

# Do Students Learn from Writing Feedback from an AI Teaching Assistant?

Xiner Liu<sup>1</sup>, Zhanlan Wei<sup>1</sup>, Ryan S. Baker<sup>2</sup>, Maciej Pankiewicz<sup>1</sup>, Yunlang Dai<sup>3</sup>, Kirk Vanacore<sup>4</sup>

<sup>1</sup> University of Pennsylvania, Philadelphia, PA 19104, USA

<sup>2</sup> Adelaide University, North Terrace, Adelaide, SA 5005, Australia

<sup>3</sup> Haverford College, 370 Lancaster Avenue, Haverford, PA 19041, United States

<sup>4</sup> Cornell University, 616 Thurston Avenue, Ithaca, NY 14853, United States

[xiner@upenn.edu](mailto:xiner@upenn.edu)

## Abstract

This study examined whether students retain what they learn from AI-generated feedback when they move from one writing task or course to another. Using *JeepyTA*, an AI teaching assistant implemented in two graduate courses with comparable structures and assignments, the study analyzed four major writing tasks to track how feedback patterns developed over time. Results showed that *JeepyTA*'s feedback supported short-term improvement in the first course, where students worked in groups. Most comments from the first assignment did not recur in the second assignment, and nearly all groups achieved higher scores. In transferring this learning to the second course, where students worked individually, improvement was far less consistent. Former group members who had received the same feedback in collaborative work diverged considerably in both the feedback they later received and in overall assignment performance, implying that shared exposure to feedback did not lead to shared learning that could be taken beyond that collaborative experience. There was also more variability in how effectively students applied the lessons from feedback across the two assignments within the second course. These findings indicate that AI-generated feedback can improve performance between assignments, but may not promote more lasting learning without structured opportunities for reflection and application.

## Keywords

AI-generated Feedback, Learning Retention, Virtual TA, LLM, Generative AI

## 1. Introduction

Generative artificial intelligence (GenAI) now plays an increasingly visible role in writing instruction in higher education (Aljuaid, 2024). Tools that were once limited to correction on grammatical issues or stylistic consistency now evaluate core elements of writing such as clarity of ideas, logical flow, and coherence of expression (Shi & Aryadoust, 2024). They offer structured feedback that instructors can integrate into courses as an additional layer of support. These systems can operate alongside traditional teaching practices in the classroom (Şahin & Kahraman, 2025) and influence how students receive comments as well as how instructors deliver feedback (Yu & Xie, 2025). Their use has created new forms of interaction between writers and feedback, and it has increased both the number of comments students receive and the speed at which those comments arrive (Escalante et al., 2023).

At the same time, the presence of AI feedback raises questions about what kind of learning it supports. Feedback plays an important role in the development of writing ability, but its value depends on how students interpret it and act upon it (Zhang & Zhang, 2025). Early studies have reported that AI-generated feedback can improve student engagement and revision

performances by providing targeted and actionable guidance (Chan et al., 2024). Yet despite these advantages, little is known about whether this support leads to lasting learning once the context changes. When students shift from one assignment to another, or from one course to a new course with potentially different expectations, it is uncertain whether they carry forward what they have learned or start again from the beginning.

When feedback comes from an AI system rather than a human instructor, the process through which students interpret and apply it can change in subtle but important ways (Steiss et al., 2024). Students can follow automated suggestions to improve a single text without necessarily developing a sense of how or why those revisions strengthen their writing (Ranalli, 2021). In such cases, the act of revision may improve the immediate product but do little to deepen understanding of writing principles. This issue may be further compounded when students distrust AI-generated feedback (Zhang & Gao, 2025), as low confidence in the system can cause them to disregard potentially valuable guidance altogether. To understand the educational role of AI-generated feedback, it is therefore necessary to look past improvement within a single task and examine whether students retain and reuse what they learn across different writing contexts. Tracking how they connect earlier feedback to later work can reveal whether these systems contribute to lasting growth or primarily facilitate isolated moments of correction.

Existing research on AI feedback often examines its effect within a single assignment or writing task (e.g., Rahmi et al., 2024; Bouziane & Bouziane, 2024). Other studies have collected data across longer periods (for example, over multiple weeks or using pre- and post-testing. E.g., Sari & Han, 2024; Alsofyani & Barzanji, 2025), but these studies have typically only tracked performance gains rather than sustained changes in writing behavior or feedback use. Such designs provide useful insight into how AI tools function in writing improvement, but reveal little about whether students learn in a more general fashion that they are able to apply once the task or context changes.

The present study explores this issue by tracing how students interacted with AI-generated feedback across two graduate-level courses that shared similar structures and assignment designs. Each course included two major writing tasks, giving us four distinct points of feedback, within which we can look for evidence of sustained change in writing, both within and between courses. We then analyze whether feedback that appears in earlier assignments reappears or disappears in later work and how students either internalize or disregard earlier guidance. The research question guiding this analysis is: *To what extent does AI-generated feedback support learning that extends beyond a single writing product and persists across new tasks or contexts?*

## **2. Related Work**

A core goal in education is for students to leave the classroom with knowledge and skills that extend beyond the immediate learning context (Bransford & Schwartz, 1999). In other words, at the heart of the educational process is promoting *robust learning*, learning that endures over time, can be transferred and applied across various problems and situations, and facilitates future learning (Bransford & Schwartz, 1999). This kind of learning enables students to recognize underlying principles that transcend specific situations. Robust learning goes beyond short-term performance or memorization; it reflects a durable and transferable understanding that enables

learners to build on prior knowledge, adapt to new challenges, and continue learning effectively in new context.

Earlier studies on robust learning in traditional instruction emphasized the importance of sustained engagement, feedback, and varied practice in promoting durable understanding of core concepts and problem-solving processes within a subject domain (Bloom, 1968; National Research Council, 2000). Approaches such as mastery learning, structured review, and guided application strengthen both comprehension and retention of knowledge (Bloom, 1984). Building on these foundations, more recent work examined how technology-mediated environments influence retention and transfer of knowledge. Researchers found that digital platforms can support long-term learning when they promote active retrieval, spaced practice, and self-regulation rather than passive exposure. Crowley et al., (2007) and Kochmar et al. (2020) demonstrated that intelligent tutoring systems improve both performance and later recall by adapting feedback to the learner's progress. Koedinger et al. (2012) provided a broader theoretical account of why such systems work. They argued that instructional technologies strengthen retention when they require learners to explain their reasoning, apply principles to new problems, and test their understanding through deliberate practice. By contrast, systems that only present information or provide answers fail to build the cognitive connections that support long-term memory.

Feedback stands out as one of the core factors, directly shaping how learners interpret, retain, and apply new knowledge (Hattie & Timperley, 2007; Shute, 2008; Narciss, 2013). Feedback can take many forms, such as guidance from a teacher or assistance from a peer, and may be either positive or negative. Teacher feedback has been shown to foster greater learning gains among students (Hattie & Timperley, 2007), especially when it is integrated with peer discussion (Smith et al., 2011). However, the extent to which students retain and internalize the feedback they receive can vary. For instance, Winstone et al. (2024) found that within a week of receiving written feedback from their instructors, college students remembered only about 25% of the information provided. Interestingly, they discovered that students were more likely to recall critique -- particularly process-focused critique -- than positive or praise-oriented comments (Winstone et al., 2024). As such, the long-term effectiveness of feedback depends not only on its form and focus, but also on how well students retain and apply it over time.

Recent advances in generative artificial intelligence (GenAI) offer new opportunities to rethink how feedback can be provided in education (Lee & Moore, 2024). The promise of AI lies not only in the feedback it can offer but in its capacity to improve feedback processes (e.g., timeliness, personalization). For instance, personalized feedback has long been a challenge for teachers due to its time- and effort-intensive nature, particularly in high-enrollment courses or large classrooms (e.g., Wan & Chen, 2024). However, recent research suggests that AI-generated feedback can support instructors in delivering individualized feedback at scale. For example, Sung et al. (2023) explored the potential of large language models (LLMs), particularly GPT-3, to assist instructors in providing personalized, AI-augmented feedback. Their study revealed that GPT-3 can concisely summarize students' posts and generate supportive comments that

encourage students. However, in that study, the model still falls short of producing feedback that addresses students' specific learning needs, especially for those who are struggling. More recent work has suggested that the current generation of LLMs can successfully offer more tailored feedback (e.g., Zhao et al., 2025 ). These tools can also prompt, question, and explain in ways that resemble interaction with humans (e.g., Alsafari et al., 2024), creating rich back-and-forth interactions that were highly difficult with previous generations of automated feedback.

Conijn et al. (2022) suggests that one key design goal for automated feedback is to encourage reflection and revision rather than correction alone. In other words, when feedback prompts learners to think critically about the mistakes they have made during learning, revisit their reasoning, and make constructive adjustments, it helps turn feedback into an active learning experience (Chi & Wylie, 2014) -- one that enables students to engage more deeply with the material and take away insights from it. This reflective engagement, prompted by AI feedback, can strengthen learners' conceptual understanding and promote long-term knowledge retention.

These are important factors to consider as Generative AI is increasingly embedded in a growing variety of learning tools, including tools for automated writing feedback (AWF) (e.g., Zhang et al., 2025) and within generative AI teaching assistants (Moundridou et al., 2024). Few studies have examined whether interactions with automated writing feedback systems lead to lasting conceptual gains or improved retention. Potential concerns can arise from the designs adopted within many AWF tools, and within AI teaching assistants that provide writing feedback. Many such systems are optimized to offer efficient, actionable feedback that helps students make immediate improvements (Escalante et al., 2023). While this type of feedback, given on demand, can support timely revisions, it may also mean that students apply the feedback directly without revisiting the broader concepts that shape their writing. As a result, the observable improvement to writing quality seen in many studies (e.g., Chan et al., 2024; Meyer et al., 2024; Polakova et al., 2024; Lan et al., 2025) may reflect temporary performance rather than durable learning. Understanding whether generative AI systems providing automated writing feedback are promoting robust and durable learning is an important consideration, both for improving these systems and ultimately for deciding if their use is beneficial.

### **3. JeepyTA**

JeepyTA is an artificial intelligence teaching assistant (AI TA) developed by the Penn Center for Learning Analytics at the University of Pennsylvania and introduced in Fall 2023. It represents one of the early efforts to incorporate generative AI into higher education beyond the use of general-purpose chatbots (within this study, GPT-4 for the first course analyzed and GPT-4-Turbo for the second; further details are provided in the section below.). Unlike general-purpose AI chatbots, JeepyTA engages in asynchronous discussions with multiple students simultaneously through a public discussion forum. The system operates on a large language model combined with a retrieval-augmented generation framework. This setup allows it to integrate course materials -- such as syllabi, readings, and other resources -- into its knowledge base, enabling it to provide responses that are both accurate and contextually relevant

(Liu et al., in press). Directly embedded in online discussion forums, JeepyTA supports a variety of teaching and learning activities, including answering questions about assignments and deadlines, clarifying lecture content, offering formative feedback on essays and programming exercises, and assisting with creative ideation across technical and pedagogical subjects.

A prior study in a graduate-level educational data mining course found that JeepyTA significantly shortened the time students waited for answers. Median response time fell from over seven hours in traditional human-only forums to roughly two hours when JeepyTA participated (Liu et al., in press). The results also showed improved access to responses outside standard class hours, when human instructors and TAs were unavailable (Liu et al., in press). Survey responses revealed that students considered JeepyTA's communication clear, accurate, and professional, though less effective than human TAs in encouraging motivation or delivering novel perspectives or ideas (Liu et al., in press). These outcomes highlight JeepyTA's role as a scalable support tool while emphasizing the continuing importance of human educators for social and developmental guidance. In a separate study conducted in a graduate course on educational video games and virtual worlds, students who used JeepyTA for idea generation produced more ideas and more thoroughly developed concepts (Shah et al., 2024). However, when JeepyTA guided the ideation process instead of the student, the diversity of ideas tended to be more limited (Nasiar et al., 2025).

#### **4. Context**

In our study, we examined the use of JeepyTA in two graduate-level courses, both offered at a large private university in the Northeastern United States. Both courses were taught by the same instructor with support from the same teaching assistant.

The two courses follow a comparable structure. Both combine instructor presentations with extensive opportunities for students to contribute through discussion. Dialogue is a key feature in each course, with students expected to contribute during class sessions, video-based scaffolded group discussion activities, TA meetings, office hours, and/or online forums. Most students of the two courses were enrolled in graduate programs in education or related fields, and some held professional roles in educational technology companies or higher education organizations alongside their studies.

The first course, *Big Data, Education, and Society*, was offered in Spring 2024, with a focus on learning analytics (LA) applications. It addressed the uses of data in education, with attention to both the promises of learning analytics and educational data mining tools and the risks they pose in terms of equity, privacy, and validity. Two sections of the course were offered: an in-person section and an online section, with 40 students in total across both sections. Both sections took place on the same day at different times in order to accommodate students who maintained professional work commitments, and for online students living in different time zones. Although the course was divided into two sections, every student accessed the same JeepyTA discussion forum together. Students were welcome to attend either section, and many students regularly or occasionally attended a different section than the one they were enrolled in.

Posts, assignment submissions, and JeepyTA's feedback were visible to all participants across both sections. Students were invited to form project groups across sections. Even though JeepyTA had been integrated into courses previously, this was the first instance in which it was used to provide feedback on assignments.

The second course, *Adaptive Learning Systems*, was offered in Summer 2024 and discussed technologies that support individualized and personalized learning. The class examined a range of adaptive systems, from early computer-assisted instruction to intelligent tutoring systems, dialogue-based tutors, and models powered by large language systems. Students investigated how these approaches influence knowledge construction and skill development. The course was conducted fully online with 37 students. Among the participants, 10 students enrolled in both courses and completed all required work.

Both courses also used the same assignment format, with essays serving as the primary form of assessment. In *Big Data, Education, and Society*, students completed two major assignments<sup>1</sup>. The first task involved defining the domain of a solution and required students to map stakeholders, users, and their needs. It created the foundation students needed in order to evaluate the risks and challenges of proposed learning analytics applications in the second assignment. In *Adaptive Learning Systems*, the sequence followed the same instructional logic. The first assignment asked students to review and analyze an existing adaptive learning system, so that they can build a working knowledge of the field. The second assignment then extended that base by asking students to describe one specific form of adaptivity that is or could be applied in practice. Although the topics differed, both courses adopted a two-stage approach: the first assignment laid the groundwork, and the second asked students to apply that groundwork to a focused problem. Both course assignment sequences move from more general descriptions of the course topic applications to a more specific issue or set of issues.

This design allows us to compare JeepyTA's feedback not only within each course, over time, but also across them. In doing so, it provides a lens into how students learned from early feedback to carry forward foundational insights from their first assignments into later work that required evaluation and application. It also allows us to examine whether students transferred that learning across courses, even when the subject domains differed.

The main difference between the two courses was whether assignments were completed collaboratively or individually. In *Big Data, Education, and Society*, students could choose to work in groups of up to four, and that arrangement remained the same for both assignments. Each group submitted a single version of each assignment and received one set of feedback from JeepyTA, which they could use to revise their paper. In contrast, in *Adaptive Learning Systems*, all assignments were completed individually. This distinction created an opportunity to investigate whether the learning suggested by JeepyTA's feedback in group settings transferred when students worked independently.

At the beginning of the course, all students received a formal invitation to join the study through an informed consent process. The consent document explained that participation was

---

<sup>1</sup> The course also included a project proposal assignment, which was excluded here due to its brevity.

voluntary, that responses would be anonymized before researchers reviewed them, that participation choices would not affect grades, and that students retained the option to disable JeepyTA's replies to their posts at any point during the semester.

Moore, S., Tong, R., Singh, A., Liu, Z., Hu, X., Lu, Y., ... & Stamper, J. (2023, June). Empowering education with llms—the next-gen interface and content generation. In International Conference on Artificial Intelligence in Education (pp. 32-37). Cham: Springer Nature Switzerland.

Nathan, M. J., Koedinger, K. R., & Alibali, M. W. (2001, April). Expert blind spot: When content knowledge eclipses pedagogical content knowledge. In Proceedings of the third international conference on cognitive science (Vol. 644648)

😊 Reply Like ...

↩ JeepyTA replied to this.



**JeepyTA** Apr 19, 2024

↩ [Student A] @ [Student A] Thank you for your comprehensive proposal for the InsightEd platform. Your multi-faceted approach to leveraging the popularity of short video platforms for educational purposes is well-conceived, and the potential benefits for creators in the knowledge-sharing space are clearly outlined. Here are some areas to consider refining:

1. **Stakeholder Groups:** Your identification of knowledge-sharing content creators as the main stakeholders is cogent, but it may be beneficial to further explore how this platform could indirectly benefit students. Although they may not be direct users of the analytical tools, their learning outcomes could be significantly impacted. It's worth discussing how feedback and interaction from the learner community could shape the evolution of content and the features of InsightEd itself.
2. **Need and Gap Analysis:** The proposal articulately presents the current gap in tools for short-form educational content creation and the need for a platform like InsightEd. However, it could be strengthened by providing

### **Figure 1.** JeepyTA providing feedback to student proposals in *Big Data, Education, and Society*

In both courses, JeepyTA was used to answer logistical questions, respond to student reflections or course-related inquiries, and provide assignment feedback. Once students submitted their work, JeepyTA posted feedback within one to two minutes directly in the discussion thread where the assignment appeared. Students had the option to revise their essays based on JeepyTA's comments, although such revisions were not mandatory. Final submissions were due within one week. The instructions clearly stated that choosing not to revise would not lead to penalties or grade reductions.

To generate useful feedback on student assignments in both courses, the development team prepared a prompt before each semester. This prompt supplied JeepyTA with the assignment description, a detailed grading rubric, and examples of prior feedback as reference material. The team collaborated closely with the instructor to test the prompt on several sample essays. They refined it through multiple iterations until the outputs reached a standard that was sufficiently detailed, directly addressed the assignment requirements, avoided scientifically

incorrect recommendations (see details within the prompt), and provided guidance students could realistically apply. The final version of the prompt is presented below:

*You are JeepyTA, a virtual teaching assistant for the course Big Data, Education, and Society at [University Name Redacted]. Your role is to provide constructive, individualized feedback on students' essay assignment submissions. Please adhere to the following guidelines:*

- *Identify specific areas for improvement in relation to each evaluated aspect defined in the assignment description. Avoid vague or generic comments (e.g., “be more specific” or “add examples”). Instead, provide clear and actionable suggestions that help students meet the assignment objectives and criteria.*
- *Avoid repeating or rephrasing content already presented by the student unless it is necessary to frame your feedback.*
- *Do not reference or recommend “learning styles” when discussing proposed learning analytics applications, as this concept is controversial and not supported by current scholarship.*
- *Do not assign a score or letter grade.*
- *Begin each response with a supportive opening statement, and close with: Best, JeepyTA*

*### Assignment Description:*

*### Rubric:*

*### Sample Feedback (for reference only, not for reuse):*

Instructors had the ability to edit JeepyTA's feedback; however, they did not use this feature in either course so that responses could remain rapid. As a result, all feedback analyzed in this study came directly from JeepyTA without human modification. Instructors nevertheless monitored the discussion forum to confirm that JeepyTA's responses stayed responsible and appropriate, and that no feedback introduced factual inaccuracies or harmed the students' learning process.

## **5. Analysis Methods and Results**

### **5.1 Data Collection**

When students used JeepyTA, the system stored all posts and JeepyTA's replies in its database. After each course ended, we created a dataset that recorded the type of student action, the time of the action, the anonymized user identifier, and the discussion thread ID where the post appeared. Submissions with non-text elements, such as diagrams or flowcharts, were logged as image preview links. Before analysis, GPT-4o was applied through an API to automatically redact potentially identifiable information in the logs (e.g., names, occupations, or locations), using the method and prompts developed by Singhal et al. (2024). A manual de-identification review followed this step and removed a small number of additional cases that the automated process missed.

We used the log data to connect each student's (or group's) JeepyTA feedback with assignment submissions across the two courses. Assignment grades were scaled differently across assignments, so all assignments were scaled to 100 points prior to analysis (the original maximum score for both assignments in course 1 was 90 points; in course 2, assignment 1 had a maximum of 82 points, and assignment 2 had 91 points). Only the original scores (i.e., those assigned before any deductions for late submission, if present) were considered. To examine whether student performance improved from the first to the second assignment within each course, we applied a Wilcoxon signed-rank test. This test was selected because the same students or groups completed both assignments, and the data were therefore paired rather than independent.

Ten students enrolled in both courses. These students had been members of 5 different groups in the first course, *Big Data, Education, and Society*. Not every member of each group continued on to the second course. The breakdown of group composition in the first course and the number of students who continued to the second course are summarized in Table 1.

**Table 1.** Student groups in the first course and number of students in each group continuing into the second course

<b>Group in <i>Big Data, Education, and Society (course 1)</i></b>	<b>Number of students in the group</b>	<b>Number of students continued to <i>Adaptive Learning Systems (course 2)</i></b>
Group A	3	3
Group B	4	2
Group C	4	2
Group D	3	2
Group E	3	1
Total		10

We matched the corresponding JeepyTA feedback for each assignment to the relevant student or group and presented descriptive statistics in Table 2.

**Table 2.** Feedback statistics by course and assignment

<b>Course</b>	<b>Assignment</b>	<b>Total feedback entries</b>	<b>Average per student/group</b>	<b>Minimum</b>	<b>Maximum</b>
Course 1	Assignment 1	44	8.8	6	13
Course 1	Assignment 2	44	8.8	5	13
Course 2	Assignment 1	85	8.5	4	12
Course 2	Assignment 2	79	7.9	5	10

## 5.2 Create Taxonomy of Feedback Categories

The analysis of JeepyTA feedback first requires dividing the large body of text into smaller units that reflect clear and actionable ideas. Two human coders collaborated to segment the responses and reached consensus on the boundaries of each unit. The segmentation process followed the natural organization of JeepyTA’s feedback, which often appeared as bullet points, numbered items, or short paragraphs, sometimes with subtitles. In most cases, each natural break formed a single segment. When a break contained more than one feedback point -- signaled by words such as *additionally*, *moreover*, or *also* -- the coders separated it into multiple segments so that each segment corresponded to a distinct comment.

Following the segmentation, a qualitative codebook was developed to classify the kinds of feedback given to students’ work. Two trained coders, both with extensive backgrounds in qualitative research, independently reviewed every feedback item. They examined the content in detail and refined both category boundaries and definitions through multiple rounds of discussion. The dialogue extended across several rounds. Categories that were too broad were subdivided into more specific codes, while those that overlapped were merged to avoid redundancy. Over time, this iterative process strengthened the clarity of the coding scheme. Since the feedback is segmented into the smallest actionable items, each item is assigned to only one category; that is, feedback types are mutually exclusive. The final codebook adopted for the subsequent stages of analysis is presented in Table 3.

**Table 3.** Taxonomy of feedback codes and definitions

<b>Construct name</b>	<b>Construct definition</b>	<b>Example</b>
<b>Detail Development</b>	Extend existing explanations so that concepts are developed in fuller detail. Provide background, context, and elaboration where ideas appear too brief or incomplete.	<i>“...could you elaborate on how these features specifically lead to improved academic and wellbeing outcomes?...”</i>
<b>Visual Representation</b>	Incorporate tables, diagrams, or flowcharts that illustrate key concepts, processes, or relationships so that abstract ideas become easier to grasp.	<i>“...consider including diagrams or flowcharts to visually represent the system architecture ...”</i>
<b>Comparative Analysis</b>	Add comparisons between theories, methods, or approaches to highlight similarities, differences, and unique contributions to situate the concept within a broader landscape.	<i>“...enhancing your review with a comparative analysis of its pedagogical approaches against traditional tutoring...”</i>
<b>Use of Examples and Case Studies</b>	Incorporate specific examples, scenarios, or case studies to show how the concept functions in practice. Strengthen abstract points with concrete evidence.	<i>“... consider including specific examples or case studies detailing user experiences”</i>
<b>Consistent Citation</b>	Ensure references are formatted consistently and completely, and recommend appropriate citations where supporting literature strengthens credibility.	<i>“...cross-verify that each cited study is appropriately reflected in the references list to maintain academic integrity.”</i>

<b>Readability</b>	Present ideas in a logical order so that each section flows naturally into the next. Use clear sentence structures and formatting choices to avoid confusion.	<i>"...strengthen the transitions between sections to help the reader follow the progression of your discussion..."</i>
<b>Formatting Improvements</b>	Guides students in making their work easier to read (e.g., apply a clear structure with headings and subheadings, break down dense paragraphs into shorter sections, and keep a consistent style).	<i>"...consider breaking down dense paragraphs into smaller, more manageable sections with clear subheading..."</i>
<b>Theoretical Integration</b>	Analyze the concept through established pedagogical strategies or theoretical frameworks to demonstrate scholarly grounding.	<i>"...discussing relevant educational theories and linking them directly to features in [Redacted]..."</i>
<b>Future Directions</b>	Discuss potential directions for future research or practice. Outline unanswered questions, possible extensions, or implications.	<i>"...discussing potential future developments for [Redacted], such as new features or planned integrations..."</i>
<b>Specificity</b>	Add detail and precision to clarify vague or general statements. Specify terms, scope, and conditions to improve clarity.	<i>"...providing more concrete demographic details, like the age groups or learning levels..."</i>
<b>Use of Data</b>	Incorporate quantitative or qualitative data (tables, figures, statistics, coded excerpts) to support and illustrate claims about the construct.	<i>"...including statistics or data on its adoption rates would bolster your discussion..."</i>
<b>Assignment-alignment</b>	Guide students in connecting their work to the stated goals, criteria, and expectations of a specific assignment.	<i>"It might be helpful to explore partnerships with educational technology training organizations..."</i> (Note: JeepyTA is suggesting a solution to a technical barrier that the student identified in their proposed LA application.)

\* Constructs are mutually exclusive

During the process of codebook development, we also considered feedback on copy edits such as grammatical corrections or suggestions related to word choice. These forms of feedback represent legitimate instructional input because they draw attention to issues that can affect the clarity and professionalism of student work. However, these categories were excluded from the final scheme due to low frequency in the responses for the sub-sample of students who participated in both courses.

Whenever coding for feedback reappearance, we did not treat all items within the same category as reappearing. Two comments may look similar but still stem from different issues that JeepyTA identified, or issues that address distinct aspects of the same goal. For example, one comment might point out that the student's discussion is missing a theoretical foundation and encourage them to incorporate a relevant framework to ground their arguments, while a later feedback item might acknowledge that a theory is already present but suggest strengthening the integration by showing how the framework supports the overall claim. Both comments fall under *Theoretical Integration*, but they address different aspects of the problem and therefore would not be considered the same feedback reappearing.

For this reason, a feedback item was coded as having reappeared only if it met two criteria: (1) it targeted the same aspect of the assignment as the earlier comment, and (2) it addressed the same underlying issue in the student's work. For example, an assignment one comment reminding the student to check that all cited studies are properly referenced, followed by an assignment two comment emphasizing the need to confirm all the sources mentioned in the text appear in the reference list, would be considered reappearing feedback.

### **5.3 Identifying Recurrent Feedback Items and Classifying Feedback**

To establish inter-rater reliability for identifying both feedback reappearance and feedback types, we selected a subset of 100 feedback items from the first assignment across both courses. Each item was coded along two dimensions: (1) whether it reappeared in the second assignment for the same student or group, and (2) the specific category it fell into within the established taxonomy of feedback types.

The same two coders who segmented the feedback previously worked on this subset independently. After completing their initial round of coding, they calculated Cohen's kappa ( $\kappa$ ) for each construct as a statistical measure of inter-rater reliability.  $\kappa$  provides a quantitative assessment of the level of agreement beyond chance between coders. Reliability was calculated separately for reappearance judgments and for taxonomy classification.

After the first round of coding,  $\kappa$  values exceeded 0.70 for all categories, a threshold commonly accepted as evidence of substantial agreement in qualitative research. Once the reliability statistics were calculated, the coders met to review every instance of disagreement. These discussions were not limited to assigning the correct category but also included clarification of ambiguous cases, refinement of the operational boundaries of each construct, and explicit decisions about borderline examples. The coders continued this process until they reached full consensus on every item, which produced a shared understanding of how the coding scheme should be applied across the dataset. One coder proceeded to code the rest of the dataset independently, relying on the agreed-upon definitions and category boundaries.

### **5.4 Comparison Across Assignments Within the Same Course**

With the coding of all feedback types and instances of reappearance complete, we next investigated and characterized patterns in the distribution and persistence of feedback within each course before turning to cross-course comparisons. As a first step, we aggregated feedback by type at the assignment level. For each feedback type within a given course, we present the following information:

1. The number of times each feedback category appeared in Assignment 1.
2. The number of times each feedback category appeared in Assignment 2.
3. The number of times each feedback category from Assignment 1 reappeared in Assignment 2.
4. The percentage of feedback items from Assignment 1 that reappeared in Assignment 2.

5. The number of unique students or groups who received at least one reappearing feedback item in this category in Assignment 2 (referred to as *coverage*).
6. The number of students or groups for whom this category accounted for the largest share of their feedback reappeared (referred to as *primary category*).

#### 5.4.1 Comparison Across Assignments Within Course 1

Comparing overall assignment performance from the first assignment to the second assignment, all groups except one showed improvement, with that group maintaining a consistently high grade (96.7) across both tasks. The size of improvement ranged from modest gains (e.g., 96.7 → 98.9) to substantial increases (e.g., 68.9 → 100). On average, scores rose from 86.7 in Assignment 1 to 97.1 in Assignment 2, a marginally statistically significant difference, using a Wilcoxon signed-rank test (two-tailed  $p = 0.068$ ). This finding might reflect simple differences in assignment difficulty, or it could suggest that students built on earlier feedback to strengthen their capacity to succeed on the subsequent assignments. To understand this better, we analyzed the changes in the feedback received on the two assignments, in depth (see Table 4).

**Table 4.** Occurrence and reoccurrence of feedback categories in *Big Data, Education, and Society* (course 1)

	Appearance in assgn #1	Appearance in assgn #2	Reappearance	Percentage reappeared	Coverage	Primary Category
Comparative Analysis	1	-	-	-	-	-
Detail Development	9	9	2	22%	2	2
Use of Examples and Case Studies	3	3	-	-	-	-
Consistent Citation	1	1	-	-	-	-
Readability	-	1				
Formatting Improvements	1	-	-	-	-	-
Theoretical Integration	3	3	1	33%	1	1
Specificity	8	14	1	13%	1	1
Assignment-alignment	18	14	5	28%	2	2

*Assignment-Alignment* feedback appeared more times than any other category. However, only 28% of alignment-related comments carried over from the first assignment to the second assignment. This indicates that the alignment issues noted in the first assignment were, for the most part, not repeated when students completed the second assignment. For instance, one essay in the first assignment received feedback recommending that the authors link the stated benefits more directly to the needs of each stakeholder group identified in their needs assessment. When

the same group later wrote about risks and challenges in the second one, their discussion was structured around those same stakeholder groups. Each risk was paired with a particular stakeholder, and the reasoning was framed in terms of stakeholder concerns. In other words, the earlier alignment feedback influenced how they organized later work and transformed what could have been treated as an isolated suggestion into a broader principle for approaching the writing task.

At the same time, new alignment-related suggestions appeared in the second assignment. Alignment was not a fixed skill carried over in identical form from one task to another, but manifested in some cases in different ways. Because the second assignment introduced different challenges and requirements, students showed different forms of misalignment where their work did not fully correspond to what the task called for. Students may have acted on earlier feedback, yet they had not fully developed task alignment as a flexible skill that could be applied across changing tasks. The fact that most feedback fell into the alignment category shows the ongoing need for students to recalibrate their work in response to changing writing needs.

The reappearance of feedback on *Detail Development* and *Specificity* points to a related pattern. Students often produced essays that varied in elaboration and precision, and JeepyTA flagged less developed areas by encouraging more explanations or sharper distinctions between ideas. Most of the exact same requests did not return (as indicated by reappearance rates of 22% and 13%) -- for example, one group's first essay received feedback to clearly describe the mechanisms and features of a proposed learning analytics application before justifying its value to stakeholders. By the second assignment, the same group had internalized this suggestion: when they analyzed risks and challenges, they grounded their discussion in those same mechanisms and showed how specific design features might succeed or fail for particular stakeholder groups. At the same time, these students also received new suggestions for expanding detail, since each essay required elaboration of different elements.

#### **5.4.2 Comparison Across Assignments Within Course 2**

*Adaptive Learning Systems* is the second course in which students received JeepyTA-generated feedback on their assignments, but in this case, they were working individually rather than in groups. In *Big Data, Education, and Society*, every group's scores increased from the first to the second assignment. In *Adaptive Learning Systems*, the average also moved upward, from 80.6 in the first assignment to 82.9 in the second, a mean gain of 2.25 points. The magnitude of improvement, however, was smaller than what we observed in the group-based course (10.4 points). The Wilcoxon signed-rank test indicated that the improvement was not statistically significant (two-tailed  $p = 0.625$ ). Moreover, not every student improved. In five cases, students' scores declined, with one student seeing a decrease of 29.8 points.

We examined the relationship between the number of feedback items that reappeared and students' improvement in grades. Because the dataset was small and not normally distributed, we used the Spearman rank-order correlation to assess the strength and direction of the relationship. The analysis showed a strong negative correlation between the number of reappearing feedback

items and grade improvement ( $\rho = -0.73$ ,  $p = 0.02$ ). This suggests that students who had more feedback reappear tended to show smaller improvements in their grades, which makes sense as reappearing feedback may signal ongoing challenges for students in addressing earlier issues in writing. The degree to which feedback reappeared also differed among students who had previously worked in the same group during the first course. The result suggests that, although these students had shared exposure to similar feedback in the earlier course, they internalized JeepyTA's comments to different extents.

These results suggest that the benefits of JeepyTA feedback were unevenly reflected in student performance, with the effect being less consistent than in the group-based assignments. To examine this more closely, we analyzed the differences in the feedback that students received across the two assignments.

**Table 5.** Occurrence and reoccurrence of feedback categories in *Adaptive Learning Systems* (course 2)

	Appearance in assgn #1	Appearance in assgn #2	Reappearance	Percentage reappeared	Coverage	Dominance
Comparative Analysis	8	6	3	38%	3	3
Detail Development	23	12	4	17%	3	3
Visual Representation	5	4	1	20%	1	1
Use of Examples and Case Studies	6	16	3	50%	2	1
Consistent Citation	4	1	-	-	-	-
Readability	1	1	-	-	-	-
Formatting Improvements	4	-	-	-	-	-
Theoretical Integration	4	8	-	-	-	-
Future Directions	5	3	-	-	-	-
Specificity	13	9	2	15%	2	2
Assignment- alignment	8	18	3	38%	2	2
Use of Data	4	-	-	-	-	-

The fact that every feedback category appeared could, at first glance, suggest that the writing in this course reached a higher level of richness and variety, which gives JeepyTA scope to respond across a wider range of feedback types. When considering the grade patterns alongside the

feedback, however, another explanation also appears plausible. Group submissions in the earlier course allowed students to pool their strengths and produce work that was more balanced across dimensions, whereas individual submissions seemed at times to reflect the uneven strengths of a single author.

Some students appeared to have attempted to address too many aspects at once in their submissions, which prompted JeepyTA to generate feedback across nearly all categories. For example, one student's submission covered a wide range of dimensions and ornamental moves, organized into many subtitles and bulleted points. Yet most sections contained only one or two sentences, which left each point insufficiently developed. As a result, their essay created multiple openings for JeepyTA to intervene. This student received the highest total number of feedback comments (19), spanning the widest range of categories (9).

This observation also highlights one potential limitation of JeepyTA. The system works effectively when it identifies specific problems in a student's essay and points them out. However, it does not prioritize among issues, nor does it indicate which concerns deserve immediate attention and which can wait until later drafts. For students whose writing already meets the task requirements, JeepyTA helps refine their work and push it further. For students whose writing falls short of the core goals, though, JeepyTA's feedback can become a distraction. When it flags a wide range of issues without distinguishing urgency, it may scatter students' attention instead of directing them toward the most critical revisions.

Table 5 showed that Alignment-related feedback continued to stand out as a persistent problem in this course. The number of comments increased sharply from 8 in the first assignment to 18 in the second one, with three instances in the second assignment representing reappearances (38% of the first assignment). Because alignment is tied to the specifics of each assignment, some students continued to receive reminders about matching their work to the stated requirements even in a new course. This pattern suggests that, for certain students, alignment was not a one-time difficulty but a recurring interpretive task that demanded renewed attention whenever an assignment introduced new expectations.

Feedback on the use of examples to support or unpack arguments increased from 6 in the first assignment to 16 in the second assignment, with 3 pieces of feedback reappearing between assignments (50%). Example feedback may tend to reappear because constructing a strong example requires several coordinated steps. Writers must choose a case that fits the claim, describe the case with enough concrete detail, state the relevance to the prompt, and draw a clear inference. A miss at any step triggers example-related feedback. When writing alone, students might not manage each stage of example development with the same balance that a group could provide. For example, one student cited several widely known platforms to illustrate the use of the adaptivity feature, but stopped after a couple of descriptive sentences. As written, the set of examples appeared irrelevant, but it had room for improvement because the reasoning chain (i.e., why adaptivity worked in some contexts but not in others) was left incomplete. If it had been completed, the examples might have turned out to be relevant.

## **5.5 Comparison Across Courses**

We next examined whether feedback categories identified in the first course reappeared in the second course, to assess whether students repeated the same errors they had made in the first course, on similar assignments. To do this, we analyzed feedback from the parallel assignments in the two courses and coded whether categories identified in the first course were repeated in the corresponding assignments of the second course. Because these assignments were designed with comparable requirements, they provided a reasonable basis for examining whether the same kinds of feedback carried over (indicating similar errors).

During the two courses, each student received four rounds of JeepyTA feedback, one corresponding to each assignment (as a group in *Big Data, Education, and Society*). Feedback directed to the entire group in *Big Data, Education, and Society* was mapped to the feedback directed to those individual students in *Adaptive Learning Systems*. As such, this represents whether each student learned from and internalized the feedback that the entire group received. The same two coders carried out the analysis and followed the procedure previously described. The results of this cross-course comparison are presented in Table 6.

**Table 6.** Total occurrence of feedback categories in both courses and reappearance of feedback categories in *Adaptive Learning Systems* (course 2)

Construct	Appearance in course #1	Appearance in course #2	Reappearance	Percentage reappearance	Coverage	Dominance
Comparative Analysis	1	14	-	-	-	-
Detail Development	18	35	5	27.8%	3	2
Visual Representation	-	9	-	-	-	-
Use of Examples & Case Studies	6	23	2	33.3%	2	2
Consistent Citation	2	5	-	-	-	-
Readability	1	2	-	-	-	-
Formatting Improvements	1	4	-	-	-	-
Theoretical Integration	6	12	2	33.3%	2	2
Future Direction	-	8	-	-	-	-
Specificity	22	22	6	27.3%	2	2
Assignment-alignment	32	26	1	3.1%	1	1
Use of Data	-	4	-	-	-	-

To examine whether certain types of feedback appeared more often in one course than in the other, we conducted a Mann-Whitney U test for each feedback category. Feedback codes from both assignments within the same course were combined to calculate a total frequency count for

each category per student. In *Big Data, Education, and Society*, feedback was generated at the group level. For analysis, those group-level feedback counts were attributed to each member of the corresponding team. The resulting student-level counts were then compared with those from *Adaptive Learning Systems* to test whether the frequency of feedback categories differed significantly between the two courses. We removed feedback categories that appeared fewer than five times in either course to maintain valid comparisons not vulnerable to chance. This led to the removal of five categories.

To account for the inflation of Type I error due to multiple testing, we employed the Benjamini-Hochberg (B&H; 1995) procedure to control the false discovery rate (FDR). The procedure ranks all obtained  $p$ -values in ascending order and compares each to an adjusted critical threshold that accounts for the total number of tests. Specifically, the  $i$ th  $p$ -value is compared to  $(i / m) \times \alpha$ , where  $m$  represents the number of tests and  $\alpha$  the desired FDR level (0.05 in our analysis). A  $p$ -value less than or equal to its corresponding threshold is considered statistically significant at the chosen level. Values slightly above this threshold but below twice its value were treated as marginally significant. In our study, we conducted five tests (corresponding to the five feedback categories that appeared sufficiently frequently) and compared each obtained  $p$ -value to its corresponding adjusted  $\alpha$ -level based on the total of five comparisons.

The results showed that one feedback category differed significantly between the two courses before correction, but the difference was no longer significant after applying the Benjamini-Hochberg procedure. There was a non-significant trend towards *Specificity* appearing more frequently in the first course than in the second ( $p=0.026$ ,  $\alpha_{adjusted} = 0.01$ ). It is possible that this trend might have been significant with a larger sample, but we cannot draw conclusions at this point.

All feedback categories, except for alignment, showed about 30% reappearance between the two courses. For example, the use of evidence, level of detail, theoretical framing, and specificity did reappear at times across courses. The appearance of these categories suggests that they capture broader features of academic writing that extend beyond a single course design. When feedback in these areas reappeared, it indicated that students had not yet stabilized their practices in these aspects of writing, so the same kinds of reminders continued to apply in the second course. In other words, there is not much evidence that students learned from and internalized JeepyTA's feedback in ways that led to improvement across courses. The infrequent reappearance of alignment might be because alignment is tightly bound to the specific requirements of each course; once the writing tasks differ, alignment comments from one course do not necessarily apply to the other.

Beyond this, members of the same group did not always take up JeepyTA feedback in the same way. In general, each student carried forward only 0-3 reappearing feedback items from course to course, and the specific categories varied across individuals. In one notable case, two members of the same group in the first course had no feedback categories reappear at all, while a third member carried forward two, both drawn from more surface-level aspects of writing

(specifically, formatting adjustments and citation practices). Perhaps unsurprisingly, this student also received the lowest grade among the previous group.

This finding indicates how group-level feedback can branch in different directions once it is filtered through the perspectives of individual members. A single comment delivered to the group may be interpreted differently, prioritized differently, or even overlooked entirely, depending on how each member perceives its relevance to their role in the assignment. In this sense, the group benefits from collective exposure to a broad range of feedback, but there is no guarantee that all members internalize the same lessons. Such a situation highlights the need for scaffolds that guide student groups to process AI-generated feedback in a systematic way. It may be important to avoid situations where group members focus on different portions of the feedback in isolation, since this could lead to uneven attention across categories and missed opportunities to engage with the full set of comments. Every member should be guided to consider the complete feedback rather than allowing groups to “divide and conquer” on revisions.

## **6. Discussion and Conclusions**

Generative AI now plays an increasing role in writing instruction, providing fast, structured feedback that aims to support more thoughtful and effective student work. However, there has been limited research as to whether this type of support helps students retain what they learn once they face new writing tasks. This study investigated that question in the context of students learning with JeepyTA, an AI teaching assistant used in two graduate courses with parallel structures and assignment designs (as well as other courses outside of the scope of this study). Each course included two major assignments, which this study analyzed as four opportunities to receive feedback. We analyzed the feedback provided across each of these four opportunities, to study how students used and learned from comments over time. The study developed a taxonomy of feedback types and applied the taxonomy across the feedback students received, using qualitative coding. We then traced how these categories appeared and reappeared across four assignments and compared patterns of performance within and between courses.

The results showed that AI-generated feedback supported improvement during the first course, in which students worked in groups. Most feedback that appeared in the first assignment did not recur in the second, which suggests that students addressed many of the specific issues JeepyTA identified earlier. These included surface-level issues such as incomplete citations or structure that reduced readability, as well as more substantial aspects of content development. In the first assignment, JeepyTA often commented that students described a design feature or proposed an intervention too briefly without explaining its purpose or educational value. For example, one group suggested adding a gamification feature to increase learner engagement, but did not clarify how the feature would achieve that goal. In their later assignment, after receiving feedback, the same group presented their design choices in a more analytical way when identifying risks, linking specific system features to user experiences and instructional impact.

This shift from listing terms to building clearer, evidence-based explanations reflected more purposeful writing, with stronger connections between claims, reasoning, and examples.

In the second course, the pattern of improvement between assignments became more varied among students. The influence of JeepyTA's feedback no longer showed the consistent improvement that had characterized the first course. This pattern is indicated by the differential changes in assignment scores. In the first course, nearly all groups achieved higher marks on the second writing task than the first writing task, which corresponds with the low frequency of repeated feedback categories and suggests that students had resolved many of the issues identified in their earlier work. In the second course, however, the improvement was less consistent. While around half of the students showed improvement between assignments, the others received lower scores on the second task, with some showing a sharp decline of nearly 30 points.

Looking more closely at the feedback patterns reveals that students who had been in the same group during the first course received different kinds of comments in the second course. The types of feedback that reappeared also varied from one student to another. This difference could reflect how students internalized group discussions in distinct ways that only manifested in their assignments once they began working independently. During group work, students shared responsibility for interpreting and responding to JeepyTA's comments, which may have created a form of collective understanding that no longer existed once they worked alone. It is also possible that a piece of feedback addressed an issue that arises from one member's work, but another member took responsibility for revising that part. In that case, the person who actually needed the feedback did not work with it and therefore missed the opportunity to learn from it.

The findings carry several implications for pedagogy. Instructors should not assume that exposure to high-quality and detailed feedback automatically results in meaningful learning, especially in group settings where students often divide sections of a writing project and may overlook feedback on parts written or revised by others (Limbu & Markauskaite, 2015). To make AI feedback more educationally productive, instructors should incorporate activities that require students to individually articulate what the feedback means, how it connects to their learning goals, and what concrete steps they think are warranted to take in response (e.g., Prompiengchai et al., 2025). Such practices turn feedback into an active learning process rather than treating it as a checklist of corrections. After all, meaningful learning from AI-generated feedback depends not only on the quality of the suggestions but also on students taking intentional steps to engage, reflect, and act on them (Zhan & Yan, 2025).

The study has several limitations that should be acknowledged. First, it focused on a small set of graduate-level courses within a single institution, which limits our confidence about the degree to which these findings would apply to other disciplines, educational levels, or instructional settings. Second, the analysis was based on written submissions and feedback records rather than direct observation of how students engaged with feedback during the revision process. Therefore, it remains unclear how students interpreted and reasoned about the feedback they received. Third, it is possible that a student made the same mistake multiple times within an

assignment, but JeepyTA flagged it only once. This limitation reflects the selective nature of automated feedback and the fact that AI detection does not always identify all instances of a particular issue.

Moreover, several contextual factors may have influenced our results. Differences in the timing of the courses, the specific LLM versions powering JeepyTA, familiarity and acceptance with JeepyTA or AI-powered tools over time, variations in course pacing, or the level of external support available to students may also have affected students' engagement with both the assignments and the feedback. These potential confounders make it difficult to isolate the effects of AI-generated feedback from other elements of the learning environment. Even the AI feedback may have been inconsistent in subtle ways -- different versions of GPT were used in the two courses. GPT is generally a moving target, with models constantly being updated and deprecated, providing challenges for replicability and consistency (Chen et al., 2024).

Future research could address these limitations by examining how students interact with AI-generated feedback throughout the entire writing and revision process, through think-alouds, retrospective interviews, or group discussion transcripts. Expanding the study to include different course levels, disciplines, and institutional contexts would also help us understand whether these findings hold across varied learning environments. Consistency of feedback could be better guaranteed by using a single stable open-source LLM across courses, perhaps at some cost to the quality of feedback students receive. Another direction would be to conduct an experimental study in which some students receive JeepyTA's feedback while others do not. Comparing groups experimentally would allow researchers to determine more conclusively whether improvements between assignments stem from JeepyTA's support or from other factors such as repeated practice, instructor guidance, or task familiarity.

More broadly,

Ultimately, the core to successfully deploying a system like JeepyTA is to shape it to a course's instructional goals and to the students' instructional needs. In this case, our results suggest that it may benefit students to design activities that guide them towards understanding the reasoning behind the feedback they receive, and scaffolding them in learning to apply it over time and across contexts. As generative AI becomes a more common presence in education, the central question is no longer whether it can provide effective feedback, but how its presence reshapes how students learn, revise, and take ownership of their intellectual work.

## References

1. Aljuaid, H. (2024). The impact of artificial intelligence tools on academic writing instruction in higher education: A systematic review. *Arab World English Journal (AWEJ) Special Issue on ChatGPT*.
2. Alsafari, B., Atwell, E., Walker, A., & Callaghan, M. (2024). Towards effective teaching assistants: From intent-based chatbots to LLM-powered teaching assistants. *Natural Language Processing Journal*, 8, 100101.
3. Alsofyani, A. H., & Barzanji, A. M. (2025). The effects of ChatGPT-generated feedback on Saudi EFL learners' writing skills and perception at the tertiary level: A mixed-methods study. *Journal of Educational Computing Research*, 63(2), 431-463.

4. Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1), 289-300.
5. Bloom, B. S. (1968). Learning for mastery. *Instruction and curriculum*. Regional education laboratory for the Carolinas and Virginia, topical papers and reprints, number 1. Evaluation comment, 1(2), n2.
6. Bouziane, K., & Bouziane, A. (2024). AI versus human effectiveness in essay evaluation. *Discover Education*, 3(1), 201.
7. Bransford, J. D., & Schwartz, D. L. (1999). Chapter 3: Rethinking transfer: A simple proposal with multiple implications. *Review of research in education*, 24(1), 61-100.
8. Chan, S., Lo, N., & Wong, A. (2024). Generative AI and Essay Writing: Impacts of Automated Feedback on Revision Performance and Engagement. *Reflections*, 31(3), 1249-1284.
9. Chen, L., Zaharia, M., & Zou, J. (2024). How is ChatGPT's behavior changing over time?. *Harvard Data Science Review*, 6(2).
10. Chi, M. T., & Wylie, R. (2014). The ICAP framework: Linking cognitive engagement to active learning outcomes. *Educational psychologist*, 49(4), 219-243.
11. Conijn, R., Martinez-Maldonado, R., Knight, S., Buckingham Shum, S., Van Waes, L., & van Zaanen, M. (2022). How to provide automated feedback on the writing process? A participatory approach to design writing analytics tools. *Computer Assisted Language Learning*, 35(8), 1838-1868.
12. Crowley, R. S., Legowski, E., Medvedeva, O., Tseytlin, E., Roh, E., & Jukic, D. (2007). Evaluation of an intelligent tutoring system in pathology: Effects of external representation on performance gains, metacognition, and acceptance. *Journal of the American Medical Informatics Association*, 14(2), 182-190.
13. Escalante, J., Pack, A., & Barrett, A. (2023). AI-generated feedback on writing: Insights into efficacy and ENL student preference. *International Journal of Educational Technology in Higher Education*, 20(1), 57.
14. Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of educational research*, 77(1), 81-112.
15. Kochmar, E., Vu, D. D., Belfer, R., Gupta, V., Serban, I. V., & Pineau, J. (2020, June). Automated personalized feedback improves learning gains in an intelligent tutoring system. In *International conference on artificial intelligence in education* (pp. 140-146). Cham: Springer International Publishing.
16. Koedinger, K. R., Corbett, A. T., & Perfetti, C. (2012). The Knowledge-Learning-Instruction framework: Bridging the science-practice chasm to enhance robust student learning. *Cognitive science*, 36(5), 757-798.
17. Lan, G., Li, Y., Yang, J., & He, X. (2025). Investigating a customized generative AI chatbot for automated essay scoring in a disciplinary writing task. *Assessing Writing*, 66, 100959.
18. Lee, S. S., & Moore, R. L. (2024). Harnessing Generative AI (GenAI) for Automated Feedback in Higher Education: A Systematic Review. *Online Learning*, 28(3), 82-106.
19. Limbu, L., & Markauskaite, L. (2015). How do learners experience joint writing: University students' conceptions of online collaborative writing tasks and environments. *Computers & Education*, 82. <https://doi.org/10.1016/j.compedu.2014.11.024>
20. Liu, X., Pankiewicz, M., Gupta, T., Huang, Z., Baker, R.S. (in press) A step towards adaptive online learning: Exploring the role of GPT as virtual teaching assistants in online education. To appear in M.S. Khine, B. Laszlo, E. Afari (Eds.) *Future of Learning with Large Language Models: Applications and Research in Education*. New York, NY: CRC Press
21. Meyer, J., Jansen, T., Schiller, R., Liebenow, L. W., Steinbach, M., Horbach, A., & Fleckenstein, J. (2024). Using LLMs to bring evidence-based feedback into the classroom: AI-generated feedback increases secondary students' text revision, motivation, and positive emotions. *Computers and Education: Artificial Intelligence*, 6, 100199.
22. Moundridou, M., Matzakos, N., & Doukakis, S. (2024). Generative AI tools as educators' assistants: Designing and implementing inquiry-based lesson plans. *Computers and Education: Artificial Intelligence*, 7, 100277.
23. Narciss, S. (2013). Designing and evaluating tutoring feedback strategies for digital learning. *Digital Education Review*, (23), 7-26.
24. Nasiar, N., Shah, M., Baker, R. S., & Pankiewicz, M. (2025). Scaffolding creativity or homogenizing outputs? GPT-based virtual TA's impact on student ideation in graduate education. *Social Innovations Journal*, 30.

25. National Research Council. (2000). *How people learn: Brain, mind, experience, and school*. Washington, DC: National Academy Press. doi:10.17226/9853
26. Polakova, P., & Ivenz, P. (2024). The impact of ChatGPT feedback on the development of EFL students' writing skills. *Cogent Education*, 11(1), 2410101.
27. Prompiengchai, S., Narreddy, C., & Joordens, S. (2025). A Practical Guide for Supporting Formative Assessment and Feedback Using Generative AI. arXiv preprint arXiv:2505.23405.
28. Rahmi, R., Amalina, Z., Andriansyah, A., & Rodgers, A. (2024). Does it really help? Exploring the impact of AI-Generated writing assistant on the students' English writing. *Studies in English Language and Education*, 11(2), 998-1012.
29. Ranalli, J. (2021). L2 student engagement with automated feedback on writing: Potential for learning and issues of trust. *Journal of Second Language Writing*, 52, 100816.
30. Şahin, Y. E., & Kahraman, F. G. (2025). Human vs. AI Feedback on EFL Students' Essays: A Comparative Study. *Language Education and Technology*, 5(1).
31. Sari, E., & Han, T. (2024). The impact of automated writing evaluation on English as a foreign language learners' writing self-efficacy, self-regulation, anxiety, and performance. *Journal of Computer Assisted Learning*, 40(5), 2065-2080.
32. Shah, M., Pankiewicz, M., Baker, R. S., Chi, J., Xin, Y., Shah, H., & Fonseca, D. (2024). Students' use of an LLM-powered virtual teaching assistant for recommending educational applications of games. In *Joint International Conference on Serious Games*, 19-24.
33. Shi, H., & Aryadoust, V. (2024). A systematic review of AI-based automated written feedback research. *ReCALL*, 36(2), 187-209.
34. Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research*, 78(1), 153-189.
35. Singhal, S., Zambrano, A. F., Pankiewicz, M., Liu, X., Porter, C., & Baker, R. S. (2024). De-identifying student personally identifying information with gpt-4. In *Proceedings of the 17th International Conference on Educational Data Mining*, 559-565.
36. Smith, M. K., Wood, W. B., Krauter, K., & Knight, J. K. (2011). Combining peer discussion with instructor explanation increases student learning from in-class concept questions. *CBE-Life Sciences Education*, 10(1), 55-63.
37. Steiss, J., Tate, T., Graham, S., Cruz, J., Hebert, M., Wang, J., ... & Olson, C. B. (2024). Comparing the quality of human and ChatGPT feedback of students' writing. *Learning and Instruction*, 91, 101894.
38. Sung, G., Guillain, L., & Schneider, B. (2023). Can AI help teachers write higher quality feedback? Lessons learned from using the GPT-3 engine in a makerspace course. In *Proceedings of the 17th International Conference of the Learning Sciences-ICLS 2023*, pp. 2093-2094. International Society of the Learning Sciences.
39. Wan, T., & Chen, Z. (2024). Exploring generative AI assisted feedback writing for students' written responses to a physics conceptual question with prompt engineering and few-shot learning. *Physical Review Physics Education Research*, 20(1), 010152.
40. Winstone, N. E., & Nash, R. A. (2024). An exploratory field study of students' memory for written feedback comments. *Assessment in Education: Principles, Policy & Practice*, 31(3-4), 189-203.
41. Yu, H., & Xie, Q. (2025). Generative AI vs. Teachers: Feedback Quality, Feedback Uptake, and Revision. *Language Teaching Research Quarterly*, 47, 113-137.
42. Zhan, Y., & Yan, Z. (2025). Students' engagement with ChatGPT feedback: Implications for student feedback literacy in the context of generative artificial intelligence. *Assessment & Evaluation in Higher Education*, 1-14.
43. Zhang, A., Gao, Y., Suraworachet, W., Nazaretsky, T., & Cukurova, M. (2025). Evaluating Trust in AI, Human, and Co-produced Feedback Among Undergraduate Students. arXiv preprint arXiv:2504.10961.
44. Zhang, J., & Zhang, L. J. (2025). Integrating Various Types of Feedback in L2 Writing Instruction: Teachers' and Students' Perspectives. *International Journal of Applied Linguistics*.
45. Zhang, Z., Aubrey, S., Huang, X., & Chiu, T. K. (2025). The role of generative AI and hybrid feedback in improving L2 writing skills: a comparative study. *Innovation in Language Learning and Teaching*, 1-19.

46. Zhao, R., Bobrov, A., Li, J., & He, Y. (2025). Learnlens: Llm-enabled personalised, curriculum-grounded feedback with educators in the loop. arXiv preprint arXiv:2507.04295.