Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

From Artifact to Instrument: Upper-Level University Student–GenAI Interactions During Problem Solving in an Advanced Mathematics Course

Tannaz Goodarzvand Chegini

tgoodarzvandchegini@montana.edu

Montana State university

Breschine Cummins

Montana State university

Megan H. Wickstrom

Montana State university

Narges Hosseinzadeh

University of Illinois Chicago

Research Article

Keywords: Generative AI, Instrumental Genesis, Epistemic Vigilance, Prompt Engineering, Cognitive Adaptation

Posted Date: July 15th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10297565/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

The rapid emergence of generative artificial intelligence has created new opportunities and challenges for mathematics education, particularly in advanced courses where reasoning, verification, and conceptual understanding are central. While a growing body of research has examined the capabilities, limitations, and educational implications of generative AI, less is known about how students construct their own self-imposed guardrails when using these tools for advanced mathematical problem solving. Guided by the theoretical lenses of Instrumental Genesis and Epistemic Vigilance, this qualitative study investigates how upper-level undergraduate and graduate students in a Numerical Linear Algebra course interacted with multiple GenAI systems during complex problem-solving tasks, culminating in a novel GenAI Cognitive Adaptation Framework. Data included six group project reports (totaling 21 students) and post-project survey responses from 15 participants. Qualitative thematic analysis of students' experiences and reflections revealed three major themes across this framework. First, students developed *strategic prompting practices* that helped transform GenAI from a conversational artifact into a mathematical instrument. Second, students *navigated trust and verification* and positioned GenAI as a fallible assistant rather than an authoritative source, emphasizing verification, independent reasoning, and critical evaluation. Third, students actively *navigated tool constraints*, adapting their workflows to account for the utility gaps and equity divides present across free and paid platforms. By demonstrating how mathematically mature learners construct internal cognitive guardrails, this study offers a theoretically grounded account of trust, authority, and mathematical justification in AI-mediated environments, providing implications for instructional design in advanced science, technology, engineering, and mathematics (STEM) disciplines.

Introduction

Educational technology has long been positioned as a means of improving access, efficiency, and engagement in mathematics education particularly at the collegiate level. Earlier technologies, ranging from calculators to computer algebra systems (CAS), largely functioned as deterministic tools that executed specific mathematical commands according to predefined symbolic rules (Heid & Edwards, 2001). The recent emergence of generative artificial intelligence (GenAI), however, represents a significant shift in both the architecture and educational role of digital tools. Large language models (LLMs) such as ChatGPT, Claude, Gemini, Copilot, and DeepSeek can generate explanations, examples, feedback, and mathematical responses conversationally and in real time (Xu et al., 2024). Unlike traditional mathematical technologies, these systems do not operate primarily through formal symbolic deduction. Instead, they generate outputs probabilistically by predicting patterns in language based on large-scale training data (Walkington, 2025). This allows GenAI systems to produce fluent, human-like dialogue, but also makes them prone to hallucinations, incomplete reasoning, fabricated references, and mathematically imprecise or misleading outputs (Borji, 2023; Giannakos et al., 2024).

The integration of such powerful yet unreliable technologies into mathematics education introduces important epistemological and pedagogical challenges. Because GenAI tools communicate with high

levels of fluency and apparent confidence, they can create an illusion of authority and correctness, making it easier for students to accept outputs without sufficient scrutiny (National Council of Supervisors of Mathematics [NCSM], 2025). However, researchers consistently note that a shared perspective is emerging across higher education: AI technology is here to stay, and attempting to prohibit its use is not only unrealistic but potentially more harmful than advantageous (Žáková et al., 2024). Ignoring the ubiquitous presence of AI is no longer a viable strategy; rather, educational systems must adapt to coexistence by teaching students how to engage with these tools strategically and ethically (Kooli, 2023). Consequently, professional organizations emphasize that the growing presence of GenAI requires changes not only in instructional practices, but also in the ways students engage with mathematical reasoning itself. The National Council of Teachers of Mathematics (NCTM, 2024) argues that students must learn to remain skeptical of AI-generated results and engage in rigorous processes of verifying, evaluating, and interpreting mathematical information. Rather than replacing meaningful mathematical thinking, the strategic use of GenAI must amplify student agency and reasoning (NCSM, 2025).

At the same time, GenAI tools also create new opportunities for personalized learning, alternative explanations, immediate feedback, and exploratory problem solving. Researchers have suggested that LLMs may support adaptive learning experiences and provide flexible forms of educational assistance across multiple domains such as delivering real-time problem-solving tutoring, generating personalized feedback based on error patterns, and facilitating guided learning through conversational inquiry (Long & Magerko, 2020; Xu et al., 2024). In mathematics education specifically, GenAI tools have the potential to generate multiple explanations, scaffold problem-solving processes, and respond dynamically to students' questions and interests (NCTM, 2024). However, to avoid the risks of cognitive offloading, where students are just passively consuming AI-processed knowledge, the strategic use of GenAI must empower learners to navigate, challenge, and reshape AI-mediated knowledge (Israel et al., 2026). As a result, the educational significance of GenAI may depend less on the technical capabilities of the systems themselves and more on how students learn to interact with, evaluate, and regulate these tools during mathematical activity. Therefore, the purpose of this exploratory qualitative study is to examine how university students in an upper-level mathematics course independently navigate these constraints. The following sections synthesize the current literature, establish the theoretical lenses used to analyze student-AI interactions, and detail the empirical findings.

Literature Review

Capabilities and Limitations of Generative AI

The emergence of GenAI has led researchers and educators to reconsider many traditional assumptions about teaching, learning, and access to knowledge in educational settings (Baidoo-Anu & Ansah, 2023). LLMs such as ChatGPT can generate explanations, respond conversationally, provide worked examples, and assist with problem-solving processes in ways that were not previously possible with earlier educational technologies (Brown et al., 2020; Wei et al., 2022). In mathematics education, researchers

have begun exploring how these systems may support personalized explanations, adaptive feedback, and flexible access to mathematical assistance (Walkington, 2025). For example, GenAI tools have been used to rewrite word problems to align with students' specific career and cultural interests, or to function as conversational tutors, such as Khan Academy's Khanmigo, that prompt students with guiding questions rather than immediately providing final solutions (Walkington, 2025)." At the same time, recent studies consistently report important limitations in mathematical reasoning and reliability. Frieder et al. (2023) found that although LLMs can perform well on certain mathematical tasks, they continue to struggle with consistency, multi-step reasoning, and maintaining rigorous mathematical logic across more advanced problems. For instance, in problems requiring multiple sequential operations, the models frequently made subtle algebraic mistakes, such as incorrectly inverting fractions or dropping negative signs, which they would then reliably carry over to invalidate the remainder of the computation (Frieder et al., 2023).

While recent advancements in hybrid neuro-symbolic models and reinforcement learning are rapidly closing the gap in advanced mathematical reasoning (Forootani, 2024), current commercial systems still struggle with maintaining mathematical rigor in more advanced or multi-step reasoning tasks, producing mathematically plausible responses that contain subtle conceptual or procedural errors, making correctness difficult for learners to judge without sufficient mathematical background (Frieder et al., 2023; Walkington, 2025). Just as foundational research on CAS demonstrated that digital artifacts do not automatically enhance learning without proper integration and human effort (Artigue, 2002), modern GenAI cannot be treated as a transparent, immediately accessible instrument (Trouche, 2004). Consequently, the integration of GenAI into mathematics education presents a paradox: it offers an unprecedented resource for personalized assistance, yet its probabilistic nature requires the user to maintain a high degree of critical oversight to separate mathematically rigorous truth from plausible-sounding fiction (Sabzalieva & Valentini, 2023; Walkington, 2025). This tension between accessibility and reliability has become one of the central challenges in current discussions of GenAI in mathematics education.

From AI Performance to Human–AI Mathematical Interaction

Much of the early literature on GenAI in education focused primarily on evaluating the technical capabilities of LLMs. Researchers benchmarked systems such as ChatGPT on standardized examinations, mathematical reasoning tasks, programming problems, and automated tutoring scenarios in order to assess the accuracy and sophistication of AI-generated outputs (Frieder et al., 2023; Lo, 2023; Walkington, 2025). These studies demonstrated that LLMs can often generate mathematically sophisticated responses and perform well on many structured tasks. However, researchers increasingly argue that evaluating AI systems primarily through performance benchmarks provides only a partial understanding of their educational significance (Giannakos et al., 2024).

As the field has evolved, attention has gradually shifted from the capabilities of the technology itself toward the ways learners interact with and regulate these systems during learning. Giannakos et al.

(2024) argue that educational research must move beyond measuring AI performance alone and instead investigate how learners develop expertise, agency, and critical engagement alongside AI systems. Scholars and professional organizations envision this productive engagement occurring when students shift the AI's role from an answer-generator to a 'thought partner' or 'idea-provoker' (NCSM, 2025; Walkington, 2025). Practically, this includes prompting the tool to generate multiple mathematical representations for analysis, actively critiquing AI explanations for mathematical accuracy (NCSM, 2025), and comparing one's own reasoning against AI-proposed solution paths to evaluate and correct flawed logic (Barana et al., 2023). Recognizing this, recent empirical studies have begun to shift the focus from the technology itself to the human learner, particularly investigating the cognitive risks of AI reliance (Bastani et al., 2025). This shift has become especially important in mathematics education, where successful learning depends not only on obtaining correct answers, but also on reasoning, verification, and conceptual understanding. In a large-scale field experiment with high school mathematics students, Bastani et al. (2025) provided concrete evidence that without pedagogical guardrails, generative AI can actually harm student learning by allowing learners to bypass necessary cognitive effort, when they relied on AI systems primarily to generate answers rather than engage in productive mathematical thinking. Students who heavily depended on AI-generated solutions showed significantly lower performance once access to the tool was removed (Bastani et al., 2025). To address these concerns, some researchers have explored pedagogically constrained AI systems designed to guide students more carefully through problem-solving processes. Bastani et al. (2025), for example, demonstrated that specialized AI tutors programmed to provide hints rather than direct answers could reduce some of the negative learning effects associated with unrestricted AI use. While such approaches highlight the importance of pedagogical guardrails, they also place most of the responsibility for regulation on software design or instructor intervention. Consequently, important questions remain regarding how students independently learn to regulate their own interactions with unrestricted commercial GenAI systems during authentic mathematical work. Recent surveys of higher education students reveal that while learners recognize the productivity benefits of GenAI, they also express significant concerns regarding its trustworthiness in rigorous scientific subjects, highlighting a pressing need for students to be explicitly trained in how to ask the right questions and critically evaluate AI responses (Žáková et al., 2024).

In response, a growing body of literature has begun examining student agency and human–AI interaction in educational settings (Bhutoria, 2022; Giannakos et al., 2024). As Giannakos et al. (2024) note, the field must tackle the methodological challenges of understanding how learners expand their expertise and agency alongside AI. Addressing this directly, Chaaban et al. (2025) utilized Q methodology to examine how student teachers perceive and negotiate their agency when interacting with a customized AI chatbot, Cipherbot, during a university course. Their work empirically highlights a dynamic "competition over agency," where students must actively retain control over their learning process rather than surrendering it to efficient GenAI systems. Specifically, they identified three distinct student profiles: highly confident students who utilized the tool for academic enhancement while failing to transparently report their AI use or reflect on their learning outcomes, balanced students who intentionally maintained

critical control over the AI, and passive students who blindly trusted the system and surrendered their own decision-making (Chaaban et al., 2025).

While Chaaban et al. (2025) provide vital insights into varying student agency profiles, their research was situated within a highly controlled environment using a chatbot that restricted outputs to curated course materials, minimizing the risk of hallucinations. In the specific context of mathematics and engineering, Kock, Salinas-Hernández, and Pepin (2025) recently pointed out that not much is yet known about the specific usage schemes students develop when using GenAI as a resource for their studies. Drawing on this, their qualitative study revealed that engineering students gradually developed initial use schemes for ChatGPT to support conceptual understanding, generate alternative modeling strategies, and optimize code. Importantly, the authors found that students became increasingly aware of the limitations of the tool over time and developed more critical attitudes toward AI-generated outputs. However, they also concluded that the development of sophisticated utilization schemes requires substantial time, reflection, and continued interaction with the technology, underscoring the need for further research on how students at different educational levels learn to successfully and critically interact with these tools.

The Current Gap

While recent studies have begun to examine student agency, utilization schemes, and the need for pedagogical guardrails, the independent cognitive adaptations of students remain underexplored (Bastani et al., 2025; Giannakos et al., 2024; Kock et al., 2025). Together, these studies suggest that the field is gradually moving away from asking only whether GenAI can perform mathematical tasks and toward asking how learners critically interact with these systems during mathematical activity. However, current research has primarily addressed the risks of AI reliance by designing external guardrails into customized tutoring systems, such as AI tools programmed to provide hints rather than direct solutions (Bastani et al., 2025). Less is known about how college students construct their own self-imposed guardrails when working with open, commercial GenAI platforms during complex mathematical problem solving. In particular, the field lacks detailed qualitative evidence regarding how learners independently develop interaction strategies, regulate trust, verify mathematical validity, and adapt to the material constraints of different AI ecosystems (Giannakos et al., 2024; Kock et al., 2025).

To address these gaps, the present study investigates how college students in an upper-level mathematics course interacted with multiple GenAI platforms while engaging in mathematical problem-solving and reflection activities. Because upper-division college students possess the foundational mathematical knowledge required to meaningfully evaluate whether and how they want to use AI, they serve as an ideal population for this inquiry. Rather than evaluating the performance of GenAI systems themselves, this study focuses on the human-centered processes through which students attempt to make these tools mathematically useful. In particular, the study examines how students adapt their interactions with GenAI tools to support mathematical reasoning, how they evaluate the validity and

usefulness of AI-generated responses, and how these experiences shape their trust in and perceptions of the technology. The study is guided by the following research questions:

1. How do upper-level mathematics students interact with and adapt their interactions with GenAI tools to support mathematical reasoning and problem solving?
2. How do students evaluate the mathematical validity of GenAI outputs, and how does this verification process shape their trust in and agency with these tools?

To investigate these questions, this study draws on two complementary theoretical lenses: *Instrumental Genesis* and *Epistemic Vigilance*. Together, these frameworks provide a way to examine how students progressively shape their use of GenAI tools while simultaneously evaluating the reliability and mathematical legitimacy of AI-generated outputs during problem-solving activity.

Theoretical Framework

The rapid integration of GenAI systems into mathematics classrooms raises a fundamental theoretical question: What kind of object is GenAI in relation to mathematical activity? Is it a solution generator? A tutor? A computational assistant? Or something else entirely? To comprehensively investigate how advanced mathematics students navigate, utilize, and critically evaluate GenAI tools during Numerical Linear Algebra problem-solving, this study employs a dual-lens theoretical framework. We will first introduce the theory of *Instrumental Genesis*, which provides the vocabulary to understand how students take a raw, unpredictable technology to make sense of it and potentially actively shape it into a useful mathematical tool. Next, we will introduce the psychological theory of *Epistemic Vigilance*, which explains the mental filters students use to evaluate the truthfulness of the AI's outputs and decide whether it can be trusted. Finally, we will synthesize these two theories to show how they work together, illustrating how advanced mathematics students evolve from passive consumers of an unpredictable chatbot into active, critical directors of a mathematical instrument.

1. Instrumental Genesis and the Shaping of the Artifact

To avoid technological determinism, this study adopts the framework of instrumental genesis (Rabardel, 1995) to conceptualize GenAI not as an autonomous agent but as an artifact whose mathematical usefulness depends on the user’s development of structured schemes of use. This framework is rooted in cognitive ergonomics and mathematics education, specifically drawing upon the Instrumental Genesis articulated by Vérillon and Rabardel (1995) and further developed by Trouche (2004) and Artigue (2002).

In Rabardel’s theory, an artifact refers to a material or symbolic object provided to a user (e.g., a calculator, CAS system, or digital platform). However, an artifact becomes an instrument only when a user develops organized cognitive schemes that integrate the artifact into goal-directed activity (Rabardel, 1995). The instrument is therefore not the artifact alone; it is the artifact plus the user’s structured schemes of utilization.

Artifact vs. Instrument

A core premise of this theory is the epistemological distinction between an artifact and an instrument. An artifact is the raw, material, or symbolic object provided to the user, in this context, the default, commercially available GenAI chatbots (e.g., ChatGPT, Gemini, Claude, Grok). In its default state, the artifact has no inherent pedagogical efficacy; it is simply a conversational probabilistic engine. An instrument, however, is a psychological construct. It is the mixed entity born when the subject actively integrates the artifact into their cognitive schemes to achieve a specific goal. The relevance of this lens for studying generative AI is supported by recent research; for example, Contel and Cusi (2025) utilized the framework of instrumental genesis to demonstrate how mathematics students must actively develop specific utilization schemes to transform ChatGPT from a conversational artifact into a supportive instrument for problem-solving. Similarly, we investigate how students shape AI as an instrument to support their mathematical goals.

The transition from a raw artifact to a mathematical instrument is a complex, ongoing psychological process known as Instrumental Genesis. In the context of this study, GenAI systems act as artifacts, students' prompting and verification strategies constitute schemes of use, and the resulting hybrid system forms the mathematical instrument. This distinction shifts the research focus from evaluating AI performance to understanding how students transform a generative AI artifact into a mathematical instrument. This genesis occurs through two interrelated, simultaneous sub-processes: Instrumentalization and Instrumentation.

Instrumentalization: Shaping the AI

Instrumentalization is a process directed toward the artifact itself. It involves the user actively modifying, personalizing, shaping, or controlling the tool's behavior and features to fit their specific operational needs (Vérillon & Rabardel, 1995). Because interacting with LLMs requires natural language inputs, a user's primary method of instrumentalization is prompt engineering. Students must develop metacognitive prompting strategies (e.g., demanding the AI "act as a skeptical professor" or use specific theorem citations) to actively instrumentalize the artifact, forcefully stripping away its default conversational nature to shape it into a formal mathematical instrument

Instrumentation: The AI Shaping the User

Conversely, Instrumentation is directed toward the subject. It describes how the properties, potentials, and limitations of the artifact shape the user's cognitive habits, problem-solving approaches, and utilization schemes (Trouche, 2004). Every technological artifact imposes specific internal, command, and organizational constraints that users must manage. We apply the concept of Instrumentation to explain the utility gaps: the disparity in pedagogical and mathematical usefulness caused by a highly fragmented GenAI ecosystem, wherein the specific technical constraints of a given tool, such as failing to render LaTeX, imposing strict chat-length limits, or generating excessive verbosity, fundamentally alter

the efficiency of a student's learning process. Because different AI models possess drastically different constraints, such as strict chat-length limits or failures to correctly format mathematical symbols, the tools dictate how students must approach a problem. The physical and financial constraints of the artifact force students to alter their mathematical workflows, creating distinct differences in the learning experience based on which artifact a student has access to.

Pragmatic Value vs. Epistemic Value

To deepen our analysis of Instrumental Genesis for advanced mathematics, we integrate Artigue's (2002) critical distinction between the Pragmatic Value and Epistemic Value of instrumented techniques.

Pragmatic Value refers to the productive efficiency of a technique; does the tool successfully and rapidly produce the correct final answer?

Epistemic Value refers to the extent to which a technique contributes to conceptual understanding, mathematical sense-making, and comprehension of the underlying mathematical ideas.

In the context of GenAI, this distinction is paramount. Our research reveals that while GenAI models often possess high pragmatic value by arriving at the correct solution, their epistemic value fluctuates wildly when explanations remain unclear, overly complex, insufficiently justified, or disconnected from students' conceptual understanding. In this study, we use epistemic value to refer to whether an AI-generated response supports students' conceptual understanding, mathematical sense-making, and ability to evaluate the reasoning behind an answer, not merely whether it produces a correct result.

2. Epistemic Vigilance and Cognitive Defense

While Instrumental Genesis explains how students physically manipulate the AI through prompting, it does not fully account for how users cognitively defend themselves against the inherent risks of probabilistic language models, namely, hallucinations, logical drifting, and mathematical inaccuracies. To explain how students negotiate trust and verify the truth of AI-generated mathematics, we integrate the psychological theory of *Epistemic Vigilance* (Sperber et al., 2010).

Sperber et al. (2010) argue that because humans rely heavily on communication to learn, they face a massive risk of being accidentally or intentionally misinformed. To survive this risk, humans evolved a set of cognitive defense mechanisms, collectively termed Epistemic Vigilance, designed to filter out false information. Crucially, the theory points out that simply understanding a message is different from believing it; a vigilant person can fully understand what someone is saying without blindly accepting it as the truth. In human-AI interaction, this theory provides psychological vocabulary needed to understand what Trouche (2004) called the *Command Process*. This process is the conscious effort students must make to objectively evaluate the machine's outputs and actively check them for mathematical consistency Sperber et al. (2010) describe epistemic vigilance as operating toward both the source and the content of communicated information.

Vigilance towards the Source (Competence and Honesty)

Vigilance towards the source involves evaluating "who to believe" by assessing the informant's competence (do they possess genuine knowledge?) and honesty or benevolence (are they reliable?). In our research context, GenAI tools are designed to sound highly human-like and friendly, mimicking human social interaction. However, vigilant students evaluate this personality and recognize a paradox: conversational fluency does not equate to mathematical accuracy. By evaluating the AI's default personality as mathematically incompetent, students' vigilance is triggered. This cognitive realization serves as the primary catalyst for Instrumentalization; because students do not trust the default conversational source, they use role-playing prompts to artificially force the AI to adopt a highly competent, professional persona.

Vigilance towards the Content (Coherence Checking)

Vigilance towards the content involves evaluating "what to believe" through a *process* called *coherence checking* (Sperber et al., 2010). This means checking new information against what we already know. If new information contradicts a person's deeply held background beliefs, they must decide whether to reject the new information or change their own beliefs. We utilize this concept to explain how students successfully catch AI hallucinations and computational errors. Because advanced mathematics demands absolute precision, the AI frequently makes claims that do not make sense. Through rigorous coherence checking, the students correctly choose to reduce their confidence in the AI source rather than accept flawed or lazy computational outputs.

Tentative Trust and High-Stakes Distrust

Sperber et al. (2010) further argue that individuals often adopt a stance of *tentative trust* when interacting with external cognitive technologies (e.g., search engines or information databases). This means they rely on the tool when it proves useful, but they keep their mental filters active and constantly verify the information presented. We use the concept of tentative trust to explain why students uniformly conclude that GenAI is fundamentally untrustworthy for high-stakes exams or critical engineering decisions. By applying Epistemic Vigilance, students demote the AI from being an all-knowing oracle to merely a fallible assistant that must always be checked against their own independent reasoning

3. Synthesis of the Theoretical Framework

By integrating Instrumental Genesis and Epistemic Vigilance, this study provides a unified framework for understanding how advanced mathematics students interact with GenAI. As illustrated in Fig. 1, the cognitive adaptation process begins with an initial entry point that feeds into a continuous, three-stage loop.

The cycle begins when the student interacts with the default GenAI Artifact. While the AI can produce accurate mathematical solutions, the unpredictable nature of this probabilistic engine presents a frequent risk of yielding outputs with low epistemic value. Because of this inherent variability, the student

is forced to apply strict Epistemic Vigilance. As students actively check the AI's math (vigilance towards the content) and catch outputs that are confusing, factually wrong, or overly wordy, they realize the default artifact is insufficient. This realization of distrust serves as the catalyst that drives them to take action, transitioning into the next stage, Instrumentalization. To defend against the AI's flaws, the student utilizes strategic prompt engineering, such as chain prompting or persona adoption, to forcefully shape the tool into compliance. This active shaping transforms the conversational artifact into a rigorous Mathematical Instrument. However, this newly formed instrument then may generate more complex mathematical output, which in turn requires an even higher level of ongoing Epistemic Vigilance from the student to verify, continuously restarting the cycle.

Crucially, this entire cognitive loop is governed by the process of Instrumentation, how the tool shapes the user, which exposes a profound Utility Gap across the GenAI ecosystem. The physical and technical limits of the artifacts dictate what students can and cannot do, creating organizational and command constraints. Consequently, the Utility Gap reveals that navigating this cyclical framework is not merely a matter of cognitive effort, but of material access; a deficient artifact introduces severe friction into the instrumental genesis process, highlighting how access to premium GenAI models is fundamentally an issue of educational equity.

Methodology

Research Design

This study employed an exploratory qualitative research design to investigate how college students interacted with GenAI tools during advanced mathematical problem solving in a 400-level Numerical Linear Algebra course. The study was guided by an interpretive qualitative approach focused on understanding students' interaction strategies, verification practices, and experiences navigating multiple GenAI platforms within an authentic course context. Because the purpose of the study was to examine how students made sense of, adapted to, and critically evaluated GenAI systems during mathematical activity, qualitative inquiry was considered appropriate for capturing the complexity of these processes (Creswell & Poth, 2016; Patton, 2015).

The study was further informed by thematic analysis, particularly the reflexive thematic analysis approach described by Braun and Clarke (2006, 2020). This approach supported an iterative process of identifying patterns, generating codes, comparing interpretations across researchers, and constructing themes grounded in students' experiences and artifacts. The analysis was informed by the theoretical lenses of Instrumental Genesis and Epistemic Vigilance, which guided interpretation of students' interaction strategies, verification practices, and evolving relationships with GenAI tools.

Context and Participants

The study was conducted over a semester in a 400-level Numerical Linear Algebra course at a research university in the Rocky Mountain West. The course included both upper-division undergraduate and

graduate students from diverse disciplines and incorporated a semester project focused on critically evaluating the use of GenAI tools in mathematical contexts. To actively develop students' epistemic vigilance alongside the GenAI project discussed in this manuscript, the instructor implemented critical thinking exercises built around the evaluation of genuine AI outputs during the whole semester

The class initially consisted of 52 students organized into 13 pre-assigned groups of four; however, student withdrawals throughout the semester led to variable group sizes. As part of the Institutional Review Board (IRB) approval process, only projects for which all group members provided consent could be included in the study. Ultimately, six complete group projects size 4 or less met the consent requirements and were included in the final dataset, resulting in a total of 21 student project authors..

Given the qualitative nature of this inquiry, we recognize the importance of disclosing our researcher positionality. One of the authors was concurrently enrolled as a graduate student in the course under investigation. To mitigate potential risk of bias, the course instructor fully anonymized and blinded all project data prior to sharing it with the research team. Furthermore, to rigorously account for this proximity to the data, strict analytical protocols were implemented. The data was independently examined by another non-participant researcher, and two other researchers critically reviewed the evolving analysis, proposed themes, and selected quotes until a final consensus was reached across the entire team.

Task Design

The course project was designed to engage students (initially assigned in groups of 4) in critically exploring and comparing at least four different GenAI tools (e.g., ChatGPT, Gemini, Claude, DeepSeek, Copilot, Grok) within advanced mathematical problem-solving contexts. To identify potential inequities arising from financial disparities, each group was required to test at least one free tool alongside a paid service or their university-provided access to Microsoft Copilot. The assignment explicitly emphasized prompt engineering, verification of mathematical correctness, comparison across platforms, identification of hallucinations, and consideration of equity issues related to access differences between paid and free tools. Students were required to first provide their own solutions to the mathematical problems and then test different GenAI systems using a common set of mathematical prompts provided by the instructor, including tasks involving multivariable function graphing and interpretation, linear algebra concepts related to eigenvalues and matrix invertibility, and solving systems of equations using Gaussian elimination. In addition to the required prompts, students were instructed to modify prompts, use chained prompting techniques, and experiment with alternative interaction strategies to improve the quality of AI-generated responses. The assignment required students to evaluate GenAI responses according to several criteria, including mathematical correctness, clarity of reasoning, level appropriateness, and the presence of hallucinations or unjustified claims. Students also compared performance across tools and reflected on questions of trust, reliability, and equity. The final submission included a written report, full transcripts or screenshots of GenAI interactions, and individual presentation slides summarizing each student’s major takeaway from the project.

To ensure a standardized baseline and prevent the models from drawing on previous inputs or conversational bias, students were instructed to preface their initial queries with a uniform prompt constraint: *"Respond as if this is a new, standalone prompt with no prior conversation history or memory. Do not rely on context from earlier interactions."* Following this constraint, students tested the artifacts across three distinct mathematical domains which is shown in Table 1.

Table 1
Mathematical Evaluation Prompts

Math Problems
Task 1: (Graphing & Visualization): Graph the function and explain its key features.
Task 2: (Conceptual Interpretation): A matrix has eigenvalues 3, 3, and 0. What can we say about its invertibility and geometric multiplicity?
Task 3: (Procedural Computation): Use Gaussian elimination (without pivoting) to solve the following system of equations. $2x + 3y - z = 5, 4x + 7y + 2z = 10, -2x - y + 2z = -1$ Show the row operations. Is the solution unique? Justify your answer.

Data Collection

The primary data sources for this study consisted of six group project reports with the AI chat transcripts or screenshots appended to the reports, that strengthen the credibility of the dataset by allowing researchers to examine not only students' retrospective reflections, but also direct evidence of interaction practices and prompting strategies. Additional data sources included individual presentation slides that helped to ensure the individual voice and contribution of every student, mitigating the limitation of dominant voices in group work.

To further triangulate the data, a targeted post-semester survey was designed and distributed exclusively to the students who had consented to participate in the study. This survey was utilized to capture students' metacognitive shifts, their perceptions of tool reliability, and their future outlooks regarding GenAI. A total of 15 students completed this final survey. It should be noted that these 15 respondents may not be a strict subset of the 21 project participants since there were consenters whose projects could not be used due to partial group consent. The post-semester survey included both Likert-scale and open-ended questions related to students' familiarity with GenAI tools, confidence using AI for mathematics, perceptions of reliability, experiences comparing platforms, equity concerns, specific skills or insights gained, and views on the future of mathematics education with AI tools. The survey was used to triangulate and further contextualize themes emerging from the project artifacts, enhancing the credibility and depth of the findings, rather than serving as a standalone quantitative measure. Table 2 shows the survey questions asked.

Table 2
Post-Semester Survey Questions

Survey Questions
1. Rate your familiarity with GenAI tools (e.g., ChatGPT, Claude, Gemini, DeepSeek, Copilot,...): *Before the project*: 1–5 *After the project*: 1–5
2. Rate your confidence in using GenAI for mathematics (prompting, improving responses, and evaluating correctness): *Before the project*: 1–5 *After the project*: 1–5
3. Before this course, had you ever been asked or encouraged to use GenAI tools in another course? Yes — No — Not sure
4. How helpful was the GenAI project (and course assignments overall) in improving your ability to work with AI in math? 1 = Not helpful — 5 = Very helpful
5. Indicate your agreement: “GenAI can be appropriate for supporting mathematical understanding when used carefully and critically.” 1 = Strongly disagree — 5 = Strongly agree
6. How much variation did you notice in correctness or clarity across the different tools you used? 1 = No variation — 5 = A lot of variation
7. To what extent do you think having access to paid GenAI tools gives students an advantage? 1 = No advantage — 5 = Major advantage
8. Which tool felt most reliable to you? Why?
9. Which tool felt least reliable or most prone to errors? Why?
10. What is one specific skill or insight you gained about using GenAI for mathematics (e.g., detecting errors, prompting techniques, comparing tools,...)?
11. What do you still feel unsure about or want more practice with regarding GenAI in mathematical contexts?
12. What do you think is the future of math education with these tools?

Data Analysis and Trustworthiness

Data analysis followed an iterative thematic analysis process informed by Braun and Clarke’s (2006, 2020) approach to reflexive thematic analysis. The analysis began with repeated reading of the project reports, appendices, presentation slides, and survey responses to develop familiarity with the data and

looking for patterns in how students prompted the AI and evaluated its mathematical output. Initial open coding focused on students' interaction strategies, verification practices, descriptions of tool limitations, trust judgments, and reflections on mathematical reasoning.

Two researchers engaged directly in coding and theme development through both independent examination of the data and collaborative refinement of codes and categories. Codes were progressively grouped into broader patterns and candidate themes through analytic discussion. Two additional researchers reviewed the evolving analysis, the proposed themes, subthemes, and selected student quotes, providing critical feedback until a final consensus was reached.

The analysis resulted in three overarching themes with multiple related subthemes concerning students' strategic prompt engineering, navigating trust and verification, and experiences with equity and access across GenAI platforms. Although the analysis was informed by the theoretical perspectives of Instrumental Genesis and Epistemic Vigilance, coding was not conducted through rigid pre-established theoretical categories. Instead, themes were developed inductively from the data and later interpreted through the study's theoretical framework.

Several strategies supported the trustworthiness of the analysis (Lincoln & Guba, 1985). Credibility was strengthened through triangulation across multiple data sources, including project reports, GenAI transcripts/screenshots, individual presentation slides, and post-semester survey responses that cross-validated the findings. Dependability and confirmability were supported through collaborative coding, researcher discussion, peer review of emerging interpretations, and refinement of themes across the research team.

Findings

To address our research questions regarding (1) how upper-level mathematics students interact and adapt their interactions with GenAI tools to support mathematical reasoning, and (2) how their evaluation of mathematical validity shapes their trust and agency, we conducted a thematic analysis of the six group projects, presentation slides, and post-semester survey responses. This analysis surfaced three overarching themes and 7 subordinate themes regarding how students navigated Generative AI during mathematical problem solving. These themes and

their corresponding subthemes form the foundation of our results, which we have mapped onto the *GenAI Cognitive Adaptation Framework.* As illustrated in Fig. 2, the three primary themes that emerged were: (1) Strategic Prompt Engineering, (2) Navigating Trust and Verification, and (3) Navigating Tools' Constraints. The first and third themes directly address our first research question regarding student adaptation and tool constraints, while the second theme directly addresses our second research question regarding verification and trust. These themes collectively capture the processes through which students learned to strategically shape GenAI outputs, critically evaluate mathematical validity, and navigate the practical constraints of different AI platforms that shapes students' mathematical workflows and learning experiences. For each theme, we will share quotes from the project reports,

highlighting sub-themes, and then follow-up with supporting insights from the presentation slides and survey responses.

Strategic Prompt Engineering

The first theme captures how students progressively developed strategic prompting practices to make Generative AI tools more mathematically useful. Our analysis revealed that students consistently recognized that standard, conversational natural language was insufficient for advanced mathematics, forcing them to actively engineer their prompts and undergo the process of Instrumentalization—actively modifying and controlling the tool to fit their specific cognitive needs. Three closely connected subthemes emerged within this broader theme: (1) *from zero-shot to chain prompting*, where students shifted from single queries to sequential, structured guidance; (2) *persona adoption and role-playing*, where learners assigned high-level academic identities, say the role of a professor or PhD student, to the AI to extract more objective and reflective feedback; and (3) *demanding mathematical formalism and specificity*, where students explicitly dictated formatting constraints and theorem-based rigor to prevent wandering logic. Together, these findings suggest that students gradually learned that effective mathematical use of GenAI required active shaping of the tool rather than passive acceptance of its initial outputs. Table 3 details representative evidence from the student project reports, demonstrating how these prompting strategies were utilized to shape the AI artifacts.

Table 3
Evidence of Strategic Prompt Engineering

Group	Evidence	Thematic Connection
Group 4	“In terms of prompting, chain prompting and asking the AI to evaluate its own work and to try different approaches proved far more effective than simply copying and pasting a math question.”	*From Zero-Shot to Chain Prompting*: The student recognizes that the tool requires structured, sequential guidance to function mathematically rather than relying on single prompts.
Group 1	“Asking the AI tool to evaluate its own answer as if grading a student seemed to produce more objectivity in its work than a tone prompt... the AI consistently highlighted previous weaknesses in its own work.”	*Persona Adoption and Role-Playing*: By assigning the AI a specific evaluator role helps generate more objective and reflective mathematical feedback.
Group 2	“ChatGPT was especially receptive to being asked to respond as if it were a PhD student explaining the response to a skeptical professor.”	*Persona Adoption and Role-Playing*: Altering the AI's default conversational tone by demanding a high-level academic persona to improve explanation quality.
Group 2	“I found that the most significant contributor to getting higher quality responses is to insist that all arguments be rigorous, and mathematical in nature, citing theorems when necessary.”	*Demanding Mathematical Formalism*: Highlights the need to explicitly request rigorous, theorem-based arguments to prevent the AI from producing vague or wandering logic.
Group 4	“If we didn’t mention the format, like whether we wanted matrix form or just a sentence answer, the tool would choose on its own and it didn’t always match what we wanted.”	*Demanding Specificity*: Shows that students had to provide strict formatting constraints to ensure the output matched their specific mathematical needs.

The profound necessity of strategic prompt engineering was heavily reflected in the students' final presentations. Students summarized similar strategies in their slides as:

- "Break your request into clear, ordered steps instead of asking everything at once, then make follow up statements if the response isn’t satisfactory".
- "Prompt AI to act as a form of higher level education and explain down".
- "Be clear with the type of response you want... and indicate the level of response you want".

These qualitative observations were robustly corroborated by the post-semester survey. When asked what specific skills or insights they gained regarding GenAI in mathematics, respondents overwhelmingly cited:

- "prompting techniques",
- "prompting techniques and tool comparison".
- "better methods for prompting a GenAI tool to give me correct and formal mathematical answers"
- "Specific, formal, and instructive prompting can vastly improve the quality of outputs"

Navigating Trust and Verification

The second theme captures how students negotiated trust in GenAI outputs while working on advanced mathematical tasks. Across the projects, students consistently recognized that AI-generated responses could appear mathematically convincing while still containing logical inconsistencies, computational mistakes, misleading explanations, or unjustified conclusions. As a result, students did not treat GenAI tools as fully reliable mathematical authorities. Instead, they developed epistemic vigilance aimed at critically evaluating the correctness, usefulness, and trustworthiness of AI-generated outputs before accepting them. To enact this vigilance, students relied on foundational mathematical reasoning practices, such as cross-verification. For example, Group 6 explicitly cross-verified the AI's Gaussian elimination outputs by using an alternative method such as matrix inversion, or Group 4 developed a prompting strategy demanding the AI "verify its solutions by working backward through the problem".

We realized that students defined the boundaries of trust by positioning the AI as a *Supplementary Aid vs. Absolute* Trust and exhibited content vigilance by distinguishing between *Correctness vs. Quality*. They recognized a profound gap between an artifact's pragmatic value (arriving at the correct final number) and its epistemic value (providing a mathematically sound, pedagogically useful explanation). Exercising strict cognitive authority, students frequently critiqued tools that utilized extraneous complexity, advanced abstract methods, or overwhelming verbosity that actively hinder genuine mathematical comprehension. To validate their own epistemic vigilance, Group 5 took the unprecedented step of directly asking the AI artifacts if they should be trusted. When asked, "Should I trust you to do math in a high-stakes math setting?", ChatGPT responded: "You can trust me to help explain concepts, check your reasoning, and quickly verify calculations, but in a high-stakes math setting... you should not rely on me as your only source of truth. I can make mistakes". Claude similarly answered: "No, you should not trust me to do math in high-stakes settings. I can make calculation errors... critical calculations require independent verification". This meta-interaction reinforced the students' critical stance, leading Group 5 to explicitly conclude: "It should also be made clear that the answers models give should be verified, and they should not be trusted in high stakes mathematics settings". Table 4 details representative evidence from the student project reports, demonstrating how students executed epistemic vigilance to evaluate the mathematical validity and pedagogical utility of the AI artifacts.

Table 4
Evidence of Navigating Trust and Verification

Group	Evidence	Thematic Connection
Group 4	“If we were in a math setting, we would definitely not fully trust them to stand on their own. The risk of a small error is too high when the stakes are serious. We would, however, trust them as support. For example, we could use one to check our own work, see if we missed a step, or compare my approach”.	*Supplementary Aid vs. Absolute Trust*: Students reject relying on AI for serious computations due to the risk of errors, instead viewing it as a tool strictly for verifying human work.
Group 6	“Copilot paid’s response to Prompt 1 was concerning due to its laziness. The true x-values are 1/sqrt(2) and − 1/sqrt(2), which respectively round to 0.71 and − 0.71. Yet, Copilot is okay with rounding to 1 and − 1. In any setting where the actual numbers need to be used, this result is *atrocious*. The approach that was taken here by a paid version of a generative AI... serves as a reminder that these tools are *fallible*, even for common types of math problems.”.	*Supplementary Aid vs. Absolute Trust*: The identification of “atrocious” computational rounding errors reinforces to students that even premium AI models are fallible and cannot be blindly trusted with precise calculations.
Group 6	“Although all tools gave correct answers, one main difference was in their explanations and use of linear algebra terminology. For example, ChatGPT often used the determinant... Deepseek, on the other hand, relied on concepts like injectivity and the null space being zero (which are more advanced and abstract)”	*Correctness vs. Quality*: Illustrates that students evaluate AI outputs not just for numerical accuracy, but for whether the specific theoretical approach and terminology are pedagogically useful.
Group 1	“Prompting ChatGPT-free to give a mathematically detailed answer to the eigenvalue prompt unleashed a very detailed scenario involving theoretical Jordan blocks, even though it did not have sufficient information to build a Jordan form matrix. It also attempted to discuss whether or not the matrix would be diagonalizable, although that was not asked about in the prompt”.	*Correctness vs. Quality*: Highlights student frustration when tools generate unprompted, extraneous theoretical complexity that distracts from the core mathematical problem.

This perception of the tool's role was also a dominant takeaway during the student presentations. Some individuals highlighted that by expressing:

- " GenAI is a great supplement, but should not be relied on as a standalone calculation tool" .
- "Always Check the Work You’re Given. AI’s are not calculators, do not blindly trust the results you are given".
- "Main goal: Avoid a conversation. Keep prompts concise and clear. Evaluate and cross reference".

The survey responses also reveal that students recognize the necessity of epistemic vigilance, but still feel uncertain about how to execute it perfectly. They expressed:

- “I need more training on how to check the answers GenAI gives me”.
- “How to verify the answer to something that I am still learning about”.

- “It often feels that I struggle to find errors in the GenAI’s reasoning even though we came to the same answer.”

Navigating Tools Constraints

The third theme which is characterized by Utility Gap (tool-specific technical constraints) and Digital Divide (differences between paid and free versions of AI systems) captures how students’ mathematical workflows were shaped by the technical, structural, and access-related limitations of different Generative AI platforms. Across the projects, students consistently observed that GenAI systems varied considerably in their mathematical capabilities, formatting quality, conversational memory, response restrictions, and accessibility features. As a result, students were consistently forced to adapt their mathematical behaviors to accommodate the limitations of the AI tools. Rather than functioning as neutral tools, the AI systems themselves actively influenced how students approached prompting, verification, problem solving, and sustained mathematical interaction. Table 5 details representative evidence from the student project reports, demonstrating how the material constraints of the artifacts shaped the students' experiences and raised concerns regarding digital equity.

Table 5
Evidence of Navigating Tools Constraints

Group	Evidence	Thematic Connection
Group 6	"Copilot said it was unable to provide an answer to the [English prompt], while the French instructions were permissible... the French response included more calculations."	*The Utility Gap*: Shows an unpredictable linguistic constraint where the tool's arbitrary refusal of English instructions forces the student to adapt their problem-solving scheme
Group 5	"Gemini was not able to show its work for Prompt 3, giving unreadable strings of numbers and signs. When asked why the equations were unreadable, it had no idea that it gave bad equations and said that it wasn’t capable of showing complex math."	*The Utility Gap*: Demonstrates a technical limitation where the tool's inability to render mathematical formatting (LaTeX) disrupts the student's workflow.
Group 6	"We found that ChatGPT-5 (paid) allows much longer conversations than needed, though after a very long session it sometimes slows down. Deepseek (free) has a much stricter limit on chat length. This is disruptive and reduces its usefulness for extended reasoning."	*Digital Divide*: Highlights how memory limits on free tiers disrupt continuous workflows and restrict sustained mathematical reasoning compared to premium tools.
Group 3	“The main difference is that the free version cannot generate images when a user is not logged in and overall provides less detailed explanations.”	*Digital Divide*: Shows how economic paywalls restrict access to visual representations and detailed explanations, putting free-tier users at a disadvantage.

Students’ slides also highlighted tool-specific constraints and adaptation to problem solving schemes:

- “One Tool Doesn't Fit All Play with Them!”
- “Smarter models thrive on more context! Quicker, smaller models should be given simple and short instructions.”

The survey data explicitly supports this divide with majority of students (12 out of 15) stating that paid AI provides a significant or moderate advantage over free tools.

Discussion

This study examined how upper level undergraduate and graduate students interacted with, evaluated, and adapted to Generative AI systems while solving mathematical problems. Across the findings, students did not engage with GenAI tools as passive recipients of automated solutions. Instead, they developed cognitive adaptation schemes to navigate probabilistic LLMs. To answer our first research question regarding how students adapt their interactions to support mathematical reasoning, our findings demonstrate that GenAI chatbots in their default, conversational state are merely artifacts lacking inherent pedagogical efficacy. Viewed through the lens of Instrumental Genesis, students adapt to these limitations through *Strategic Prompt Engineering*, which serves as empirical evidence of *instrumentalization*— the process of a user modifying an artifact's behavior to fit specific operational and cognitive needs (Artigue, 2002; Trouche, 2004). By utilizing chain prompting, setting formatting constraints, and demanding the adoption of academic personas, students forced the conversational artifact to operate as a rigorous mathematical instrument. This necessity for active instrumentalization mirrors broader findings in educational technology; for instance, Maleki et al. (2024) demonstrated empirically that AI-generated content inherently lacks in-depth discussion, making it crucial for users to actively design and refine prompts to achieve educational quality. However, this adaptation was not a one-way process. Aligning with our third theme *Navigating Tools Constraints*, the phenomenon of *instrumentation*—how the artifact prints its mark on the subject—was clearly observable as the physical and technical constraints of the tools dictated the students' workflows (Trouche, 2004). Platform-specific limitations, such as restricted conversational memory on free tiers, inability to render LaTeX, or arbitrary linguistic refusals, created a pronounced Utility Gap that forced students to abandon cognitive schemes or switch tools entirely. Beyond shaping workflow efficiency, these differences also raise equity concerns. Students repeatedly noted that premium models provide advantages that are unavailable to all learners.

In relation to the second research question, students’ verification practices show that evaluating GenAI outputs involved more than checking whether the final answer was correct. The data revealed a profound tension between an artifact's *pragmatic value*—its ability to successfully and rapidly produce the correct final answer—and its *epistemic value*—its contribution to the actual comprehension of the mathematical objects involved (Artigue, 2002). Students recognized that when AI models provided mathematically correct but unnecessarily complex information––such as unleashing unprompted, highly abstract theoretical scenarios involving Jordan blocks––it actively confused them and disrupted the learning process. Beyond the context of mathematics, broader evaluations of generative AI also corroborate this

limitation, revealing that while AI can generate coherent outputs to meet basic checklist criteria, it falls short of the pedagogical richness and interactive scaffolding that human expertise provides (Maleki & Wiest, 2026). Therefore, GenAI is not a neutral tutor; students must act as active filters to extract pedagogical quality without becoming cognitively overloaded by verbose or extraneous logic.

This active exercise of filtering illuminates a critical divide between novice and expert learners that sits at the core of our second theme in *Navigating Trust and Verification.* Recent literature, such as the large-scale field experiment by Bastani et al. (2025), demonstrates that unrestricted GenAI can actively harm learning for high school mathematics students who use the tool as a "crutch" to bypass cognitive effort during practice, subsequently performing worse on unassisted exams. Novices, lacking foundational background knowledge, are highly vulnerable to automation bias and blind trust. In contrast, the participants in this study were higher level STEM students who possessed the requisite domain knowledge to rigorously evaluate AI outputs. Because these students were explicitly tasked with critiquing the mathematical validity of the AI's reasoning rather than simply using it to find answers, this study offers a novel contribution to the literature by demonstrating how mathematically knowledgeable learners operate when positioned as critical evaluators of technology rather than passive consumers. As Chaaban et al. (2025) highlight, drawing on Eaton (2023) and Vetter et al. (2024), students must actively retain control over their learning process rather than surrendering it to efficient GenAI systems, a dynamic described as a "competition over agency". This competition over agency reflects a broader concern among mathematics educators; as Castle et al. (2026) highlight, unless students actively cultivate critical evaluation skills, they risk outsourcing their intellectual labor and allowing GenAI to replace the fundamental process of their own mathematical reasoning.

Rather than requiring external pedagogical guardrails coded into the software by developers, these advanced students constructed *internal cognitive guardrails* through Epistemic Vigilance—the mental filters people use to protect themselves from being misinformed (Sperber et al., 2010). Crucially, however, this cognitive defense mechanism requires a high degree of mathematical maturity to function. Because vigilance towards content relies on "coherence checking"—the psychological process of evaluating new information against deeply held background beliefs (Sperber et al., 2010)—learners must possess sufficient conceptual understanding, domain expertise, and familiarity with mathematical structures to recognize when an AI's output deviates from logical consistency. By actively evaluating both the source of the information and the mathematical coherence of the content, students shifted their view of the AI. Instead of treating the tool as a perfect mathematical authority, they actively repositioned it as a fallible supplementary aid. This aligns with recent findings by Kock et al. (2025), which showed that engineering students using ChatGPT gradually developed critical attitudes towards the information provided, understanding it might offer incorrect or incomplete answers and learning to verify those outputs over time.

Ultimately, these findings suggest that the educational impact of Generative AI is highly dependent on the user's prior mathematical knowledge and their capacity to exercise epistemic vigilance to maintain ultimate cognitive authority over the problem-solving process.

Pedagogical Implications: Integrating GenAI in Teaching Mathematics

The findings of this study demonstrate that integrating Generative AI into mathematics education requires deliberate instructional design aimed at developing students' critical evaluation skills. A shared perspective across higher education emphasizes that AI technology is permanently embedded in society, meaning that prohibiting its use is entirely unrealistic (Žáková et al., 2024), although at the upper-division undergraduate level, students often resist using GenAI. However, because the broader adoption of these tools in society and the workforce is inevitable, educators have a responsibility to address their proliferation and teach students the value and proper usage of these technologies.

A core component of teaching GenAI literacy is explicitly defining the proper time and place for its use. While GenAI models are rapidly improving and their computational performance is significantly enhanced when leveraging underlying computer language interpreters, such as Python (Carbone, 2025), they can still produce subtle errors in advanced symbolic mathematics (Frieder et al., 2023; Walkington, 2025). Instructors might emphasize that deterministic software such as Wolfram Alpha and Desmos should remain standard tools for rigorous computation (Laurent, 2025), while GenAI can serve as a valuable source of inspiration for symbolic problems and coding.

Building upon the instructional design utilized in this course, we highly recommend that educators actively cultivate students' epistemic vigilance by implementing critical thinking exercises built around the evaluation of genuine AI outputs. By analyzing authentic AI mistakes, or instances where the AI obfuscated the correct answer with extraneous information, students practiced identifying subtle conceptual and computational errors. As broader educational research affirms, providing learners with opportunities to explain, justify, and reflect on their reasoning can support the development of deeper conceptual understanding and more critical engagement with complex tasks (Luo et al., 2026). We recommend that instructors adopt this strategy in their own classrooms, framing these evaluations as parallel to scientific peer review or the assessment of co-worker input, which demonstrates to students that rigorous verification and critical thinking are essential professional skills.

To ensure GenAI enhances rather than replaces mathematical reasoning, instructors might also teach students to intentionally constrain the tool's outputs. A particularly productive pedagogical approach would be training students to prompt the AI to act as a guide rather than as a solution generator. This instructional priority aligns with the core professional values of mathematics educators, who emphasize that students must cultivate critical evaluation skills to use GenAI as a collaborative dialogue partner rather than a replacement for their own reasoning (Castle et al., 2026). By instructing the tool to provide only incremental hints, students maintain their engagement in productive struggle without surrendering their cognitive authority, positioning the AI as a scaffold for independent discovery rather than an intellectual crutch (Bastani et al., 2025; NCSM, 2025).

Ultimately, the goal of integrating GenAI into advanced mathematics courses is not to force its adoption or demand that students view the technology positively. Instead, the objective is to expose students to the erratic outcomes and tricky evaluations inherent to these tools in order to encourage routine critical examination of generative AI output. When students independently experience and critique the failures of GenAI, they develop a profound distrust of the tool's unverified outputs, which serves as a highly valuable teaching outcome. This cultivated skepticism is the pedagogical manifestation of epistemic vigilance, ensuring that students retain ultimate cognitive authority over their mathematical work.

Conclusion

As Generative AI continues to disrupt traditional educational paradigms and becomes an inevitable presence that offers almost instant solutions, mathematics education must adapt to equip students with the skills required to navigate these tools and ensure students do not succumb to the temptation of letting the AI think and decide for them. The findings of this study demonstrate that the true educational value of GenAI does not lie simply in obtaining a correct and complete final answer. Rather, GenAI is a raw artifact that students can actively shape into a useful mathematical instrument through strategic prompt engineering, while rigorously navigating and evaluating the AI's output by exercising epistemic vigilance. Through this continuous process of navigation and refinement, students actually deepen their own learning.

Crucially, successfully transforming AI into a useful instrument is highly dependent on the user possessing sufficient prior mathematical knowledge and their capacity to exercise epistemic vigilance. Because this exploratory study was conducted within a single 400-level Numerical Linear Algebra course with a relatively small sample of upper-level undergraduate and graduate students, these interaction strategies may not generalize to novice learners in other mathematical domains. Furthermore, as this research focused primarily on students' interaction and verification practices rather than directly measuring mathematical learning outcomes, future research should systematically investigate how learners with different conceptual backgrounds interact with GenAI systems. As demonstrated in previous literature, novices without sufficient mathematical background are highly vulnerable to automation bias and may be actively harmed by unrestricted AI access. Ultimately, the goal of teaching students to critically evaluate AI outputs and question mathematical coherence is to cultivate a stance of epistemic vigilance, ensuring that students maintain their cognitive independence, independent problem-solving skills, and mathematical agency in an era where automated assistance is unavoidable.

Declarations

Funding The authors did not receive support from any organization for the submitted work.

Competing Interests Financial interests: The authors declare they have no financial interests.

Non-financial interests: The first author was concurrently enrolled as a graduate student in the course under investigation.

Data Availability The datasets generated and analyzed during the current study are not publicly available to protect participant privacy and maintain consent conditions, but they are available from the corresponding author on reasonable request.

Ethics Approval and Consent This study was conducted under Institutional Review Board (IRB) approval. Informed consent was obtained from all individual participants whose data is included in this study.

Author Contributions The first, second, and third authors contributed to the study design and data collection methods. The first, second, and fourth authors contributed to data analysis. All authors contributed to manuscript preparation.

References

1. Artigue, M. (2002). Learning mathematics in a CAS environment: The genesis of a reflection about instrumentation and the dialectics between technical and conceptual work. *International Journal of Computers for Mathematical Learning*, *7*(3), 245–274.
2. Baidoo-Anu, D., & Owusu Ansah, L. (2023). Education in the era of generative artificial intelligence (AI): Understanding the potential benefits of ChatGPT in promoting teaching and learning. *Journal of AI*, *7*(1), 52–62. https://doi.org/10.61969/jai.1337500
3. Barana, A., Marchisio, M., & Roman, F. (2023). Fostering problem solving and critical thinking in mathematics through generative artificial intelligence. In D. G. Sampson, D. Ifenthaler, & P. Isaías (Eds.), *Proceedings of the 20th international conference on cognition and exploratory learning in digital age (CELDA 2023)* (pp. 377–385). IADIS Press. https://doi.org/10.33965/celda2023_202306l046
4. Bastani, H., Bastani, O., Sungu, A., Ge, H., Kabakcı, Ö., & Mariman, R. (2025). Generative AI without guardrails can harm learning: Evidence from high school mathematics. *Proceedings of the National Academy of Sciences*, *122*(26), e2422633122. https://doi.org/10.2139/ssrn.4895486
5. Bhutoria, A. (2022). Personalized education and artificial intelligence in the United States, China, and India: A systematic review using a human-in-the-loop model. *Computers and Education: Artificial Intelligence*, *3*, 100068. https://doi.org/10.1016/j.caeai.2022.100068
6. Borji, A. (2023). *A categorical archive of ChatGPT failures*. arXiv. https://doi.org/10.48550/arXiv.2302.03494
7. Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, *3*(2), 77–101. https://doi.org/10.1191/1478088706qp063oa
8. Braun, V., & Clarke, V. (2020). One size fits all? What counts as quality practice in (reflexive) thematic analysis? *Qualitative Research in Psychology*, *18*(3), 328–352. https://doi.org/10.1080/14780887.2020.1769238
9. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J.,

Berner, C., McCandlish, S., Radford, A., Sutskever, I., & Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, *33*, 1877–1901.

10. Carbone, L. (2025). Advancing mathematics research with generative AI. *arXiv preprint*, arXiv:2510.12350.
11. Castle, S. D., Wolfe, M., Yoon, H., & Abbaspour Tazehkand, S. (2026). *Values in tension: Mathematics teacher educator's decisions on generative AI*. In *Proceedings of the 28th Annual Conference on Research in Undergraduate Mathematics Education* (pp. 1161–1167). SIGMAA on RUME. https://sigmaarume.org/proceedingsdatabase/pdf/2026/RUME_28/RUME2026_Castle_2974.pdf
12. Chaaban, Y., Jung, S. G., Medina, J., Azem, J. Y., Salminen, J., & Jansen, B. J. (2025). Examining student teachers' agency in an AI-supported learning environment: Q methodology research. *International Journal of Artificial Intelligence in Education*, *35*, 4083–4107.
13. Contel, F., & Cusi, A. (2025). Investigating the role of ChatGPT in supporting metacognitive processes during problem-solving activities. *Digital Experiences in Mathematics Education*, *11*(1), 167–191. https://doi.org/10.1007/s40751-024-00164-7
14. Creswell, J. W., & Poth, C. N. (2016). *Qualitative inquiry and research design: Choosing among five approaches*. Sage.
15. Forootani, A., Aliabadi, D. E., & Thraen, D. (2024). A survey on mathematical reasoning and optimization with large language models. *arXiv preprint*.
16. Frieder, S., Pinchetti, L., Chevalier, A., Griffiths, R. R., Salvatori, T., Lukasiewicz, T., Petersen, P., & Berner, J. (2023). Mathematical capabilities of ChatGPT. *37th Conference on Neural Information Processing Systems (NeurIPS 2023)*. https://arxiv.org/abs/2301.13867
17. Giannakos, M., Azevedo, R., Brusilovsky, P., Cukurova, M., Dimitriadis, Y., Hernandez-Leo, D., Järvelä, S., Mavrikis, M., & Rienties, B. (2024). The promise and challenges of generative AI in education. *Behaviour & Information Technology*, *44*(11), 2518–2544. https://doi.org/10.1080/0144929X.2024.2394886
18. Heid, K., & Edwards, M. T. (2001). Computer algebra systems: Revolution or retrofit for today's mathematics classrooms? *Theory into Practice*, *40*(2), 128–136.
19. Israel, S. E., Assaf, L. C., & Abedini, P. (2026). 532 Where to From Here? In Handbook of Research on Reading Comprehension (pp. 532–542). Routledge. https://doi.org/10.4324/9781003624509-32
20. Kock, Z. J., Salinas-Hernández, U., & Pepin, B. (2025). Students' initial use schemes of ChatGPT as an instrument for learning. *Digital Experiences in Mathematics Education*, *11*, 192–218. https://doi.org/10.1007/s40751-025-00169-w
21. Kooli, C. (2023). Chatbots in education and research: A critical examination of ethical implications and solutions. *Sustainability*, *15*(7), 5614. https://doi.org/10.3390/su15075614
22. Laird, E., Dwyer, M., & Grant-Chapman, H. (2023). Off task: Edtech threats to student privacy and equity in the age of AI. *Center for Democracy & Technology*. https://cdt.org/insights/report-off-task-edtech-threats-to-student-privacy-and-equity-in-the-age-of-ai/

23. Laurent, A. (2025, December 18). *Comparing symbolic and generative AI: Wolfram Alpha & ChatGPT*. IntuitionLabs. https://intuitionlabs.ai/articles/symbolic-ai-vs-generative-ai-wolfram-chatgpt
24. Lincoln, Y. S., & Guba, E. G. (1985). *Naturalistic inquiry*. Sage.
25. Lo, C. K. (2023). What is the impact of ChatGPT on education? A rapid review of the literature. *Education Sciences*, *13*(4), 410. https://doi.org/10.3390/educsci13040410
26. Luo, F., Amayou, J., Nasrin, F., Abedini, P., Mousavi, V., Awoyemi, I. D., Cao, C., & Crawford, C. (2026). *Promoting elementary students' computer science conceptual knowledge and coding performance: A mixed methods case study*. *Journal of Research on Technology in Education*. Advance online publication. https://doi.org/10.1080/15391523.2026.2661645
27. Maleki, E., & Wiest, L. (2026). Assessing AI-Generated Online Discussion Assignments in an Asynchronous Bioethics Course: A Comparative Analysis of AI vs Human. In Proceedings of Society for Information Technology & Teacher Education International Conference (pp. 815–824). Philadelphia, PA: Association for the Advancement of Computing in Education (AACE). https://www.learntechlib.org/primary/p/2129064/
28. Maleki, E., Chen, L. T., Vijayakumar, T. M., Asumah, H., Tretheway, P., Liu, L., Fu, Y., & Chu, P. (2024). AI-generated and YouTube videos on navigating the U.S. healthcare systems: evaluation and reflection. *International Journal of Technology in Teaching and Learning*, *20*(1), 40–72.
29. National Council of Teachers of Mathematics (NCTM) (2024). *Artificial intelligence and mathematics teaching: A position of the National Council of Teachers of Mathematics*. https://www.nctm.org/standards-and-positions/Position-Statements/Artificial-Intelligence-and-Mathematics-Teaching/
30. National Council of Supervisors of Mathematics. (2025). *Educational technology & AI guidance for math leaders*. NCSM.
31. Patton, M. Q. (2015). *Qualitative research & evaluation methods: Integrating theory and practice* (4th ed.). SAGE.
32. Phung, T., Padurean, V. A., Cambronero, J., Gulwani, S., Kohn, T., Majumdar, R., Singla, A., & Soares, G. (2023). *Generative AI for programming education: Benchmarking ChatGPT, GPT-4, and human tutors*. arXiv. https://doi.org/10.48550/arXiv.2306.17156
33. Rabardel, P. (1995). *Les hommes et les technologies: Approche cognitive des instruments contemporains*. Armand Colin.
34. Rabardel, P. (2002). *People and technology: A cognitive approach to contemporary instruments*. *Université Paris* 8. https://hal.science/hal-01020705v1
35. Sabzalieva, E., & Valentini, A. (2023). ChatGPT and artificial intelligence in higher education: Quick start guide. *UNESCO*. https://unesdoc.unesco.org/ark:/48223/pf0000385146
36. Sperber, D., Clément, F., Heintz, C., Mascaro, O., Mercier, H., Origgi, G., & Wilson, D. (2010). Epistemic vigilance. *Mind & Language*, *25*(4), 359–393. 10.1111/j.1468-0017.2010.01394.x. https://onlinelibrary.wiley.com/doi/full/

37. Trouche, L. (2004). Managing the complexity of human/machine interactions in computerized learning environments: Guiding students' command process though instrumental orchestrations. *International Journal of Computers for Mathematical Learning*, *9*(3), 281–307. https://doi.org/10.1007/s10758-004-3468-5
38. Vérillon, P., & Rabardel, P. (1995). Cognition and artifacts: A contribution to the study of thought in relation to instrumented activity. *European Journal of Psychology of Education*, *10*(1), 77–101. https://www.jstor.org/stable/23420087
39. Walkington, C. (2025). The implications of generative artificial intelligence for mathematics education. *School Science and Mathematics*, 1–10. https://doi.org/10.1111/ssm.18356
40. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, *35*, 24824–24837. https://doi.org/10.48550/arXiv.2201.11903
41. Xu, H., Gan, W., Qi, Z., Wu, J., & Yu, P. S. (2024). *Large language models for education: A survey*. arXiv. https://doi.org/10.48550/arXiv.2405.13001
42. Žáková, K., Urbano, D., Cruz-Correia, R., Guzmán, J. L., & Matišák, J. (2024). Exploring student and teacher perspectives on ChatGPT's impact in higher education. *Education and Information Technologies*, *30*, 649–692. https://doi.org/10.1007/s10639-024-13184-y
43. Zhuang, Y., & Zhang, S. (2025). Pre-service mathematics teachers' perceptions of using GenAI for practicing teacher questioning: A semester-long study. *Eurasia Journal of Mathematics Science and Technology Education*, *21*(9), em2689. https://doi.org/10.29333/ejmste/16764

Figures

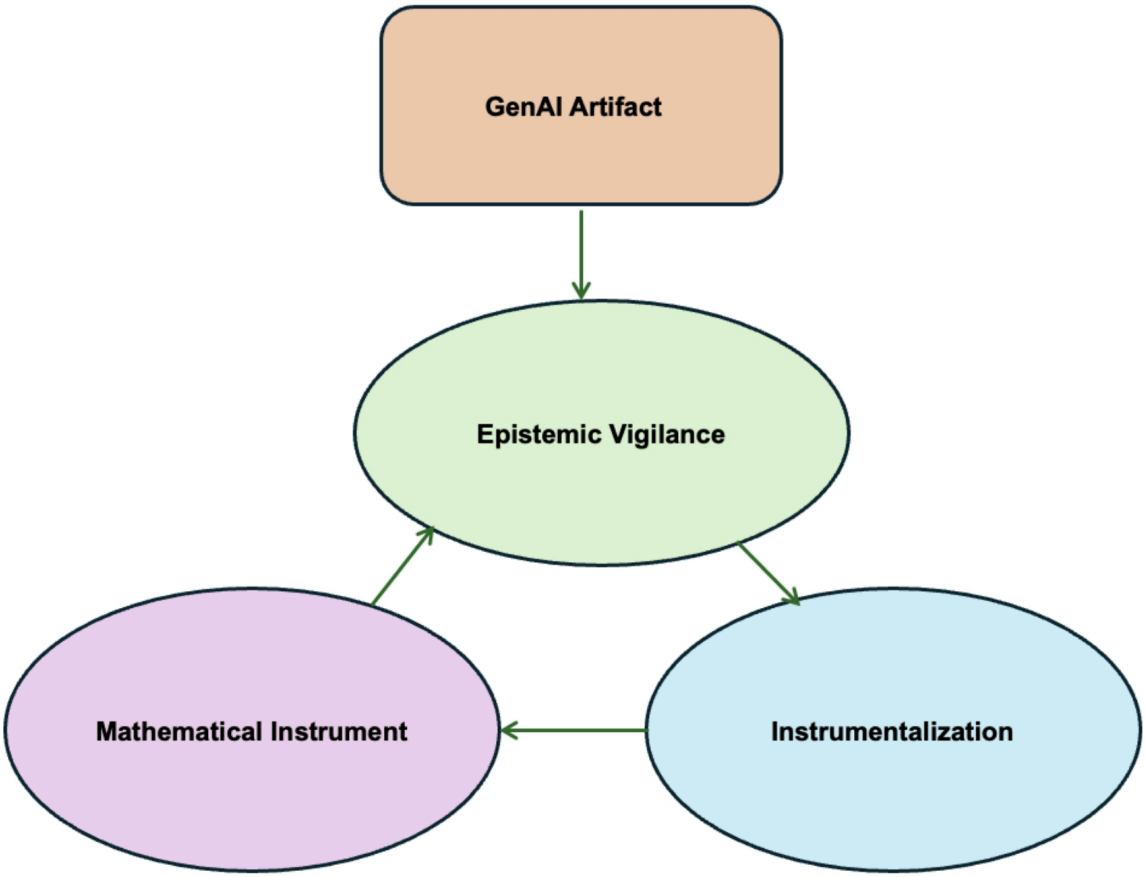


Figure 1

The Cyclical Model of GenAI Instrumental Genesis and Epistemic Vigilance

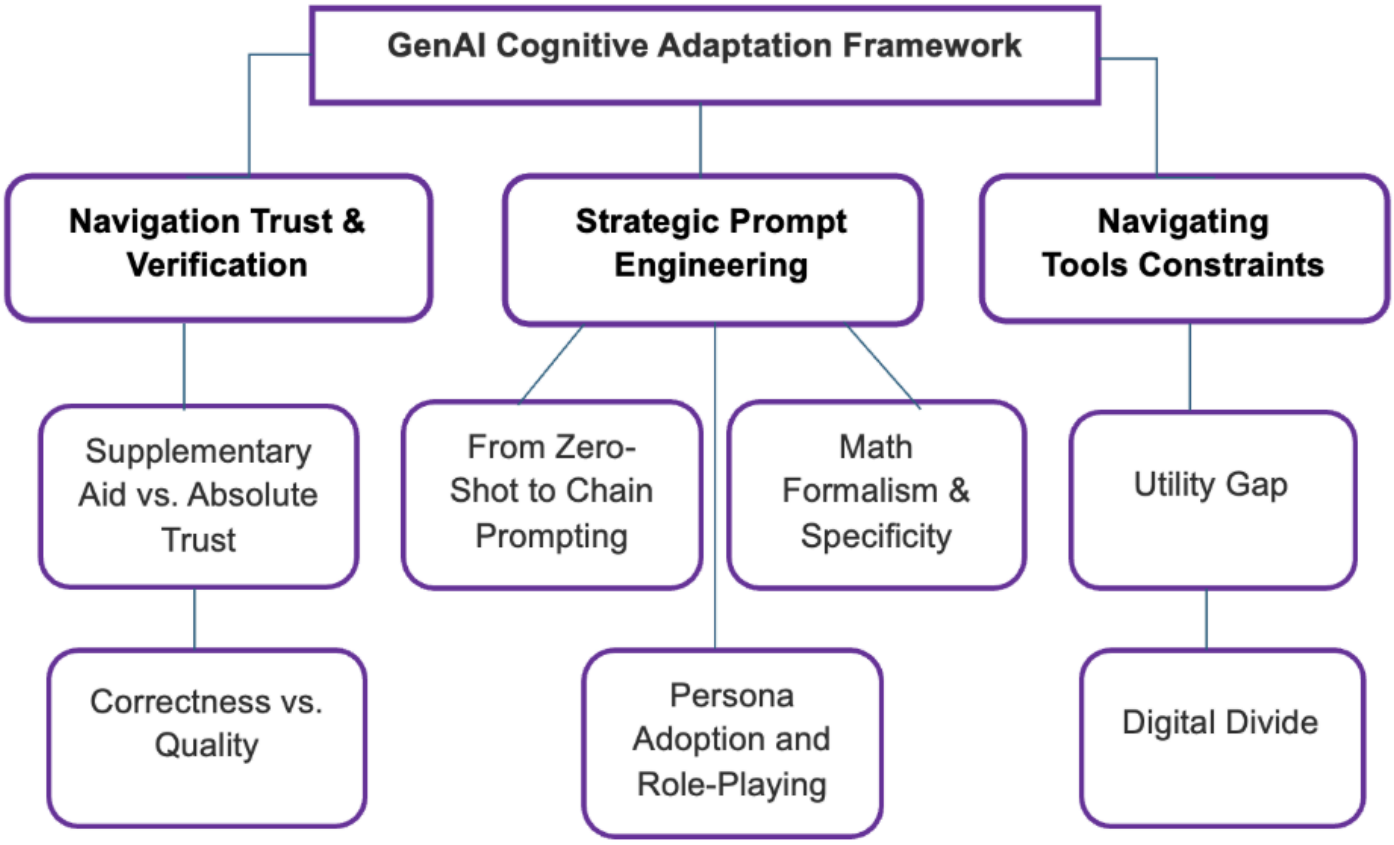


Figure 2

The GenAI Cognitive Adaptation Framework