

An Exploratory Pilot Study of Compose With AI: Scaffolding Critical Evaluation and Ethical Integration of AI-Generated Content in Disciplinary Writing

Dr. Amy Hutchison, University of Alabama; Orcid: [0000-0003-1776-4552](https://orcid.org/0000-0003-1776-4552)

Dr. Lori Bruner, University at Albany; Orcid: [0000-0001-8845-0206](https://orcid.org/0000-0001-8845-0206)

Dr. Erdogan Kaya, George Mason University; Orcid: [0000-0002-3211-3259](https://orcid.org/0000-0002-3211-3259)

Abstract

This study explored the design and exploratory evaluation of *Compose With AI*, a web-based platform guiding grades 4-8 students (piloted with grades 5-7 participants) in ethically integrating AI-generated content into science-focused writing. Using a Design-Based Implementation Research (DBIR) approach, researchers analyzed student think-aloud protocols and teacher interviews to identify factors enhancing and inhibiting critical AI evaluation. Findings indicate that *Compose With AI*'s structured, scaffolded approach helped students demonstrate emerging abilities to critically evaluate AI-generated content, even with no prior AI experience. The platform's step-by-step guidance, including prompt engineering and iterative refinement, appeared to foster metacognitive engagement and ethical AI use. A key feature, the "keep, change, omit" function, supported responsible decision-making about incorporating AI-generated content. The teacher dashboard offered a comprehensive view of students' writing processes, fostering transparency and accountability. This exploratory study contributes early design insights into how educational technology might support students in integrating AI-generated content into their writing while cultivating their evaluative reasoning. *Compose With AI* offers one promising design for scaffolding ethical AI-assisted writing and critical evaluation of AI-generated content, though larger and more diverse studies are needed to examine its broader applicability.

Keywords: *AI in education, generative AI, disciplinary writing, STEM, critical evaluation, prompt engineering, ethical AI use, AI pedagogy*

Funding

This research was made possible by funding from the US National Science Foundation Grant number 2337969.

Introduction

The transformative potential of Artificial Intelligence (AI) technologies across various sectors of society has generated significant interest and debate, particularly in the field of education (Bennett & Abusalem, 2024; Guo et al., 2024; Yan et al., 2024). The swift adoption of generative Natural Language Processing (NLP) tools, such as ChatGPT (Hu, 2023; Trust et al., 2023), has created an urgent need to understand how to support students in using this technology effectively. Still, generative AI tools such as ChatGPT hold significant potential for supporting students' learning. For example, generative AI can be used for building students' background knowledge and explaining complex concepts, which is crucial for comprehending disciplinary texts such as scientific articles or textbooks (Hutchison, 2024; Zhai, 2022). These capabilities offer new avenues for supporting student learning and engagement with scientific content. Yet, the rapid pace of adoption often outpaces our understanding of appropriate pedagogical uses of such tools and of the risks involved. Without well-developed critical evaluation skills, students risk accurately assessing the content produced by chatbots or, worse, replicating inaccurate or biased information in their academic work. Students need to possess well-developed critical evaluation skills to accurately assess the content produced by chatbots (U.S. Department of Education, 2023).

As AI continues to evolve at an exponential rate, there is a critical need for instructional materials and tools to guide teachers and students in the safe and ethical use of AI. Current trends suggest that many students use AI to complete assignments without properly evaluating or rephrasing the content (Cu & Hochman, 2023; Roose, 2023). While some educational institutions have responded with restrictive bans, such policies hinder students' ability to

cultivate fundamental AI literacy and necessary content evaluation competencies. A more productive approach involves developing instructional tools that equip students with the skills for ethical and effective AI integration. This need is particularly acute in disciplinary writing, crucial for communication and meaning-making in disciplines such as science (Hand et al., 2018; Keys et al., 1999). However, disciplinary writing is often underemphasized in science classrooms (Huerta & Garza, 2019), and the increasing reliance on AI for academic writing (Chan & Hu, 2024) exacerbates this challenge. The transition from informal writing to the formal demands of scientific discourse presents a significant challenge for middle-grade students, but AI tools could potentially alleviate such challenges if properly scaffolded and used in a way that does not allow students to over-rely on generative AI.

To address these gaps, we developed *Compose With AI*, a web-based platform designed to guide students in grades 4–8 in ethically integrating AI-generated content into science-focused writing. This paper presents an exploratory pilot evaluation focused on the pedagogical usability of the tool. Drawing on standards from human and computer interaction education research, we define pedagogical usability as the interface’s ability to support both ease of navigation and the specific learning goals of the task—in this case, fostering critical AI evaluation and disciplinary writing skills. This concept draws on Ardito et al.’s (2006) framework for usability in learning environments, which distinguishes between technical usability (ease of use) and pedagogical usability (the degree to which a tool’s design actively supports intended learning outcomes). Unlike general usability heuristics (e.g., Nielsen, 1994), pedagogical usability requires that interface affordances be evaluated in relation to disciplinary learning goals, not merely navigational efficiency. Accordingly, we examined not only whether students found the platform

easy to use, but whether its design features actively scaffolded the critical thinking and disciplinary writing skills the platform was designed to cultivate.

Thinking “With” AI: Critical Skills Students Need

Amid the potential for AI to support students’ writing, there is also a significant need for students to understand how to critically evaluate AI-generated content and think “with” AI rather than letting AI think for them or simply copying and pasting the first response they get. Existing critical evaluation strategies for web-based information provide a starting point for guiding students critical evaluation of AI-generated content, but these strategies must be adapted to the specific context of AI. For example, lateral reading (Wineburg & McGrew, 2019) and media literacy strategies (e.g., Esmacili et al, 2019) provide useful starting points. Because students often struggle to evaluate online information (e.g., McGrew et al., 2018), we modeled these strategies with short, embedded videos within the *Compose With AI* platform. For purposes of the current study, we consider that thinking *with* AI requires several skills and approaches, which we consider next.

Prompt Engineering

At the foundation of critical engagement with AI lies the skill of prompt engineering. As Liu and colleagues (2023) suggest, this involves crafting precise, contextually appropriate prompts that guide AI to produce relevant and accurate information. Prompt engineering is not a static process but requires continuous refinement. It is a cyclical exercise where students must assess the adequacy of the AI’s initial output and iteratively adjust their prompts. This process allows for greater control over the output and ensures the AI generates the desired content. This continual process can be challenging for students, especially since well-developed pedagogical strategies for developing and evaluating prompts are still emerging (Walter, 2024). Even so,

recent research indicates that learning to generate and evaluate the relevance of AI prompts may help students better understand the limitations of AI (Theophilou et al., 2023). Accordingly, the Compose With AI platform was designed to teach students about prompt engineering and prompt refinement.

Critical Evaluation

Beyond the creation of prompts, students must be adept at critically evaluating the content generated by AI. This involves the ability to identify and select appropriate strategies evaluating the relevance and accuracy of a source and checking for potential bias (Bozkurt, 2024). Evaluating the accuracy and relevance of a source is a challenging task regardless of the text format, but is nonetheless an essential skill for conducting inquiry within most academic disciplines (Stanford History Education Group, 2016; McGrew et al., 2018). However, the specific context and limitations of AI-produced content may make it even more challenging for students to evaluate the accuracy and relevance of content. For example, AI output is dynamic, which means that a student may repeatedly get different output for the same query. Further, AI-generated content often lacks the nuanced contextual understanding that human authorship provides. Students evaluating such content need to be particularly adept at interpreting the context and relevance of information without the supporting cues typically available in human-authored materials (Luckin et al., 2022). This can complicate students' efforts to critically assess the applicability or validity of AI-produced information in specific contexts. Moreover, AI-generated content can produce information that appears genuine and well-structured, often mimicking the style and tone of expert discourse. This veneer of authenticity can make it difficult for students to discern subtle inaccuracies or inconsistencies (Alasadi & Baiz, 2023). The perceived authority of AI outputs demands a sophisticated level of scrutiny from students,

who must deploy advanced evaluative skills to distinguish between persuasive language and substantive content.

Integration With Authentic Thought

Effective use of AI in writing also necessitates integration of AI-generated content with students' own ideas. It is crucial for students to synthesize AI contributions with their authentic thoughts, ensuring that the final output reflects their personal voice and disciplinary understanding. This integration is not merely a mechanical process but involves critical reflection and analytical reasoning, as students evaluate how AI-generated suggestions align with their intended message and purpose. The *Compose With AI* platform was designed to address these pedagogical needs by providing structured scaffolding that guides students through each stage of engagement with AI.

Theoretical Framework Guiding the Design of Compose With AI

The development of the *Compose With AI* platform is grounded in several interconnected theoretical frameworks that inform its design and implementation, providing the foundation for understanding how students can effectively and ethically integrate AI-generated content into their academic writing. Central to the platform's design is Constructivist Learning Theory (Piaget, 1980), which emphasizes active engagement and personal meaning-making in learning. This theory posits that students build their knowledge base and the worldviews through lived experiences and subsequent reflections on those direct experiences (Bereiter, 1994). In the design of the *Compose With AI* platform, this theory is operationalized by encouraging students to engage actively with AI-generated content, thereby enabling them to construct their own understanding of disciplinary writing. The platform guides students through a process that requires them to interact with, evaluate, and reflect on AI-produced texts, which encourages them

to reconcile new information with their existing knowledge base. This interaction ensures that students are not mere passive recipients of pre-constructed knowledge, but, rather, active participants in the learning process. By requiring students to assess and integrate AI-generated content into their work, the platform embodies the essence of constructivism, allowing students to build on their prior knowledge and modify their understanding in light of new experiences. This approach is intended to not only support the development of critical thinking skills and scientific literacy but also foster students' understanding of the ethical use of AI in academic writing.

Cognitive Load Theory (Sweller & Chandler, 1991) also informed the platform's design by ensuring that tasks are broken down into manageable segments, reducing cognitive overload. This is achieved through a user-friendly interface and a logical progression of tasks, from crafting prompts to evaluating content, enabling students to focus on learning without becoming overwhelmed. By minimizing extraneous cognitive load, the platform maximizes the potential for meaningful learning.

This evaluation focuses on two primary research questions:

1. What specific strategies within the scaffolded interface enhance or inhibit students' ability to critically evaluate and ethically integrate AI-generated content?
2. To what extent do teachers and students find the platform to have high pedagogical usability for supporting disciplinary writing?

By offering a comprehensive view of the writing *process* rather than just the final product, this pilot study contributes to ongoing debates about writing with generative AI by exploring a scaffolded alternative to AI bans.

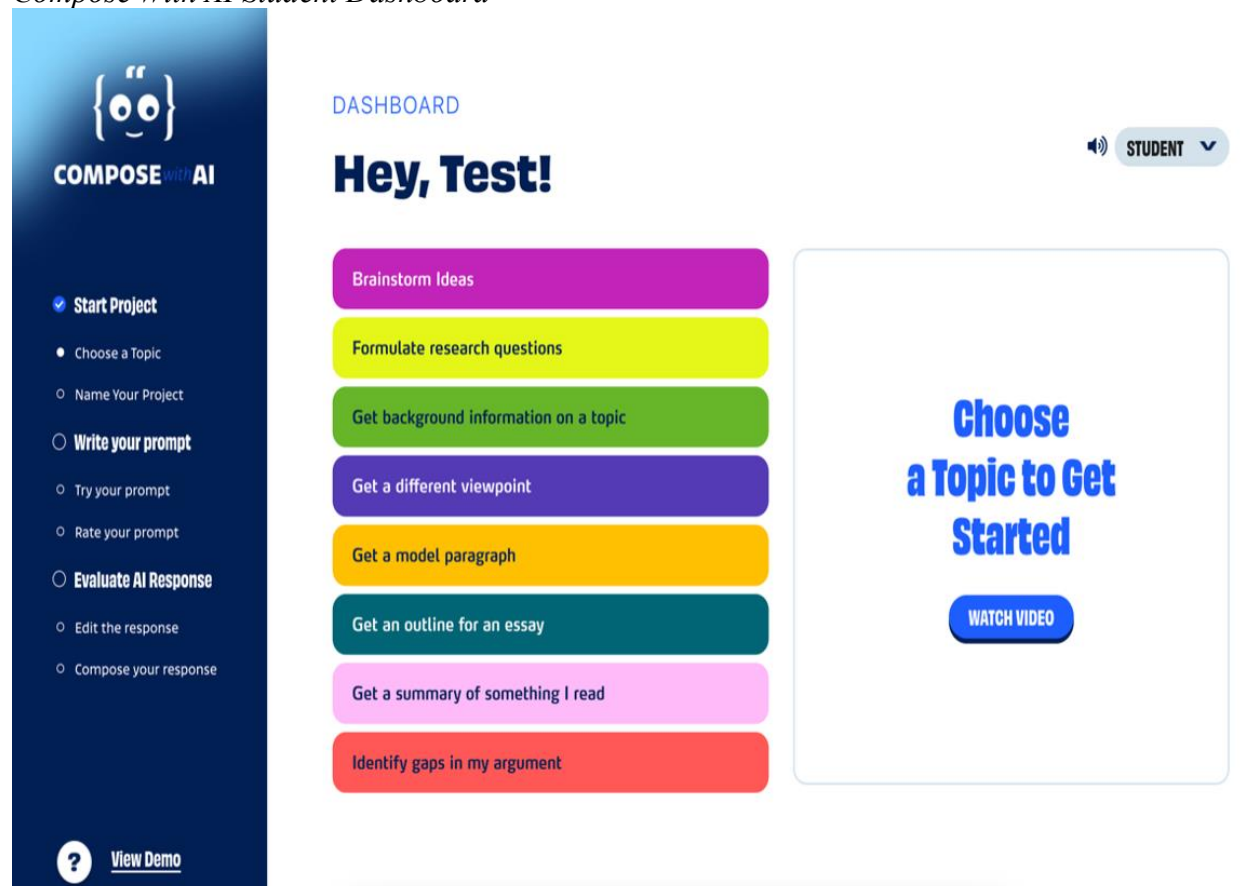
Methodology

Compose With AI

Compose With AI is a web-based platform designed to guide students in grades 4-8 through the process of ethically integrating AI-generated content into their science-focused writing. The platform also includes a teacher dashboard that enables teachers to track all the steps students take in the platform (see Figure 7). Once logged in, as shown in Figure 1, students are greeted with a dashboard showcasing various writing tasks and AI integration options. The main menu presents a range of topics aligned with disciplinary writing (see Figure 1). Each option is accompanied by a brief description and an instructional video, providing clear guidance on each task-type.

Figure 1

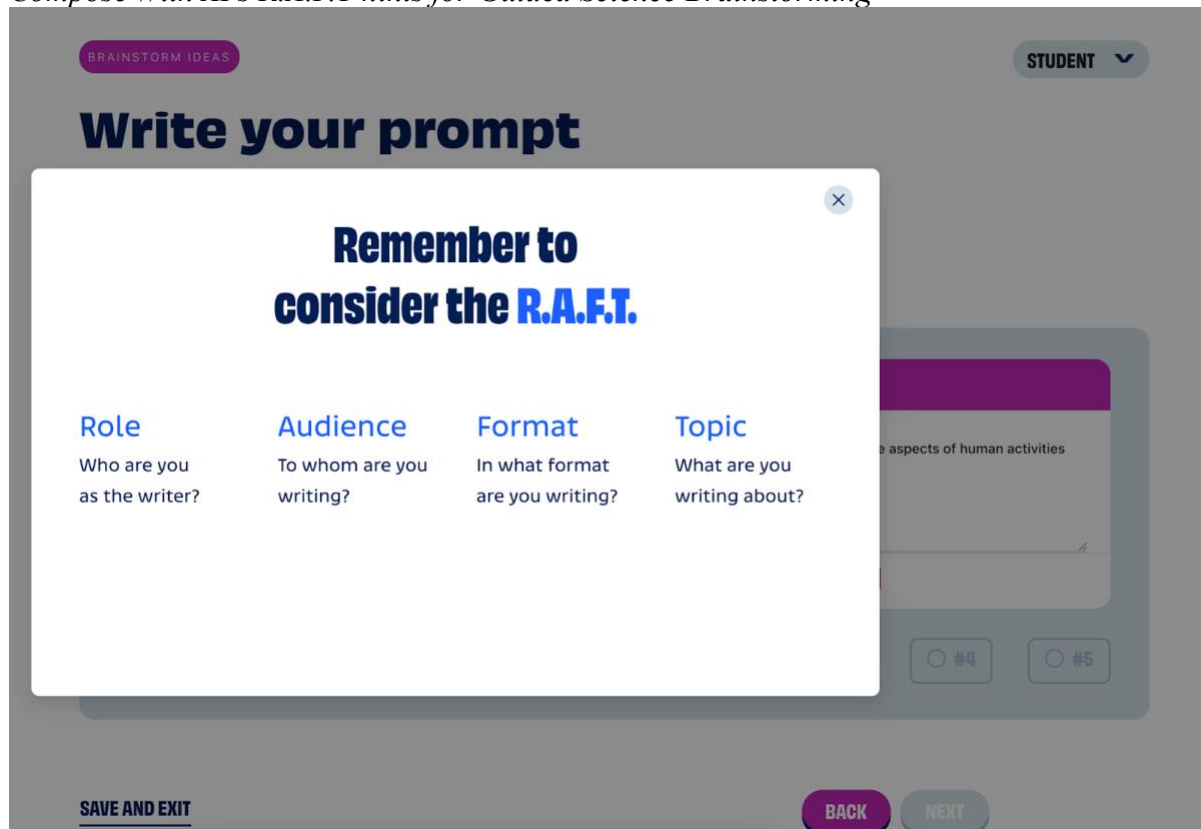
Compose With AI Student Dashboard



For instance, when students click the “Brainstorm Ideas” option, *Compose With AI* employs a structured approach to help students craft effective AI prompts to effectively brainstorm science focused ideas. The platform guides students through the R.A.F.T. technique (Role, Audience, Format, Topic), helping them craft specific and relevant queries relevant to brainstorming as illustrated in Figure 2.

Figure 2

Compose With AI’s R.A.F.T hints for Guided Science Brainstorming



Students can test their prompt by using an integrated feature that copies it from the Compose With AI platform to ChatGPT or another preferred platform, which can be changed in the Teacher Dashboard. Researchers chose not to integrate ChatGPT directly due to school restrictions on certain platforms. Once an AI response is generated, Compose With AI guides students through a critical evaluation process, prompting them to assess the content for accuracy and relevance, as shown in Figure 3. At this step, embedded short videos modeled how to apply

strategies such as lateral reading and bias checking before students made keep/change/omit decisions. Students must apply at least two evaluation strategies and report their findings before progressing to the next step.

Figure 3
Compose With AI's Critical Evaluation Dashboard

FORMULATE RESEARCH QUESTIONS

STUDENT ▾

Evaluate The AI Response

EVALUATE THE CREDIBILITY, USEFULNESS AND ACCURACY OF THE AI RESPONSE

Select at least (2) methods

Read each sentence by itself and check for accuracy WATCH VIDEO	Compare the facts in the response to other sources (lateral reading) WATCH VIDEO
Does the response contain any bias? WATCH VIDEO	Do you understand all the words? WATCH VIDEO
Did the AI tool provide reputable sources to support the response? WATCH VIDEO	Does the response make sense to me? WATCH VIDEO
Does the response cover all aspects of the prompt? WATCH VIDEO	Is the response appropriate for my audience? WATCH VIDEO

[SAVE AND EXIT](#) [BACK](#) [NEXT](#)

After critically evaluating the AI-generated content for credibility, usefulness, and accuracy, students are advanced to a screen where they select which information they can use but will modify, which information they have checked and will use as it is, and which information they will not use (see Figure 4).

Figure 4
Dashboard for Selecting Information to Use in Response.



COMPOSE *with AI*

- ✓ **Start Project**
 - Choose a Topic
 - Name Your Project
- ✓ **Write your prompt**
 - Try your prompt
 - Rate your prompt
- ✓ **Evaluate AI Response**
 - Edit the response
 - Compose your response


[View Demo](#)

Information & text I will keep, change, or omit

[REVIEW THE AI RESPONSE](#)

Reviewing for accuracy and finding areas to add your own insight is important when working with AI tools. Use the boxes below to help form your response by adding the appropriate info to each box. Click 'Save My Entries' before moving on to the next step.

AI Response

Flamingos are fascinating birds with a range of unique physical features and behaviors that contribute to their survival in their specialized environments. Here's a detailed overview of how these adaptations help them thrive:

Physical Features Long Legs and Neck: Flamingos have long, slender legs and necks, which are adapted for wading in shallow water and reaching food in mud or water. Their legs allow them to stand in deeper waters while keeping their bodies above the surface, and their necks are flexible enough to enable them to feed on small organisms in the water.

Specialized Beaks: Flamingos possess a distinctive beak with a unique filtering mechanism. Their beaks are bent downward and have specialized lamellae (tiny comb-like structures) that help filter out mud and small organisms from the water. This adaptation is crucial for their diet, which primarily consists of algae, brine flies, and other microorganisms.

Coloration: The pink or reddish coloration of flamingos comes from pigments called carotenoids in their diet. The coloration serves as a visual signal for mate selection during the breeding season. The brighter and more vibrant the plumage, the more attractive they are to potential mates. This color also provides some camouflage in their natural habitats, blending with the reddish hues of algae and the shallow waters they

Information I can use, but will change

Add the info here...

Information I have checked and can use as-is

Coloration: The pink or reddish coloration of flamingos comes from pigments called carotenoids in their diet. The coloration serves as a visual signal for mate selection during the breeding season. The brighter

Information I will not use

Add the info here...

SAVE AND EXIT

BACK

NEXT

Finally, students are guided in composing a response, and response options vary based on which type of task they selected at the beginning. Figure 5 shows a response example that includes using the information they selected to keep or omit to write a claim and provide evidence. Students are also prompted to cite their sources, including AI.

Figure 5
Compose With AI Screen for Keeping, Changing or Omitting AI-Generated Content

The image shows the 'Compose With AI' interface. On the left is a dark blue sidebar with a logo of a face in curly braces and the text 'COMPOSE with AI'. Below the logo is a list of steps: 'Start Project' (checked), 'Choose a Topic', 'Name Your Project', 'Write your prompt' (checked), 'Try your prompt', 'Rate your prompt', 'Evaluate AI Response' (checked), 'Edit the response', and 'Compose your response'. At the bottom of the sidebar is a 'View Demo' button. The main area is divided into three sections. The top section has three tabs: 'Information I can use, but will change', 'Information I have checked and can use as-is', and 'Information I will not use'. The middle section contains a text box with the prompt: 'Astrophysicist: I study celestial bodies and the universe's physical properties.' The bottom section has a 'SAVE AND EXIT' button. The right-hand panel has four sections: 'My claim (topic sentence)' with a text box containing 'today I found jobs that involve space exploration. such as Aerospace Engineer: I design spacecraft and rockets that launch missions into space.' and a 'Save' button; 'Evidence (key details)' with a text box containing 'Add the info here...' and a 'Save' button; 'Summary' with a text box containing 'Add the info here...' and a 'Save' button; and 'Sources (including AI)' with a text box. At the bottom right are 'BACK' and 'NEXT' buttons.

For teachers, *Compose With AI* tool offers a detailed teacher dashboard. This feature allows teachers to monitor student progress, view ongoing projects, and identify areas where additional support may be needed (Refer to Figure 6). Teachers can also review students' processes and assess the evolution of their prompts. The dashboard presents an overview of class activity, individual student work, and options for providing feedback. The most important feature of the teacher dashboard is that teachers can see the student's entire process, including the prompt they used, prompt revisions, evaluation strategies, information they opted to keep or omit, and the content that was generated by AI. This feature enables teachers to evaluate students' work based on the entire writing process rather than a singular written product at the end. It also empowers students to be transparent with how they used AI. The dashboard includes a built-in rubric that teachers can use to evaluate the students entire thinking process alongside the final written product (see Figure 7).

Figure 6
Compose With AI Teacher Dashboard

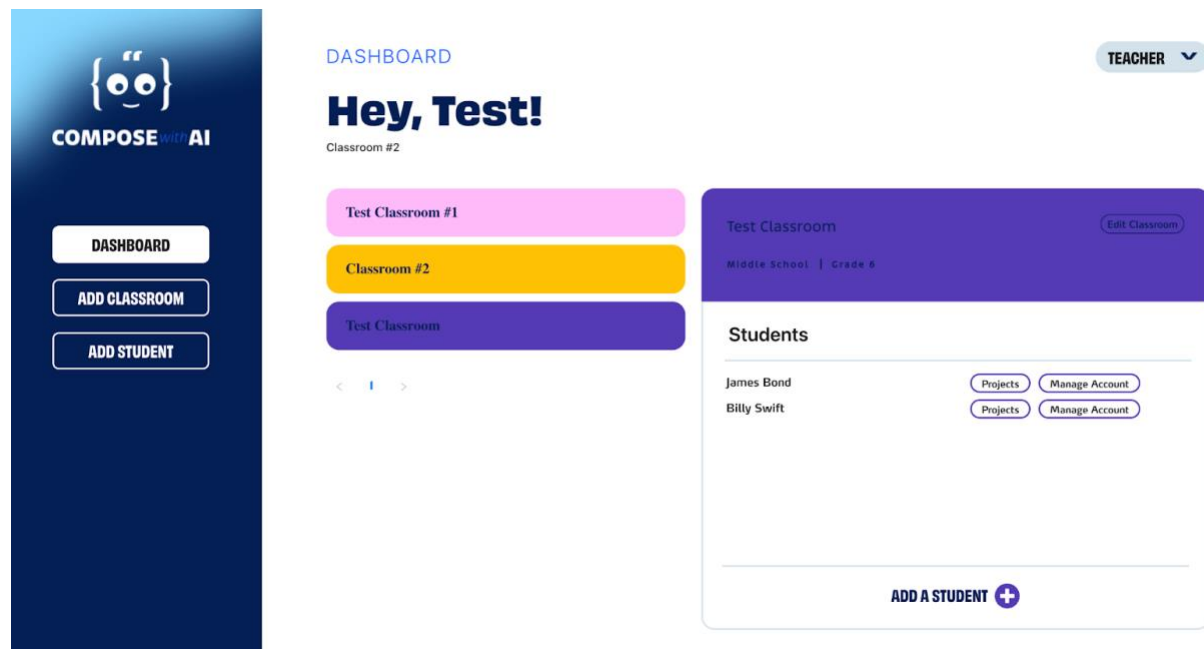


Figure 7
Rubric included in the Teacher Dashboard.

Evaluate the student's project using the following criteria:	NOT AT ALL (1)	SOME (2)	NAILED IT! (3)
I used the RAFT strategy to make purposeful changes to my prompt to better align with my goal(s) or question(s)	☐	☐	☒
My FINAL prompt matches my goal(s) or question(s)	☐	☐	☒
I used strategies to evaluate the accuracy and relevance of the content generated by AI Score as follows: Used 0-1 strategy (1); Used 2 strategies (2); Used 3 or more strategies (3)	☐	☐	☒

TOTAL SCORE 9
SUBMIT

Evaluate the student's project using the following criteria:	NOT AT ALL (1)	SOME (2)	NAILED IT! (3)
I showed which AI information I could use and not use	☐	☐	☒
I showed which facts I checked	☐	☐	☒
I have a strong topic sentence	☐	☒	☐
I provided evidence or key details related to my topic	☐	☒	☐
I cited my sources (including AI)	☐	☐	☒

TOTAL SCORE 9
SUBMIT

Research Design, Data Collection and Analysis

We followed a multi-phase Design-Based Implementation Research (DBIR) process to evaluate the pedagogical usability of the platform and its features (Fishman et al., 2013), analyzing data as follows: (1) *design phase* (i.e., collected and analyzed teacher interview data

about the platform design as the platform was being developed), (2) *delivery phase* (i.e., collected and analyzed student think-aloud data using verbal protocol analysis), (3) *refinement phase* (i.e., refined the platform based on delivery phase data and evaluated our revisions through qualitative analysis of data collected through a second round of student think-alouds, and (4) *retrospective analysis phase* (i.e., after the second round of student think-alouds, all data were combined and analyzed qualitatively).

To make the DBIR design process auditable and transparent, Table 1 presents a design trace summarizing the major iterations of the platform, the empirical evidence that motivated each change, and the observed consequences in subsequent data collection phases.

Table 1

Design Trace: Major Iterations of the Compose With AI Platform

Phase / Version	Design Change	Empirical Rationale	Observed Consequence
Design Phase (v1.0)	Initial platform build: open-ended evaluation strategy selection; all strategies available simultaneously; embedded videos optional and unprompted.	Teacher interviews (design phase) indicated students in grades 4–8 have limited experience evaluating digital content; teachers anticipated students would need more explicit structure to make meaningful strategy choices.	Informed initial design refinements prior to Round 1 student think-alouds; established the baseline against which student behaviors were assessed.
Refinement Phase (v1.0 → v1.1)	Strategy selection revised to require a minimum of two choices; students could no longer proceed without selecting and applying at least two evaluation strategies.	Round 1 think-alouds showed students applying all available strategies indiscriminately, suggesting unlimited choice produced overload rather than meaningful engagement (consistent with Cognitive Load Theory).	Round 2 think-alouds showed more deliberate, contextually appropriate strategy selections; requiring a minimum appeared to reduce indiscriminate use without eliminating student agency.

Phase / Version	Design Change	Empirical Rationale	Observed Consequence
Refinement Phase (v1.1 → v1.2)	Embedded instructional videos repositioned to appear more prominently at key decision points (strategy selection and keep/change /omit screens); visual cues added to draw attention to video availability.	Both rounds showed students bypassing videos without prompting, even when uncertain; students who watched videos (when prompted) demonstrated better understanding of the targeted strategy.	Video engagement remained low; identified as a persistent design challenge requiring more structural integration (e.g., required micro-viewing before progression) in future iterations.
Retrospective Analysis Phase	19 initial codes consolidated to 13 final codes; themes mapped to Constructivist Learning Theory and Cognitive Load Theory to ground interpretive claims in established frameworks.	Cross-phase analysis revealed no contradictions between concurrent and retrospective verbal reports; themes were consistent across both rounds of data collection.	Four overarching themes identified (step-by-step guidance, required evaluation, persistent visual display, hesitance toward authoritative AI output) directly informed implications for future design and research.

During the think-aloud verbal protocol, students were presented with a 45-60 minute think-aloud task that was preceded by a short tutorial on the platform and the evaluation strategies suggested in *Compose With AI* (e.g. lateral reading, check for bias). During the task, students had access to videos and tips built into *Compose With AI* to explain the potential uses of AI and evaluation strategies provided within the platform. We utilized a verbal reporting methodology (Ericsson & Simon, 1980, 1984/1993; Fox, Ericsson, & Best, 2011; Pressley & Afflerbach, 1995) to gain access to students' cognitive processes both during the task (concurrently) and after the task (retrospectively). For the task, students: 1) were guided to practice thinking aloud before starting to ensure they understood think-aloud expectations, 2)

received a prompt (e.g., “Choose one animal (for example: a penguin). How does this animal’s physical features or behavior help them survive in their environment?”), and 3) followed the steps in *Compose With AI* while thinking aloud about their selections. During the task, we recorded students’ screens as they worked and prompted them to think aloud about how they are navigating the task and using the strategies provided by the *Compose With AI* platform. Immediately following the task, we asked students to comment on specific aspects of the task that the researcher wanted to probe further.

In addition to student testing, three teachers reviewed the platform and participated in semi-structured, one-on-one interviews. Teachers provided feedback on the platform design and the revisions to the *Compose With AI* platform after the initial round of student think-alouds. During the interviews, we asked teachers three open-ended questions about: (1) the features they thought might work well for students and support students’ learning, (2) the features that might be difficult for students to use, find challenging, or may inhibit their learning, and (3) features they might add to better support students’ learning.

Data Analysis

Data analysis involved a two-stage process focusing on concurrent and retrospective verbal protocols. Students’ concurrent think-aloud protocols, collected during the task, underwent qualitative thematic analysis (Braun & Clarke, 2006). Researchers first reviewed all transcripts and screen recordings to thoroughly familiarize themselves with the data. Next, think-aloud data were organized by episodes within task segments, with each screen that students worked on in the platform serving as a task segment. Within each task segment, researchers analyzed the data for episodes demonstrating how components of the task or platform inhibited or enhanced students’ ability to complete the task. Episodes were then given initial codes to

indicate the inhibiting or enhancing factors. This process resulted in 19 codes, which were then evaluated to determine patterns and frequency within the data, which led to 13 final codes that were consistent across participants. Next, the codes were mapped onto students' actions within the *Compose With AI* platform, as observed in the screen recordings, to establish a clear link between verbal reports and observable behaviors. Finally, the codes were viewed through the lenses of Constructivist Learning Theory and Cognitive Load Theory as a way to group them into the themes. This process is shown in Table 2. The following themes reflect students' cognitive processes and strategies: importance of step-by-step guidance, importance of requiring evaluation, importance of persistent visual display, and hesitance to question authoritative sources.

Table 2

Sample Codes, Theoretical Lenses, and Rationale for Theme Groupings

Theme	Sample codes	Theoretical lens	Rationale for grouping
<i>Importance of step-by-step guidance</i>	• Reliance on structured sequence • Confusion when steps are skipped or unclear • Positive affect when progression is clear • Request for explicit next-step direction	Cognitive Load Theory	Sequential task segmentation reduces extraneous cognitive load, freeing working memory for the core evaluative task. Students' verbal reports indicated that the step-by-step structure reduced confusion and allowed them to focus on content rather than navigation.
<i>Importance of requiring evaluation</i>	• Selection of evaluation strategies • Cross-referencing AI output with outside sources	Constructivist Learning Theory	Requiring students to actively evaluate AI-generated content operationalizes constructivist principles. Students reconcile new information with existing knowledge rather than

Theme	Sample codes	Theoretical lens	Rationale for grouping
	• Noting discrepancies between AI output and prior knowledge • Verbalizing criteria for accepting or rejecting AI content		passively accepting output. Think-aloud data showed students constructing understanding through evaluation (e.g., lateral reading, bias checking).
<i>Importance of persistent visual display</i>	• Reference to previously displayed information • Reduced re-reading or backtracking behavior • Verbal acknowledgment of on-screen decision history • Leveraging visible keep/change/omit selections in final response	Cognitive Load Theory	Persistent display of prior decisions functions as an external memory aid, reducing germane cognitive load by offloading recall demands. Screen recordings confirmed students consistently referenced the persistent display rather than attempting to reconstruct prior choices.
<i>Hesitance to question authoritative sources</i>	• Deference to AI output without verification • Expressed uncertainty about ability to evaluate AI content • Reluctance to modify or omit AI-generated information	Constructivist Learning Theory	From a constructivist perspective, meaningful learning requires students to actively interrogate new knowledge against their existing schemas. Hesitance to question AI represents a failure to engage in this negotiation.

Immediately following the task, students participated in a brief, semi-structured interview to probe specific aspects of their task performance. These retrospective verbal reports were analyzed to determine if any contradictions existed between what was observed between the

concurrent and retrospective verbal reports. However, the focus of the retrospective recalls was on explicitly identifying students' reflections on their strategies, challenges encountered, and overall experience with the platform. No contradictions were found between the concurrent and retrospective verbal reports, but the retrospective reports provided additional support for the themes that had already been identified.

Teacher data were analyzed similarly, using an open coding approach (Strauss & Corbin, 1998) to identify emergent themes and patterns reflecting teachers' perspectives. Researchers began by familiarizing themselves with the transcripts through multiple readings to gain a thorough understanding of the content. Then, we employed open coding to identify meaningful segments of text, which were then assigned 11 individual codes that captured key ideas or themes. We derived codes inductively from the data, without reliance on pre-existing categories or frameworks. Finally, through iterative comparison, we grouped codes into related categories, which in turn, led to the identification of overarching themes.

Authors collaboratively analyzed the data to ensure credibility and shared interpretation. Triangulation was achieved by comparing and cross-checking themes across all qualitative data sources including student think-alouds and teacher interviews. Inter-rater reliability was high (Cohen's $\kappa > .80$), and any discrepancies were resolved through discussion until full consensus was reached. This collaborative process strengthened the coherence and trustworthiness of the findings.

Participants

Participants included four students, all female, consisting of two students in grade five, one student in grade six, and one student in grade seven. None of the participants had ever used ChatGPT before and reported only hearing about AI, but not knowing how to use it or what it is.

The teachers' backgrounds were as follows: (1) a female fifth-grade teacher in a self-contained elementary school classroom, with 20 years of teaching experience; (2) a female seventh- and eighth-grade English teacher, with 17 years of teaching experience; and (3) a male eighth-grade technology teacher, with 21 years of teaching experience. This small, exploratory sample (four students in grades 5-7 and three teachers) was intended to generate early design insights rather than generalizable findings.

The *Compose With AI* platform was implemented only for pilot research and was not connected to any external AI APIs. Students copied prompts manually into a *Compose With AI* interface to avoid transmitting any personal or learning data to third-party cloud services. All student work (prompts, revisions, and rationales) was stored on secure servers and was accessible only to the research team. For research purposes, all datasets were pseudonymized, stored in password protected computers.

Results

Design Elements That Supported Students' Critical Engagement with AI-Generated Content

Results from the student think-aloud sessions highlighted that one of the most beneficial features of *Compose with AI* is its structured step-by-step guidance. Students are led through the following sequential steps: (1) crafting an initial prompt, (2) refining the prompt with guided questions, (3) testing the prompt in ChatGPT, (4) assessing the relevance of the generated content and revising the prompt if needed, (5) verifying the credibility and accuracy of the AI output, (6) deciding which AI-generated content should be modified, omitted, or used as is, and (7) formulating a response. This step-by-step flow was highly valuable for students, with users reporting that it made the process feel easy and fun. For example, Andrea, a fifth grader, stated, "For a class assignment. I mean, it works really well and it's really fun, I do a nine [on a

scale of 1-10]. Well, it depends. If you've never done it before and your teacher's telling you to do it, give it a nine. But if you've done it before it's a lot easier. I'd give it a 10." None of the participants had previously been taught how to evaluate online content. The researchers observed that the step-by-step process worked well as an introduction to the entire process of searching for information all the way to composing a final response. For users who completed a second task and think-aloud, researchers also observed that the users were able to provide more critical responses on the second task, indicating the platform becomes easier to use with repeated exposure.

One of the helpful structured steps was the guidance provided to students for generating a prompt. Melissa, a fifth grader, was asked to use the *Get a Model Paragraph* selection to write a paragraph about how penguins' physical features or behaviors help them survive. At this step in the platform, students are guided through a series of five steps to generate and refine their prompt. Figure 8 shows the prompt guidance that students receive during the prompt generation phase and how Melissa refined her prompt each time she received additional guidance. As can be seen, Melissa's prompt improved in specificity as she received additional prompt guidance. When Melissa tried her final prompt and generated a response with ChatGPT, her face became animated and she said "That's cool! I like the way it just gave me information, the AI thing did." In her retrospective recall, Melissa stated, "In the beginning, I didn't quite know what a prompt was, but then I understood," indicating that the platform itself helped her learn as she went along.

Figure 8

Example of How A Student Prompt Changed At each Step of the Prompt Guidance

Prompt Guidance	Prompt Provided by Student
#1. Determine the most important keywords. EXAMPLE Give me ideas about the positive and negative aspects of human activities in a local ecosystem.	How does a penguin's physical features or behavior help them survive?
#2. Check to see if your prompt includes a specific format. EXAMPLE Write an outline about the positive and negative aspects of human activities in a local ecosystem.	Write a summary about how a penguin's physical features or behaviors help them survive?
#3. Determine if you want to ask AI to write from a specific point of view. EXAMPLE Write an outline about the positive and negative aspects of human activities in a local ecosystem from an activist's point of view.	Write a summary about how a penguin's physical features or behaviors help them survive? Write it like a kid writing to their teacher.
#4. Check to see if your prompt includes a specific length. EXAMPLE Write a 2 page outline about the positive and negative aspects of human activities in a local ecosystem from an activist's point of view.	Write a one paragraph summary about how a penguin's physical features or behaviors help them survive? Write it like a kid writing to their teacher.
#5. Determine if you need to include additional details or relevant context. EXAMPLE Write an outline about the positive and negative aspects of human activities in a local ecosystem from the point of view of an activist who is recruiting other people to volunteer to help reduce impact on the local	Write a one paragraph summary about how a penguin's physical features or behaviors help them survive? Write it like a kid writing to their teacher. Add fun facts.

Another crucial feature for enhancing students' ability to critically engage with AI content was the platform's integrated evaluation strategies with embedded videos explaining each strategy (Figure 3). Initially, students could select multiple strategies; however, this was revised to require only two selections after the first round of think-alouds showed students attempting to use all methods, regardless of relevance.

Even the youngest participants found the evaluation strategies to be useful. For example, fifth-grader Andrea, tasked with researching space-related careers, used "Compare facts to other sources" (lateral reading) and "Do you understand all the words?" She cross-referenced the AI-generated content with the Bureau of Labor Statistics, validating some careers but questioning others, including "space architect," due to its unfamiliar context. She commented: "Aerospace engineer was on the AI thing, so I know that one's right. And I didn't see any of the other ones, so I don't really know." Melissa continued searching what she deemed to be reputable sites and found several of the careers that were on the list generated by ChatGPT. Before moving to the next evaluation strategy, she decided to note "I found some, but not all of these. No space architect" in the box that required her to indicate the result of the evaluation strategy she used (see Figure 9). Andrea also commented on the use of the word architect in the list generated by ChatGPT. She was unable to find "space architect" on any of the sites she found when she did her parallel reading. She noted that she knows the meaning of the word architect and she doesn't think it applies to space.

Figure 9

Example of Strategy Provided for Evaluating an AI Response

Evaluate The AI Response

AI Response

Astronaut: I get to travel to space and conduct experiments firsthand.
Aerospace Engineer: I design spacecraft and rockets that launch missions into space. Astrophysicist: I study celestial bodies and the universe's physical properties.
Planetary Scientist: I analyze planets and moons to understand their characteristics. Mission Control Specialist: I manage and monitor space missions from Earth. Satellite Technician: I build and maintain satellites that orbit Earth and other celestial bodies. Space Biologist: I research how space environments affect living organisms. Space Architect: I design habitats and structures for use in space.

Compare the facts in the response to other sources (lateral reading)

WATCH VIDEO

☐

The facts are comparable to other sources

☒

Some facts are different compared to other sources

Notes

I found some, but not all of these. No space architect.

CANCEL

SAVE METHOD

As students moved through the platform, the page that asked students to evaluate whether to keep, change, or omit AI-generated information (shown in Figure 4) was shown to enhance students' abilities to critically evaluate the content they generated. As shown in Figure 4, the AI-generated content is shown on each screen as the student progresses through the platform. Students demonstrated strong reasoning skills at this step. For example, Andrea chose to dismiss the information about space architects, arguing that, "... Architects make buildings, like they design buildings, and a spaceship is not a building. I'm not sure about this one." Thus, she

copied information about space architects into the box labeled “Information I will not use.” She also reasoned that her parallel reading led her to find physicists, but not astrophysicists. Thus, she decided to copy this information into the box labeled “Information I can use but will change.” Finally, she copied several careers that she had found in her web search into the box labeled “Information I have checked and can use as-is.” On the whole, this step in the writing process required students to sort all of the AI-generated information into categories. As shown through Andrea’s example, students’ responses at this phase required them to engage in critical reasoning and construct their own understanding of the AI response, setting the stage for them to exercise caution in how they integrated AI-generated content into their final composition.

Finally, researchers observed that the persistent visual display, one of the features designed to reduce cognitive load, appeared to be an enhancing feature. Specifically, the platform provides a persistent visual display of the students’ decisions and information that follows them throughout the platform. For example, as Figure 5 shows, the information students have chosen to keep, change, or omit is visually displayed on the screen as users compose their final response. This constant visibility acted as a reminder and a guide, reducing the need to recall previous decisions. The think-alouds revealed the usefulness of this feature at several points. For example, during the final stage of her second task, Martina, the sixth-grade participant, remarked “I mean, for the last question, I could just copy this and put it in the summary, because I have already decided on this information..” This remark is indicative of her awareness of the persistent display of prior decisions and her inclination to leverage it for task completion. Similarly, researchers observed students consistently referencing the persistent visual information throughout all phases of the task.

All student users reported the platform to be useful and stated that it was either “fun” or “cool” to use *Compose With AI*. All student test users scored the overall experience of using the platform as either a 9 or 10 (on a scale of 1-10).

Approaches and Strategies that Inhibited Students’ Ability to Critically Evaluate AI-Generated Content

The primary barrier to students’ progress on the platform was that some of the steps were not self-explanatory. Although *Compose With AI* automatically runs prompts in ChatGPT, students’ lack of familiarity with generative AI hindered their willingness to evaluate the content generated by AI. For example, all four participants initially expressed mild resistance to evaluating and modifying the response generated by AI, perceiving the ChatGPT output as authoritative. Two of the students even reported feeling unqualified to make changes to the responses.

The platform’s requirement for students to choose at least two strategies to evaluate the credibility and accuracy of their AI-generated content was not intuitive because students had limited prior experience critically evaluating digital content. This lack of familiarity meant that all students required additional guidance throughout the process, as the design of *Compose With AI* assumed familiarity with critical evaluation. One participant expressed an interest in having a virtual guide assist her as she navigated the platform. While the availability of built-in instructional videos may have been beneficial, students did not readily engage with them and often required prompts to watch the videos, suggesting that integrating video resources more effectively could enhance students’ understanding of the platform’s functionality. Requiring users to watch these videos may lead to a better comprehension of the next steps in the process, ultimately improving student’s abilities to critically evaluate content.

Teacher Perceptions of the Pedagogical Usability of *Compose With AI*

Advantages of Integrating Support for Using AI

All three teachers noted the benefits of teaching students to use AI tools ethically as part of the curriculum. Every teacher we interviewed—noted that students are entering their classrooms familiar with AI tools like ChatGPT, and that most students have experience using these tools outside of school. Though, this was not the case with students in our sample. As one teacher remarked: “I haven’t been able to find something like this [*Compose With AI*] that helps kids use AI to their advantage; most kids are just finding things online at home that are like, ‘here’s AI- have fun.’”

All three teachers also noted that they have participated in professional development related to AI tools within the past calendar year, including prior to the start of the 2024-2025 academic year. Teachers reported that these professional development experiences focused on explaining the purpose and function of AI tools, but did not specifically address how to use them to support instruction. As one teacher explained, “This [*Compose With AI*] would be huge for a district that needs a curriculum or a program to teach kids how to use AI effectively.” Notably, none of the teachers worked in districts that banned AI usage in school.

Pedagogical Usability of the Compose With AI Platform

All three teachers noted that *Compose With AI* demonstrated strong pedagogical usability, with teachers noting that its interface design actively supported the learning goals rather than merely being easy to operate. One teacher commented: “The tool is kid-friendly. It keeps information organized into bite-sized chunks that won’t be too overwhelming for students.” This comment reflects more than a navigational preference; it indicates that the interface’s chunked structure directly supported the cognitive scaffolding the platform was designed to provide by reducing extraneous load and allowing students to focus on the evaluative reasoning required at each step. This same teacher further remarked that the graphics and colors were “eye-catching”

and would be appealing to students in the middle grades. She explained, “it [the colors and graphics] made me excited to use it.” Similarly, another teacher remarked that the tool “allowed students to explore AI in a hands-on way,” and noted that the platform provided opportunities to experiment and learn through guided interaction—suggesting the interface itself functioned as a pedagogical scaffold, not simply a content delivery vehicle.

Teachers also remarked on specific aspects of the design. While our student participants did not readily watch the embedded videos, two teachers took note of them. One teacher noted that she loved the video examples and how researchers used AI to create the dialogue. She commented: “It felt very fitting for this project.” Another teacher remarked that the videos were short but clear, and she imagined students could “get the gist of what to do without having to sit there for a long time and getting bored.” Both teachers also expressed appreciation for how the videos aligned with the tool’s primary objective: helping students understand how to use AI effectively and ethically.

A third teacher commented specifically about the “prompting” aspect of the tool. She admitted she had never used AI prompts before and found the step-by-step procedure to be extremely supportive of the process. She noted: “The way it was broken down helped me better think about the types of things you need to say to get the response you’re looking for. This was my favorite part of the tool.” She also noted how the structured approach to prompt engineering could be particularly beneficial to students who are new to working with AI. “I can see how this would be a great way to scaffold students’ learning,” she added.

Comments about Using AI for Writing

The final theme that emerged was about the affordances of using AI-generated content to support students’ writing. The teachers noted that *Compose With AI* provides guidance for many

forms of science writing that are common in the middle grades: narrative and/or informational essays, opinion writing, and inquiry/research reports. One teacher noted that opinion writing is especially difficult for students, and found the “get a different perspective” feature to be particularly useful. She noted that students are “50/50” on their ability to do this, which often hinders their ability to effectively write opinion pieces. “It’s like having a writing coach right there with the kids,” she said about the tool. “They can see how to break down an argument...something that would usually take tons of individual instruction. This gives them that extra one-on-one support.”

All teachers remarked that *Compose With AI* helps students understand that AI-generated tools still require them to have writing skills and craft in order to use these tools effectively. As one teacher explained, “It [*Compose With AI*] opens my eyes to the idea that AI is a great tool and this [the *Compose With AI* platform] is a good way to show students how to apply critical thinking skills to their writing. It can ‘better’ what they come up with, but they still have to use their own analysis to figure that out.” Another teacher noted: “This [*Compose with AI*] really shows that AI isn’t a replacement for the writing process. It’s like AI gives them a starting point, but they have to be in charge of shaping and improving it.”

Discussion

This study explored the design and exploratory evaluation of *Compose With AI*, a web-based platform guiding students (grades 4–8) in ethically integrating AI-generated content into science-focused writing. Our findings reveal key factors enhancing and inhibiting students' abilities to critically evaluate and effectively use AI for academic purposes. These findings demonstrate that a structured and scaffolded approach, such as that provided by *Compose With AI*, offers a promising early proof of concept for fostering responsible AI literacy in disciplinary

writing, though further research is needed to establish broader effects. The study was designed to address the pedagogical need for students to understand how to critically evaluate AI-generated content and engage in “thinking with AI,” which requires skills such as prompt engineering, critical evaluation of output, and integration with authentic thought. Here we highlight implications of our key findings and paths for future research.

Effectiveness of Platform Features in Fostering Critical Engagement

A key finding is the significant value students placed on the platform's structured step-by-step guidance. This sequential approach, moving from crafting prompts to evaluating content and finally composing a response, was reported by students as making the process feel "easy and fun," even for those with no prior AI experience. This finding provides initial support for the idea that breaking down disciplinary writing tasks that integrate AI may reduce cognitive load, perhaps enabling students to produce compositions that better reflect their own voice and thoughts. The progression through distinct stages may have minimized extraneous cognitive load, enabling students to engage more effectively with the core task of evaluating and integrating AI content.

Prompt engineering, identified as foundational to critical engagement with AI, was explicitly supported by Compose With AI through guided questions helped students craft more specific and relevant queries, and students like Melissa demonstrated improved prompt specificity through this iterative guidance. This process not only helped students generate better content but also helped them learn *what* a prompt was and how to refine it, indicating that structured tools may be effective for teaching complex skills like prompt engineering. This is an important finding since pedagogical approaches for teaching prompt engineering are still relatively unexplored.

Beyond prompt creation, the platform facilitated critical evaluation through integrated evaluation strategies. Although students initially struggled with selecting strategies and did not always engage with the videos, the requirement to apply at least two strategies pushed them to consider different ways of assessing content. This finding reinforces the necessity of teaching traditional source evaluation skills, such as assessing credibility and accuracy by comparing information across multiple sources, even when working with AI. Further, our results directly address concerns about the potential for AI to hallucinate or generate misinformation (Alasadi & Baiz, 2023; Su & Yang, 2023) by providing an approach for students to critically evaluate AI outputs and assess information accuracy and reliability. Accordingly, *Compose With AI* provides a foundation for addressing concerns that researchers and educators should adapt the ecosystems used in science fields to leverage generative AI's benefits while mitigating potential harms (Qadir, 2023).

Crucially, the "keep, change, omit" function appeared to enhance students' critical evaluation abilities. This step required students to actively sort AI-generated information based on their assessment of its accuracy and relevance, engaging in critical reasoning and constructing their own understanding. Andrea's example of dismissing "space architect" information because "a spaceship is not a building" and deciding to change information based on her parallel reading demonstrates the sophisticated level of critical thought this feature elicited. This function effectively scaffolded the integration of AI-generated text with students' authentic thoughts, requiring them to decide how AI output aligned with their task and existing knowledge before incorporating it. This entire negotiation process also echoes the constructivist notion that learning is an active endeavor where students negotiate meaning and understanding by

interacting with content, thereby facilitating deeper learning and retention of knowledge (Ültanır, 2012).

The persistent visual display of AI-generated content and students' decisions (keep, change, omit) throughout the platform was another beneficial feature. It reduced the need for students to remember previous choices, potentially minimizing cognitive load and allowing them to leverage prior evaluations when composing their final response.

Compose With AI also addresses unique challenges in disciplinary writing, particularly in scientific writing where students are expected to master technical vocabulary, follow specific formats, and develop skills in argumentation and evidence-based reasoning. Disciplinary literacy literature underscores the importance of genre-specific scaffolding (e.g., Graham & Harris, 2006), which is critical in STEM fields where students are expected to learn to communicate complex ideas clearly. We believe that *Compose With AI* has potential to support students in improving academic writing by helping students get genre-specific models for the types of writing they need to produce in school settings.

Challenges Encountered

Despite the successful scaffolding, the study also highlighted challenges. Students initially perceived AI output as authoritative and expressed mild resistance to evaluating or modifying it, sometimes feeling unqualified to make changes. This finding suggests that students may lack prior experience critically evaluating *any* digital content, and the perceived intelligence or complexity of AI may exacerbate this reluctance. Furthermore, AI-generated content presents unique evaluation challenges, such as its dynamic nature and lack of human-authored context, which can make discerning inaccuracies difficult despite a veneer of authenticity. These findings demonstrate that explicit instruction on why and how to question AI output is essential. The

requirement to select evaluation strategies was not intuitive due to students' limited prior experience with digital content evaluation. While videos were available, students did not readily engage with them. These findings indicate a need for more explicit guidance when working with the platform, potentially requiring students to interact with instructional content or integrating support more seamlessly into the workflow.

Teacher Perspectives on Pedagogical Usability

Teachers interviewed unanimously recognized the benefits of teaching students to use AI tools ethically, acknowledging that students are already encountering AI outside of school. They noted a lack of professional development on *how* to use AI to support instruction and viewed *Compose With AI* as a valuable resource to fill this gap.

Teachers found the platform to have strong pedagogical usability, reporting that its interface design actively supported the disciplinary learning goals rather than functioning as a neutral delivery mechanism. They appreciated its organized, “bite-sized” approach, interpreting the chunked structure not merely as user-friendly design but as a form of cognitive scaffolding that allowed students to focus attention on evaluative reasoning at each step. They saw the prompt guidance as particularly beneficial for students new to AI. Notably, teachers articulated the platform’s pedagogical value through the metaphor of the “writing coach”—the idea that the interface itself took on an instructional role by structuring students’ thinking about argument, perspective, and evidence in ways that would otherwise require substantial individual teacher attention. Importantly, teachers felt that *Compose With AI* effectively demonstrated that AI is a tool to support, not replace, the writing process, requiring students to apply their own critical thinking and analysis to shape the final product. The positive teacher feedback supports the platform’s potential for effective AI integration in disciplinary writing instruction and offers a

novel approach to evaluating students' writing processes, providing teachers with a comprehensive view of students' entire writing journey with AI, rather than just the final product.

By focusing on the entire writing process rather than just the final product, we believe the teacher dashboard addresses concerns about cheating, plagiarism, and/or over-reliance on AI, promoting accountability and ethical AI usage. Our findings highlight the importance of creating instructional strategies to teach students about responsible uses of AI rather than simply banning students from using it. This approach aligns with the broader educational discourse on integrating AI in classrooms effectively and ethically, as noted in studies by Luckin et al. (2022) and Holmes et al. (2022), which emphasize the necessity of equipping students with the skills to navigate AI technologies responsibly.

Implications for Practice, Policy, and Future Research

The findings have several implications. First, this pilot study offers preliminary indications that explicit, scaffolded instruction may support students in developing the AI literacy skills needed to evaluate and ethically use AI-generated content. Platforms like Compose With AI represent one approach to delivering this instruction, guiding students through complex processes like prompt engineering and critical evaluation, though larger studies are needed to confirm these patterns. Educators should focus on teaching students to think "with" AI by emphasizing the iterative nature of prompt refinement and the necessity of critically assessing and integrating AI output with their own thinking. The teacher dashboard feature, which allows monitoring of the entire student process (prompts, revisions, evaluations, choices), supports a pedagogical shift towards evaluating the *process* of AI integration, not just the final product, fostering transparency and accountability. This study also supports the argument that restrictive AI bans may be less productive than approaches that equip students with skills for ethical and

effective AI integration. Educational policies should encourage the development and implementation of tools and curricula that build AI literacy, rather than simply preventing access.

Further research is needed to explore how to best address students' initial reluctance to question AI output. Investigating pedagogical strategies or platform features that build students' confidence in evaluating digital content, including AI, is essential. Exploring how to integrate instructional videos or other support mechanisms more effectively into the workflow could enhance student engagement with critical evaluation strategies. Future studies could also examine the long-term impact of using platforms like Compose With AI on students' AI literacy development across different disciplinary contexts and grade levels, and how these tools might support more sophisticated forms of "thinking with AI."

Limitations

The small participant sample limits the generalizability of the findings. Accordingly, this study should be viewed as an exploratory design-based pilot that provides an initial proof of concept for supporting science-focused writing with AI. While the sample is small, the use of multiple, complementary data sources--including student screen recordings (behavioral evidence), concurrent think-aloud protocols (real-time cognitive evidence), retrospective interviews (reflective evidence), and teacher interviews (practitioner evidence)--provided a multi-faceted view of how the platform functioned in practice and strengthened the credibility of the findings within the limits of this pilot. That said, because no quantitative learning outcomes were collected and the study did not include a comparison or control condition, conclusions are descriptive and process-oriented, reflecting students' and teachers' observed interactions rather than measured effects. Future research should include larger and more diverse samples across

grade levels, demographic contexts, and school settings to strengthen external validity and examine differential impacts. In addition, longer-term or classroom-embedded studies are needed to evaluate the durability of learning gains and the evolution of students' evaluative reasoning over time. Future design iterations could embed required micro-videos, progress indicators, and lightweight analytics (e.g., counts of prompt iterations and distribution of keep/change/omit decisions) to better connect design features with evidence of learning. Finally, students' initial reluctance to evaluate AI-generated text despite guidance indicates a need for continued research on pedagogical strategies that promote confident, critical engagement with AI-produced content.

Conclusion

In conclusion, *Compose With AI* offers a promising, exploratory approach to helping students develop critical evaluation skills for AI alongside disciplinary writing skills. Within this small pilot, students demonstrated emerging abilities to critically evaluate and integrate AI-generated content when supported by structured scaffolds such as prompt guidance and the keep/change/omit workflow. By providing step-by-step support for complex tasks like prompt refinement and content evaluation, the platform illustrates one potential pathway for educators to introduce responsible, transparent AI use in classroom writing activities. These findings are preliminary but point to the value of continued design-based work examining how scaffolded AI tools can cultivate critical and ethical engagement with generative technologies in education.

References

- Alasadi, E., & Baiz, C. (2023). Generative AI in education and research: Opportunities, concerns, and Solutions *Journal of Chemical Education*, 100(8), 2965–2971. <https://doi.org/10.1021/acs.jchemed.3c00323>
- Ardito, C., Costabile, M. F., De Marsico, M., Lanzilotti, R., Levialdi, S., Roselli, T., & Rossano, V. (2006). An approach to usability evaluation of e-learning applications. *Universal Access in the Information Society*, 4(3), 270–283. <https://doi.org/10.1007/s10209-005-0008-6>
- Bennett, L., & Abusalem, A. (2024). Artificial intelligence (AI) and its potential impact on future of higher education. *Athens Journal of Education*, 11(3), 195-212.
- Bereiter, C. (1994). Constructivism, socioculturalism, and Popper's World 3 Educational Researcher, 23 (7), 21-23.
- Bozkurt, A. (2024). Why generative AI literacy, why now and why it matters in the educational landscape? Kings, queens and GenAI dragons. *Open Praxis*, 16(3), pp. 283–290. <https://doi.org/10.55982/openpraxis.16.3.739>
- Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative research in psychology*, 3(2), 77-101.
- Chan, C. K. Y., & Hu, W. (2024). Exploring students' perspectives on generative AI-assisted academic writing. *Education and Information Technologies*, 29, 14481–14510. <https://doi.org/10.1007/s10639-024-12878-7>
- Cu, M. A., & Hochman, S. (2023). Scores of Stanford students used ChatGPT on final exams, survey suggests. *Preuzeto s* <https://stanforddaily.com/2023/01/22/scores-of-stanford-students-used-chatgpt-on-final-exams-survey-suggests>.
- Ericsson, K. A., & Simon, H. A. (1980). *Verbal reports as data*. MIT press.
- Ericsson, K. A., & Simon, H. A. (1984/1993). *Protocol analysis: Verbal reports as data*. MIT press.
- Esmaeili, A., Rahimi, S., & Moradi, M. (2019). The relationship between information literacy and the ability of library users to spot fake news based on the components of IFLA infographic. *Librarianship and Information Organization Studies*, 30(1), 7-26.
- Fishman, B. J., Penuel, W. R., Allen, A. R., Cheng, B. H., & Sabelli, N. O. R. A. (2013). Design-based implementation research: An emerging model for transforming the relationship of research and practice. *Teachers College Record*, 115(14), 136-156.

Fox, M. C., Ericsson, K. A., & Best, R. (2011). Do procedures for verbal reporting of thinking have to be reactive? A meta-analysis and recommendations for best reporting methods. *Psychological Bulletin*, 137(2), 316–344. <https://doi.org/10.1037/a0021663>

Graham, S., & Harris, K. R. (2006). Strategy instruction and the teaching of writing. *Handbook of writing research*, 5, 187-207.

Guo, S., Zheng, Y., & Zhai, X. (2024). Artificial intelligence in education research during 2013 2023: A review based on bibliometric analysis. *Education and Information Technologies*, 1-23.

Hand, B., Shelley, M. C., Laugerman, M., Fostvedt, L., & Therrien, W. (2018). Improving critical thinking growth for disadvantaged groups in elementary school science: A randomized controlled trial using the science writing heuristic approach. *Science Education*, 102(4), 693–710. <https://doi.org/10.1002/sce.21341>

Holmes, W., Porayska-Pomsta, K., Holstein, K., Sutherland, E., Baker, T., Shum, S. B., ... & Koedinger, K. R. (2022). Ethics of AI in education: Towards a community-wide framework. *International Journal of Artificial Intelligence in Education*, 32, 1-23. <https://doi.org/10.1007/s40593-021-00239-1>

Huerta, M., & Garza, T. (2019). Writing in science: Why, how, and for whom? A systematic literature review of 20 years of intervention research (1996–2016). *Educational Psychology Review*, 31(3), 533–570. <https://doi.org/10.1007/s10648-019-09477-1>

Hutchison, A. (2024). Make artificial intelligence your friend, not your foe, in the literacy classroom. *The Reading Teacher*. <https://doi.org/10.1002/trtr.2296>

Keys, C. W., Hand, B., Prain, V., & Collins, S. (1999). Using the science writing heuristic as a tool for learning from laboratory investigations in secondary science. *Journal of research in science Teaching*, 36(10), 1065-1084.

Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., & Neubig, G. (2023). Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9), 1-35.

Luckin, R., Cukurova, M., Kent, C., & Du Boulay, B. (2022). Empowering educators to be AI ready. *Computers and Education: Artificial Intelligence*, 3, doi.org/10.1016/j.caeai.2022.100076

McGrew, S., Breakstone, J., Ortega, T., Smith, M., & Wineburg, S. (2018). Can students evaluate online sources? Learning from assessments of civic online reasoning. *Theory & research in social education*, 46(2), 165-193. doi.org/10.1080/00933104.2017.1416320

Nielsen, J. (1994). *Usability engineering*. Morgan Kaufmann.

Piaget, J. (1980). The psychogenesis of knowledge and its epistemological significance. In

Piatelli-Palmarini (Ed.), *Language and learning* (pp. 23-34). Cambridge, MA: Harvard University Press.

Pressley, M., & Afflerbach, P. (1995). *Verbal protocols of reading: The nature of constructively responsive reading*. Lawrence Erlbaum Associates.

Qadir, J. (2023). Engineering education in the era of ChatGPT: Promise and pitfalls of generative AI for education. In *2023 IEEE Global Engineering Education Conference (EDUCON)* (pp. 1–9). IEEE. <https://doi.org/10.1109/EDUCON54358.2023.10125121>

Roose, K. (2023, January 12). Don't ban ChatGPT in schools. Teach with it. *The New York Times*. <https://www.nytimes.com/2023/01/12/technology/chatgpt-schools-teachers.html>

Shanahan, C., & Shanahan, T. (2014). The implications of disciplinary literacy. *Journal of Adolescent & Adult Literacy*, 57(8), 628-631.

Shanahan, T., & Shanahan, C. (2008). *Teaching disciplinary literacy to adolescents: Rethinking content-area literacy*. Harvard Educational Review, 78(1), 40-59.

Stanford History Education Group. (2016). Evaluating information: The cornerstone of civic Online reasoning [Executive Summary]. Stanford University. Retrieved from <https://stacks.stanford.edu/file/druid:fv751yt5934/SHEG%20Evaluating%20Information%20Online.pdf>

Strauss, A., & Corbin, J. (1998). *Basics of qualitative research: Techniques and procedures for developing grounded theory* (2nd ed.). Sage.

Su, J., & Yang, W. (2023). Unlocking the Power of ChatGPT: A Framework for Applying Generative AI in Education. *ECNU Review of Education*, 6(3), 355-366. <https://doi.org/10.1177/20965311231168423>

Sweller, J., & Chandler, P. (1991). Evidence for cognitive load theory. *Cognition and Instruction*, 8, 351–362.

Theophilou, E., Koyutürk, C., Yavari, M., Bursic, S., Donabauer, G., Telari, A., Testa, A., Boiano, R., Hernandez-Leo, D., Ruskov, M., Taibi, D., Gabbiadini, A., & Ognibene, D. (2023). Learning to Prompt in the Classroom to Understand AI Limits: A Pilot Study. In R. Basili, D. Lembo, C. Limongelli, & A. Orlandini (Eds.), *AIXIA 2023 – Advances in Artificial Intelligence* (pp. 481–496). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-47546-7_33

Trust, T., Whalen, J. & Mouza, C. (2023). Editorial: ChatGPT: Challenges, Opportunities, and Implications for Teacher Education. *Contemporary Issues in Technology and Teacher Education*, 23(1), 1-23. Waynesville, NC USA: Society for Information Technology & Teacher Education. Retrieved November 10, 2024 from <https://www.learntechlib.org/primary/p/222408/>.

Ültanır, E. (2012). An epistemological glance at the constructivist approach: Constructivist learning in Dewey, Piaget, and Montessori. *International journal of instruction*, 5(2).

U.S. Department of Education, Office of Educational Technology. (2023). *Artificial intelligence and the future of teaching and learning: Insights and recommendations*.
<https://www2.ed.gov/documents/ai-report/ai-report.pdf>

Walter, Y. (2024). Embracing the future of Artificial Intelligence in the classroom: the relevance of AI literacy, prompt engineering, and critical thinking in modern education. *International Journal of Educational Technology in Higher Education*, 21(1), 15. doi.org/10.1186/s41239-024-00448-3

Wineburg, S., & McGrew, S. (2019). Lateral Reading and the Nature of Expertise: Reading Less and Learning More When Evaluating Digital Information. *Teachers College Record: The Voice of Scholarship in Education*, 121(11), 1-40. <https://doi.org/10.1177/016146811912101102> (Original work published 2019)

Yan, L., Greiff, S., Teuber, Z., & Gašević, D. (2024). Promises and challenges of generative artificial intelligence for human learning. *Nature Human Behaviour*, 8(10), 1839-1850.

Zhai, X. (2022). *ChatGPT user experience: Implications for education*. Available at SSRN 4312418.