education*". This methodology involves the dialogue in joint reflection between collaborating researchers. The researchers must work as a team to reevaluate their research by reflecting on their own past experiences and insights. This is done only by reflecting on differing life experiences; thus, the researchers must be at varying stages or in different areas of research[21]. For example, in our investigation, Author 2 serves as a Reinforcement Learning from Human Feedback (RLHF) Administrator, while Authors 1 and 3 are mathematics teacher educators.

The data collected for duoethnography come from representations of memories, meaning-making, and re-understandings constructed through conversations, as well as new realizations derived from reflection on the study. For example, if we are to investigate understanding of fractions, we could engage in a dialogue about how we learn fractions, what has worked in the past, what has not, and develop a new understanding of fraction teaching based on writings and conversations. Duoethnography is a fluid process in which the researchers work together to construct a new understanding. Latz and Murray explained, "In dialogue with another, I/we will be able to revisit, retell, and re-understand my/our stories, something not possible through self-reflection alone" (p. 2)[23].

The data analysis for duoethnography is conducted through dialoguing and reflecting on that dialogue. Individuals construct their understanding in the moment, then, after transcribing, reevaluate what has come to pass, and construct the results. The researcher must reflect on what was done and what was collected to construct their own understanding of the content. However, the data collected were the researchers' memories and experiences. Thus, the experience is more fluid. This form of methodology differs from autoethnography in that reflection alone cannot yield the same understanding. Having the other researcher give counter-narratives or differing experiences helps to develop new knowledge[21].

Manuscripts using duoethnography are written as a dialogue. Some manuscripts also include an explanation of the research but present their findings and discussions as conversations. Readers are expected to construct their understanding of the concept through reading the dialogue constructed by the co-researchers. Traditional Duoethnography expects its readers to construct meaning for themselves rather than have it explicitly stated[21]. Duoethnography offers researchers the opportunity to develop and explore previously unexpressed concepts. Researchers with varying degrees of experience in a single area can offer knowledge

and support and facilitate learning through reflection. Thus, duoethnography is a form of qualitative research that allows two researchers to explore and reflect on a topic, constructing their own meaning and understanding [24].

3. Methodology

To conduct this research, we employed duoethnography [21–25]. This form of qualitative research examines participants' perspectives in their natural settings. Our findings stemmed from informal, spontaneous, rich, and in-depth interviews that focused on understanding how genAI is trained through personal experiences with RLHF, while contrasting this understanding with TE's understanding, informed by research and teaching in mathematics classrooms [24, 25].

Author 1 conducted an exploratory interview with Author 2, an expert in genAI training. We engaged in a reflective, progressive problem-solving process to better understand the effects of genAI and explore ways to address the issue for teachers. Our interview, the primary data source, focused on real-world situations and provided detailed descriptions. We used our personal stories to gain a clearer understanding of genAI in the classroom and how our experiences shaped the way we view the subject today. After gathering, analyzing, and interpreting the interview, we worked together to understand the data and provide implications for the classroom.

3.1. Participants

The participants in this study include the authors. Author 1 is an Associate Professor of Mathematics Education at James Madison University with over 14 years of experience teaching mathematics content and methods courses at the university level. She earned a doctoral degree in Mathematics Education from the University of Georgia, a master's degree in applied mathematics, and a Master's in Elementary Education from the University of South Carolina. Her research focuses on how mathematics teachers select, critique, and adapt educational resources from online sources, specifically genAI.

Author 2 worked as an RLHF admin for over four years. His educational background is in journalism (ABJ) from Mass Media Arts at the University of Georgia, with an early career in new media production focused on online learning. His interest in data science stemmed from his experiences in media production, specifically in big data analytics with platforms like YouTube and Facebook Ads. Hours spent poring over engagement analytics set him on the path of investigating how professionals make sense of all this data. This coincided nicely with ChatGPT's explosion in popularity, and as the field heated up, he found an opportunity to work with a company that was collecting training data for LLMs. He produced training data for over a year before moving into an administrative role. He liked to summarize his work with machine learning as this cheekily: "I started out trying to figure out how to get machines to do what we want them to; now I try to figure out how to get other humans to get machines to do the same."

Author 3 is an experienced educator, EdTech founder, and instructional designer with a Ph.D. in Mathematics Education from the University of Georgia. His expertise spans teacher training, eLearning management systems, and the integration of digital technologies in mathematics instruction. With a background in computer science and mathematics education, he has taught at both secondary and tertiary levels in South Africa, China, and the United States, where he has worked extensively in teacher education and curriculum development.

Currently based in the United Kingdom, Author 3 is a School Success Coach at Sparx Learning, an online homework EdTech company that specializes in supporting secondary school teachers in integrating technology into their instructional contexts. His work focuses on developing scalable, research-informed educational solutions that enhance teaching and learning outcomes. His research interests include decentralized education models and the application of AI in education

The differences within our similarities made this duoethnography meaningful. The conversation between Author 1 and Author 2 was analyzed and triangulated with Author 3's perspective. Thus, participation in this study, which involves interviews, suggestions, feedback, and reflection, will enhance its significance.

3.2. Data Collection

We collected data from an interview transcript. To improve the quality of our discussion, we can stray from the central topic and allow the interview to develop fully. We as-

sisted each other in remembering past events by sharing our own autobiographical memories. These recollections created connections that we analyzed from our different perspectives, resulting in a more comprehensive self-study [25].

This fully IRB-approved study included all authors, and we recorded connections, insights, reactions, and preliminary themes that emerged during the interview. This written thought catalog allowed us to debrief and dig deeper into our conversation. Some editing was done to the transcript to improve readability and focus. The main artifacts associated with this study are the transcript, our emails, and our reflections. Thus, our dialogue served as the primary tool for data collection.

To ensure the validity of our data, we conducted member checking, with Author 3 serving as an independent third party to validate the implications of these findings. All three authors reread the transcript and followed up with clarifying questions or suggestions for editing. Additionally, we maintained an audit trail, documenting all key thoughts throughout the research process and noting when they occurred and why. Collaboration was an essential aspect of this duoethnography. Each author shared a differing perspective that challenged each member, resulting in a richer and more thorough understanding of our internal feelings and views on genAI. Lastly, we engaged in peer debriefing by communicating online about thoughts, reflections, ideas, and new findings [25].

3.3. Data Analysis

Our virtual communication and self-reflections were intended to identify themes in our dialogue through coding. These themes are listed individually in the findings section of this study and supported by additional, related references to ensure validity. The themes identified in this self-study supported the development of both participants as individuals, academics, and teachers. Personal transformation through critical reflection is the primary goal of duoethnography, which also serves to measure the study's effectiveness [25]. Such a transformation was documented in a reflective journal and a metacognitive hypertext, which was saved on a password-protected computer. The objective of the data analysis was to identify themes that facilitate a deeper understanding of genAI as a tool for enhancing students' mathematical comprehension, as well as its limitations.

The transcribed interview included the findings and discussions generated throughout the interview. Interview questions, found in **Appendix A**, were developed to guide the interview and ensure all research questions were thoroughly addressed. These questions were instrumental to the learning as they required me to think about the goals for the interview and decide what we wanted to know.

We used thematic analysis to identify the themes constructed in the duoethnography [25]. The interview was independently reviewed by each author, with each coding for specific major patterns and meanings. When the data presented new patterns, emerging themes were constructed from the participants' responses. The authors discussed their constructed themes and came to a consensus around how to describe the data. We reached 100% consensus on themes and meanings. After identifying the categories for all responses in the interview, we developed a codebook using a process and shared it with the other researcher (see **Table 1**). While these four preliminary themes (training genAI from human feedback, importance of errors, genAI as a people pleaser, know when to trust the tool, and genAI not going away) were intended to be building blocks as we continued to analyze the transcription via coding, we realized that they were the best way to describe the data and were complete. We used these themes throughout the analysis, shaping the overall findings.

Table 1. Themes with Definitions.

Theme	Definition
Training GenAI from Human Feedback	Data describes how human feedback is used to train the genAI tool.
Importance of Errors	Data describes when errors are used when using genAI training.
GenAI as a People Pleaser	Data describes responses by genAI that focus on pleasing the user.
When to Trust the Tool	Data describes when genAI is accurately creating reliable responses.
GenAI Is Not Going Away	Data describes lasting nature of genAI use.

Lastly, metacognitive hypertext, or personal memos, that we included in the transcribed interviews, allowed us to connect our thoughts and identify our opinions on the topics and strategies discussed. The analysis of collected data confirmed our preparedness and ability to effectively utilize the genAI tools.

4. Results

4.1. Training GenAI from Human Feedback

Author 1: Thank you for talking with me about AI. As a math teacher educator, I have observed my pre-service teachers using AI to complete homework and getting completely inaccurate answers. This really made me question how generative AI tools really create answers and how they are trained to determine if a response is appropriate. So, I contacted "Author 3" about this, and he mentioned knowing you and that you actually worked as a trainer of generative AI tools. What is your role? Does it relate to LLMs?

Author 2: I am a training administrator who focuses on collecting data from human subjects for machine learning. Our most common tool is reinforcement learning from human feedback (RLHF), though we also use side-by-side (SxS) evaluations and other labeling techniques. It all boils down to the same general idea: My job is to figure out how to get human raters to generate valuable data for training LLMs.

Author 1: So what does that look like? I can train teachers through classes, but I know it's more complicated when you're talking with computers instead of individuals. Does it include uploading information?

Author 2: It is common knowledge that LLMs are trained on massive quantities of data, but it is more than just uploading the works of William Shakespeare and the last 50 years of NYSE records. Once a basic language model is operational, additional functionality is trained through enormous amounts of human-powered A/B testing. For example, you will take two versions of a model (call them "Alpha" and "Bravo"), and a human will ask them a question ("What is bigger, Saturn or Uranus?"). In an ideal case, Alpha says Saturn, while Bravo says Uranus. The human then labels that conversation in some way, indicating that Alpha's response was better than Bravo's, and all of that data is fed back into the next generation of models, so the next time they encounter a similar question, they know Saturn is a better answer. This process occurs with hundreds to thousands of human trainers across hundreds of thousands of interactions.

Author 1: Wow, that's a lot of trainers working on one model. Are all genAI models trained this way?

Author 2: Yes, most major models are trained in similar ways. The variance in capability across models stems from the quantity and quality of their training data (as well as the scale of processing infrastructure available to a certain degree). One developer may be focused on making a model exceptionally accurate in its stated facts, so they will contract many training data points focused on accuracy. Another developer's primary concern may be ensuring their model cannot be used for nefarious purposes, so they will devote many resources to training data that reinforces content safeguards. Moreover, others may still want their model to be hyper-specialized for a specific industry, so they will focus on collecting training data from qualified industry professionals. However, most developers generally want the same thing–safe, accurate, and valuable models.

4.2. Importance of Errors

Author 1: What is your opinion on the tool's accuracy? I have observed that ChatGPT has issues with explaining multiplication problems to me, so I am a little wary of using it in my life. Do you use these tools, or is there one you trust more than another?

Author 2: In my professional work, I do not rely much on specific genAI tools because 1) I am aware of their limitations in terms of reliability, and 2) I do not want to propagate certain biases or patterns that manifest in the models inadvertently. You have to be careful of circular training that almost manifests as a sort of "inbreeding" with LLMs. For example, if you have a model that always responds with a three-bullet-point plan, and you ask it for ideas on how to train it further, it will likely give you a three-bullet-point plan that reinforces that behavior.

Author 1: So if it makes these errors, what do you view as your role in this process?

Author 2: My key position is as an intermediary between developers and trainers. Essentially, a developer comes to us with a specific objective ("We need X amount of training data focused on Y"), and it is my role to administer an operation that will satisfy that objective, from identifying

qualified human raters to structuring a task that will facilitate appropriate data collection.

Author 1: Does the data collection process come from errors?

Author 2: Errors are significant for training! Most of the value a trainer provides is not just in recognizing when an LLM makes an error, but specifically in encouraging the creation of errors to build a record of how an LLM should or should not respond. The critical element is that we need training data showing a significant divergence between an excellent and a lousy response. The most significant "learning" does not happen when you tell a machine over and over "2 + 2 = 4" (and indeed not when you tell it over and over "2 + 2 = 5"); it occurs when one version of a model says "2 + 2 = 4" *while* another version says "2 + 2 = 5" in that same situation.

Author 1: Then how do you find errors?

Author 2: We specifically instruct human raters to recognize, produce, and appropriately label those errors. If it is a specific type of error that we are targeting ("for some reason, this model hallucinates facts about celebrities; place the models in situations where they are likely to make such a hallucination"), there is usually direct guidance on how to label that error for training purposes. If it is something more spontaneous, there are more general flags raters can send up.

Author 1: What is the next step to "fix" these issues in the LLMs?

Author 2: It is usually the job of people like me to comb through as many of those reports I mentioned above as possible to see if the source of the error can be pinpointed and communicated to the developers accurately ("We notice the models struggle to interpret graphs with more than three colors," etc.). At that point, developers may adjust the models or collect more training data targeting that specific error to train out that behavior.

Author 1: What does this look like? Can you give me an example?

Author 2: Many people know the infamous "Sure, I can help with that..." LLMs speak.' That is an excellent example of what developers are aware of as an issue and seek to fix through RLHF. When responses become too canned or formulaic, a developer will call for more training data, either emphasizing a unique voice for the LLM or omitting the pleasantries entirely. So then a bunch of conversations are collected where someone asks how fast an armadillo can run, and Alpha says, "Sure, I can help with that" (and that is labeled as an inadequate response). In contrast, Bravo says, "Armadillos can run up to 30 miles per hour" (and that is labeled as a reasonable response), and those labels will teach the next version of the model, "It is better just to answer the question without the preamble."

4.3. GenAI as a People Pleaser

Author 1: When talking with students, teachers learn different tricks that help them communicate better. What insight do you believe you could provide others that is not common knowledge to the public?

Author 2: My most significant point is that LLMs do not "know" anything and cannot "reason" how we typically use that word. They provide the answer that will most likely elicit a positive response from their rater (and this is meant to mirror what is most likely to elicit a positive response from an intended user).

Author 1: So, you are saying that the tools are people pleasers. Since it wants to answer the user correctly, does this mean that it could learn the right answer?

Author 2: Not necessarily. LLMs reflect the quality and biases of the training data their developers collect. One salient example is when you ask DeepSeek about what occurred in Tiananmen Square in 1989. At best, LLMs are susceptible to the blind spots of the humans who develop and train them. At worst, they can lend an air of legitimacy to attempts at censorship and the dissemination of misinformation.

Even if you assume an LLM developer has done their level best at creating an accurate and unbiased model, the "motivation" the model has to give users the answer the model thinks they want to hear is a significant pitfall. Every major model defaults to a "yes and no" approach, regardless of how infeasible a user's request may be. This is an area that's being improved on, and safeguards are being trained into the models to prevent them from recommending recipes for meth or napalm. However, the very nature of LLM development puts people-pleasing in the driver's seat. Models tend to tell you what you want to hear, which makes them hazardous sources to put faith in.

A low-stakes example: If you ask a sensible friend to pick out a white outfit for a wedding, they will likely admonish you not to do so. Ask a model, however, and you will probably get an answer with hallucinated justifications for why you can get away with wearing white at a wedding, even if it is not traditionally socially acceptable.

A higher-stakes example: If you ask a statistician to make a series of high-level insights on a data set with an *n* of 8, they will likely not comply because practically no one would consider this a feasible sample size. Ask a model, however, and even if there is a disclaimer, you are still likely to get a response along the lines of "Sure, I can help with that..."

Author 1: So don't assume that if it has an answer, it is correct.

Author 2: Right. It will give you a response even if it has to make it up itself.

4.4. Know When to Trust the Tool

Author 1: What does this mean for educators or students who use LLMs? Should we say not to use it?

Author 2: The value of an LLM depends heavily on the specific query posed or the role it is being asked to fill. They perform well in a Wikipedia role (giving a general overview of a given topic with avenues for further research). They can even handle simple 1- or 2-step mathematical operations that are likely to have been carried out before in their training data. However, they tend to struggle with novel scenarios where they lack previous training interactions to reference, which is especially obvious when asked to work with multiple distinct concepts at once.

Here is an illustration. I would trust, "What is the probability of drawing the Queen of Hearts from a standard deck of playing cards on the first try?" This is a simple question with a simple answer, which you will likely find in a textbook. The model likely has training data on how to solve this correctly.

However, I would NOT trust the response for something along the lines of, "Three cards are lying face down in front of me. Two are face cards, and there are no spades. No card is next to a card of the same color. What are the odds that the card on the left is the Queen of Hearts?" This complex problem combines logical reasoning, spatial understanding, and probability in a unique way that the model is unlikely to have encountered in its training data. Therefore, applying the proper approach to solving it will be difficult.

Author 1: Teachers need to learn when they should or shouldn't trust the LLMs. When do you suggest educators and students trust the tools?

Author 2: Teachers can trust LLMs as great tools for brainstorming and getting exposure to subjects for self-directed research, but they make risky primary sources. Working with an LLM should almost always be considered a *preliminary* step, not the final one—and all factual claims made by an LLM should be scrutinized. Now, this iterative, reflective approach to machine assistance can be a strength for educators!

A great use case would be a high school teacher asking an LLM to generate problems for their class to work through. I would not recommend simply passing those problems out to the students! However, if the teacher reviews these machine-assisted problems, they will likely find, "Oh, this is too easy," "this is too hard," or "this is just right!" By analyzing the machine output and identifying those specific elements that are desirable/undesirable, the teacher may recognize areas for improvement in their assignments or instruction.

Alternatively, an educator might assign an LLM a problem it is likely to struggle with and then share its output with their students as a teaching tool: "There is a clear error in step 3. What should have been done differently?"

4.5. GenAI Is Not Going Away

Author 1: What about the future of these tools? I have been told by other AI researchers that within the next 5 years, it will be able to do the mathematics that it can't do now. What is your perspective on the future of the LLMs market?

Author 2: Data scientists more qualified than I am convinced artificial general intelligence is on the horizon, and from the massive amount of capital expenditure I have seen reported, winning that "space race" is a top priority for the biggest corporations on the planet. I do not want to be alarmist, but I think it is worth considering that for-profit corporations fund most LLM development. I think humankind can reap many benefits from its development; it just needs to be understood that the benefit may be a byproduct rather than the main objective.

It is easy (and, I argue, valid) to draw parallels to the development of the Internet, social media, and smartphones.

Massive companies are making these remarkable technological leaps, enabling us to do incredible things. However, many contend that the profit-motivated approach to technological development has very adverse side effects (increased polarization/radicalization, exacerbation of mental health issues, etc.).

LLMs will continue to become more advanced, but they will do so in ways that primarily benefit their shareholders. Be mindful, and perhaps even a little wary, of that.

Author 1: Coming back to the ideas of teachers, what should teachers be aware of that might be coming up based on these future developments?

Author 2: LLMs are not going away, and they should not be ignored. We need to understand that students have access to these tools and focus on teaching them to use them thoughtfully rather than as an "easy out." LLMs should encourage *more* thought, not less. Do not ban ChatGPT—start educating students on where it came from, how it works, what it is suitable for, and what it is not. Brainstorm ways to involve students in an iterative, back-and-forth approach to machine assistance that allows them to leverage themselves.

That said, the three factors anyone with a passing interest in genAI development needs to pay attention to are profit motivation, regulation, and economic impact. Ask yourself,

1. What will make FAANG (Facebook, Amazon, Apple, Netflix, and Google) the most money?
2. What bills will lawmakers pass limiting/incentivizing genAI development?
3. What does the job market look like in 10 years if genAI makes even 5% of the current workforce redundant?

Author 1: That is a lot of workers. Does that mean that the public just needs to assume that in the future, these jobs and bills created by the genAI tools will just have false information?

Author 2: To clarify a finer point, I mentioned earlier about LLMs telling you what to hear—this is the primary mechanic behind hallucinations (errors). I have had to explain to people that a model is not "lying" to you. It mimics human conversation, almost like a small child would. It encounters a situation, tries to match the context to something in its training data, and returns an answer that most closely mirrors what a human would respond to it. There are many potential points of failure: perhaps the human interacting with the model miscommunicated, the model misread the situation, the human trainer who rated the model was misinformed or made a rating mistake, etc. Considering all of this, it is astonishing how accurate and robust these tools are. I continue to find it some of the most impressive work that humanity has ever achieved. We must collectively recognize how much human effort goes into developing LLMs and how much human error is imparted in that process.

Author 1: Thank you so much! I truly appreciate your time. I have really learned a lot from this conversation.

Author 2: Happy to help!

5. Implications

The transcript described the opinion of one genAI trainer. Still, his insights were valuable to the educational community and provided specific implications for supporting our mathematics teachers in the future. From the findings, specific themes arose:

1. GenAI was trained by human feedback.
2. Errors need to be identified to refine the AI tool's response.
3. GenAI was designed to be a people-pleaser, so it will always provide a response, even if it does not have an answer.
4. Teachers need to know when to trust the tool and when to find their information from another source.
5. GenAI is not going away, so TEs need to address these concerns now.

From these findings, specific implications occur for the education community.

5.1. GenAI Reflects the Biases of Its Trainers

Since humans train the genAI, it inherits their biases. This explained why many researchers have found clear bias in the responses from genAI tools[4]. Since humans inherently favor specific methods for solving problems based on their educational backgrounds, genAI mirrors this tendency and may focus on a single method even when multiple solution strategies exist.

The data revealed that human error influences the genAI's responses. The failure of genAI models can stem from several sources, such as the user miscommunicating

their intent, the genAI misreading the situation, or the human trainer who rated the model being misinformed or making a rating error. Therefore, genAI can exhibit biases and failures due to its human implementation and training.

5.2. The Importance of Errors

The data indicated that errors and finding errors are central to training LLMs; thus, as TEs, we should continue to look for and find errors to support this development and assume that errors will occur, particularly in mathematics. When TEs ask genAI mathematical questions, the genAI will not think; instead, it will predict based on its training and the ratings created by past developers.

5.3. GenAI Needs Documented Problems

The data also indicated that genAI can handle simple, documented problems (e.g., solving equations and calculating probabilities). However, it struggles with multi-step reasoning and novel problems. However, genAI is optimized to be "helpful," which sometimes leads to incorrect but convincing answers. TEs can use this knowledge to create problems for their students and show the mathematical errors that this genAI can construct. TEs must focus on providing students with novel problems that support their mathematical learning, pushing them beyond what they already know.

5.4. TEs Pay Attention

The data also indicated that education needs to pay attention to and be aware of three significant factors in the creation of genAI: profit motivation, regulation, and economic impact. It is necessary and responsible to identify who profits or benefits from the use of these tools. Moreover, teachers are responsible for using these tools in their classrooms; therefore, they should analyze the benefits versus risks of using genAI tools and what they could provide to their students.

Teachers need to examine the regulations, or lack thereof, created by lawmakers to determine whether this tool should be used in the classroom. Teachers are responsible for the impact of their activities on their students, so they must conduct a risk-benefit analysis to determine whether the risks to the student population are minimal.

6. Limitations

This study, based on the perspective of a single generative artificial intelligence trainer, has several key limitations. Since the findings rely on a single person's experiences and interpretations, they cannot capture the full range of how genAI trainers operate across organizations, model architectures, and governance frameworks. This narrow focus risks overgeneralizing personal practices or beliefs as representative of the entire field. It also limits the ability to verify claims about how genAI systems are trained, as no other voices or comparisons are available to confirm or challenge the account. Additionally, insights into genAI's impact on education may reflect the trainer's own philosophical beliefs, disciplinary background, or level of technical knowledge rather than offering a balanced or evidence-based view. Finally, predictions about the future of artificial intelligence based on a single expert's perspective are inherently limited because they cannot encompass the full range of expertise spanning industry, policy, ethics, pedagogy, and research. As a result, while such a study can provide valuable insider insights, we are aware that conclusions and implications should be viewed as evidence of existence rather than definitive.

7. Conclusions

Overall, this investigation provided TEs with more information and warnings about genAI. TEs must consider the following factors: who is profiting, what regulations have been imposed on these tools, who is affected, and who could benefit from this information. GenAI is a tool, not a replacement for teachers, and should be used for brainstorming and content generation, rather than as a final answer source. Teach students to use genAI critically by checking facts, thinking deeply, and staying aware of biases. Thus, TE needs to empower teachers and students with genAI, rather than letting genAI think for them. In conclusion, we must be critical of the use of these tools in the future and remain informed to support our students to the best of our abilities.

Author Contributions

Conceptualization, A.G.S. and P.S.; methodology, A.G.S.; validation, A.G.S., D.M.C. and P.S.; formal analysis, A.G.S., D.M.C. and P.S.; investigation, A.G.S.; data cura-

tion, D.M.C.; writing—original draft preparation, A.G.S.; writing—review and editing, A.G.S., D.M.C. and P.S. All authors have read and agreed to the published version of the manuscript.

Funding

No funding was used for this study.

Institutional Review Board Statement

The study was approved by the Institutional Review Board of James Madison University (IRB-FY25-466) on February 17, 2025.

Informed Consent Statement

Informed consent was obtained from all subjects involved in the study.

Data Availability Statement

Data is available upon request from the corresponding author.

Conflicts of Interest

The authors declare no conflict of interest.

AI Use Statement

During the preparation of this manuscript, the authors used Grammarly solely for language refinement. The authors subsequently reviewed and edited the content as necessary and take full responsibility for the final content of the published article.

Appendix A. Duoethnography Interview Questions

1. Please describe your job and how it relates to LLMs.
2. Explain how LLMs are developed and trained.
3. How does this relate to genAI chatbot models, and is there a specific genAI tool you work with?
4. How do you view your role in this process?
5. How do you find errors?
6. What do you do when you notice the LLMs create an error?
7. What is the next step to "fix" these issues in the LLMs?
8. Can you give me an example of what happens in your work with LLMs?
9. What insight do you believe you could provide others that is not common knowledge to the public?
10. What do you wish people knew about LLMs?
11. What does this mean for educators or students who use LLMs?
12. How should educators or students view LLMs?
13. What is your perspective on the future of the LLMs market?
14. What should teachers be aware of that might be coming up based on these future developments?
15. Is there anything else you would like to share?

References

[1] Lyden, S., Trischuk, M., Millar, B., et al., 2025. Exploring a generative AI-assisted curriculum development approach. In Proceedings of the 36th Australasian Association for Engineering Education Annual Conference, Brisbane, Australia, 7–10 December 2025. Available from: https://aaee2025.org/microsite/pdf/full-paper_224.pdf

[2] Sawyer, A.G., 2024. Artificial intelligence chatbots as a mathematics curriculum developer: Discovering preservice teachers' overconfidence in ChatGPT. International Journal on Responsibility. 7(1), 1. DOI: https://doi.org/10.62365/2576-0955.1106

[3] Kim, Y.R., Park, M.S., Joung, E., 2025. Exploring AI integration in math education: Preservice teachers' experiences and reflections on problem-posing activities with ChatGPT. School Science and Mathematics. 126(1), 9–23. DOI: https://doi.org/10.1111/ssm.18336

[4] Wu, G., 2023. 8 big problems with OpenAI's ChatGPT. Available from: https://www.makeuseof.com/openai-chatgpt-biggest-probelms/ (cited 2 February 2025).

[5] Ding, L., Li, T., Jiang, S., et al., 2026. Students' perceptions of using ChatGPT in a physics class as a virtual tutor. International Journal of Educational Technology in Higher Education. 20, 63. DOI: https://doi.org/10.1186/s41239-023-00434-1

[6] Preiksaitis, C., Rose, C., 2026. Opportunities, challenges, and future directions of generative artificial intelligence in medical education: Scoping review. JMIR Medical Education. 9, e48785. DOI: https://doi.org/10.2196/48785

[7] Cruz, D.G., Frasso, S., Bedner, S., et al., 2024. Chat-

ting with ChatGPT: Artificial intelligence assisting with interdisciplinary mathematics lessons. The Banneker Banner. 36(2), 3–11. DOI: https://doi.org/10.65517/24V36N203

[8] Sharma, N., 2025. Generative AI, conversion AI, and chatbot—A breakdown. Available from: https://www.nimbleappgenie.com/blogs/generative-ai-vs-conversational-ai-vs-chatbot/ (cited 2 February 2025).

[9] Turing, A.M., 1950. Computing machinery and intelligence. Mind. 49(236), 433–460. DOI: https://doi.org/10.1093/mind/LIX.236.433

[10] International Business Machines Corporation, 2024. What are large language models (LLMs)? Available from: https://www.ibm.com/topics/large-language-models (cited 2 February 2025).

[11] U.S. Department of Education, Office of Educational Technology, 2023. Artificial Intelligence and the Future of Teaching and Learning: Insights and Recommendations. U.S. Department of Education, Office of Educational Technology: Washington, DC, USA. Available from: https://www.ed.gov/sites/ed/files/documents/ai-report/ai-report.pdf

[12] National Council of Teachers of Mathematics, 2024. Artificial intelligence and mathematics teaching. Available from: https://www.nctm.org/standards-and-positions/Position-Statements/Artificial-Intelligence-and-Mathematics-Teaching/ (cited 2 February 2025).

[13] Murphy, R.F., 2019. Artificial Intelligence Applications to Support K–12 Teachers and Teaching: A Review of Promising Applications, Challenges, and Risks. RAND Corporation: Santa Monica, CA, USA.

[14] Pinheiro Privette, A., 2024. AI's challenging waters. Available from: https://cee.illinois.edu/news/AIs-Challenging-Waters (cited 2 February 2025).

[15] Baum, J., Villasenor, J., 2023. The politics of AI: ChatGPT and political bias. Available from: https://www.brookings.edu/articles/the-politics-of-ai-chatgpt-and-political-bias (cited 2 February 2025).

[16] Pfeiffer, D., 2024. Five growing concerns about generative AI for librarians and information professionals. Available from: https://www.choice360.org/libtech-insight/four-growing-concerns-about-generative-ai-for-librarians-and-information-professionals/ (cited 2 February 2025).

[17] Sawyer, A.G., Aga, Z.G., 2024. Counterexamples to Demonstrate Artificial Intelligence Chatbots' Lack of Knowledge in the Mathematics Education Classrooms. Available from: https://amte.net/sites/amte.net/files/Connections%20%28Sawyer%29.pdf (cited 3 May 2025).

[18] Golda, A., Mekonen, K., Pandey, A., et al., 2024. Privacy and security concerns in generative AI: A comprehensive survey. IEEE Access. 12, 48126–48144. DOI: https://doi.org/10.1109/ACCESS.2024.3381611

[19] James Madison University, 2025. Executive order bans the use of Deepseek. Available from: https://www.jmu.edu/news/computing/2025/02-12-deepseek-executive-order.shtml (cited 13 December 2025).

[20] Krutka, D.G., Heath, M.K., Willet, K.B.S., 2019. Foregrounding technoethics: Toward critical perspectives in technology and teacher education. Journal of Technology and Teacher Education. 27(4), 555–574. DOI: https://doi.org/10.70725/560948mgghtd

[21] Norris, J., Sawyer, R.D., Lund, D., 2012. Duoethnography: Dialogic Methods for Social, Health, and Educational Research. Left Coast Press: Walnut Creek, CA, USA.

[22] Rapke, T.K., 2014. Duoethnography: A new research methodology for mathematics education. Canadian Journal of Science, Mathematics and Technology Education. 14(2), 172–186. DOI: https://doi.org/10.1080/14926156.2014.903317

[23] Latz, A.O., Murray, J.L., 2012. A duoethnography on duoethnography: More than a book review. The Qualitative Report. 17(36), 1–8. DOI: https://doi.org/10.46743/2160-3715/2012.1736

[24] Norris, J., Wilson, T., Oberg, A., 2008. Duoethnography. In: Given, L.M. (Ed.). The SAGE Encyclopedia of Qualitative Research Methods. SAGE Publications: Thousand Oaks, CA, USA. pp. 233–236.

[25] Breault, R.A., 2016. Emerging issues in duoethnography. International Journal of Qualitative Studies in Education. 29(6), 777–794. DOI: https://doi.org/10.1080/09518398.2016.1162866