Computers and Education: Artificial Intelligence 9 (2025) 100497

Contents lists available at ScienceDirect

Computers and Education: Artificial Intelligence

journal homepage: www.sciencedirect.com/journal/computers-and-education-artificial-intelligence

Assessing students' DRIVE: A framework to evaluate learning through interactions with generative AI

Manuel Oliveira *, Carlos Zednik, Gunter Bombaerts, Bert Sadowski, Rianne Conijn

Department of Industrial Engineering and Innovation Sciences, Eindhoven University of Technology, the Netherlands

ARTICLE INFO

Keywords:
Learning
Assessment
Academic writing
Generative AI

ABSTRACT

As generative AI (GenAI) transforms how students learn and work, higher education must rethink its assessment strategies. This paper introduces a conceptual framework, DRIVE, and a taxonomy to help educators evaluate student learning based on their interactions with GenAI chatbots. Although existing research maps student-GenAI interactions to writing outcomes, practice-oriented tools for assessing evidence of domain-specific learning beyond general AI literacy skills or general writing skills remain underexplored. We propose that GenAI interactions can serve as a valid indicator of learning by revealing how students steer the interaction (Directive Reasoning Interaction) and articulate acquired knowledge into the dialogue with AI (Visible Expertise). We conducted a multi-methods analysis of GenAI interaction annotations ($n = 1450$) from graded essays ($n = 70$) in STEM writing-intensive courses. A strong positive correlation was found between the quality GenAI interactions and final essay scores, validating the feasibility of this assessment approach. Furthermore, our taxonomy revealed distinct GenAI interaction profiles: High essay scores were connected to a "targeted improvement partnership" focused on text refinement, whereas high interaction scores were linked to a "collaborative intellectual partnership" centered on idea development. In contrast, below-average scores were associated with "basic information retrieval" or "passive task delegation" profiles. These findings demonstrate how the assessment method (output vs. process focus) may shape students' GenAI usage. Traditional assessment can reinforce text optimization, while process-focused evaluation may reward an exploratory partnership with AI. The DRIVE framework and the taxonomy offer educators and researchers a practical tool to design assessments that capture learning in AI-integrated classrooms.

1. Introduction

The emergence of generative artificial intelligence (GenAI) in higher education has fundamentally disrupted traditional assessment in higher education. This is especially true in academic writing, where GenAI can produce text that is increasingly indistinguishable from human work (e.g., Casal & Kessler, 2023; Fleckenstein et al., 2024; Gao et al., 2023). In this new reality, conventional output-focused assessment methods become unreliable measures of student learning (Swiecki et al., 2022; Yan et al., 2024, pp. 101–111). When the final product no longer provides a clear signal of a student's knowledge or skills, the focus must shift to the process of learning itself. Analyzing the dialogue between a student and a GenAI system can provide a more transparent record of this process. Specifically, this record allows us to assess two critical aspects of learning. First, we can observe how students actively steer the AI, showing evidence of their directive reasoning. Second, we can identify how students make their acquired domain-specific knowledge and skills visible through how they deploy these within these interactions.

To the best of our knowledge, educators lack a practical validated tool for evaluating student learning from interactions with GenAI systems in authentic classroom settings. This paper addresses this need by proposing and testing the validity of a framework, which we call DRIVE (Directive Reasoning Interaction and Visible Expertise), built specifically to assess these two components of student engagement. Although existing studies provide initial insights by documenting student-GenAI interaction patterns (e.g., Cheng et al., 2024; J. Kim, Yu, Lee, & Detrick, 2025; Nguyen et al., 2024; Yang, Fan, et al., 2025), this foundational work has largely been conducted in controlled settings, and initial explorations in authentic classrooms focusing on tool development

* Corresponding author. Human Technology Interaction, Department of Industrial Engineering and Innovation Sciences, Eindhoven University of Technology, Den Dolech, Eindhoven, 5200MB, the Netherlands.
E-mail address: m.j.barbosa.de.oliveira@tue.nl (M. Oliveira).

https://doi.org/10.1016/j.caeai.2025.100497
Received 15 July 2025; Received in revised form 12 November 2025; Accepted 13 November 2025
Available online 13 November 2025

rather than on frameworks for assessing learning (e.g., M. Kim et al., 2024). We developed our framework using an approach aligned with the principles of Design-Based Research (Reeves et al., 2005; The Design-Based Research Collective, 2003), working in a setting with high ecological validity: university courses where students used GenAI to complete graded assignments.

The present work focuses on academic writing in general, with an emphasis on argumentative writing. This form of writing requires developing debatable theses with logical evidence and counterarguments (Toulmin, 1958), encompassing both skills GenAI replicates easily (e.g., text generation, basic argumentation) and struggles with (e.g., complex problem-solving, critical thinking; Cash & Oppenheimer, 2024). A systematic analysis of student-GenAI engagement in this context allows us to identify the types of interactions associated with evidence of learning. To that end, this study first seeks to validate our framework by testing whether its process-focused scores align with traditional essay scores. Second, we identify the specific interaction patterns that are most strongly associated with high-quality work, providing educators with a practical and accessible tool for designing assessments that capture the learning process in AI-integrated classrooms.

2. Background

2.1. The skill of argumentative writing

To contextualize the development of the assessment framework and associated taxonomy, we must first consider the nature of the academic skill it aims to evaluate: argumentative writing. Argumentative writing represents a core academic skill that extends beyond mere text composition to also involve critical thinking, evaluation of evidence, and logical reasoning (Andrews, 2015; Newell et al., 2011). Traditionally, assessment of argumentative writing has focused on evaluating the final product of a student's assignment (i.e., an essay) often according to a grading rubric which focuses on examining structural elements, coherence, use of evidence, and logical progression of arguments (Ferretti & Graham, 2019). However, the integration of GenAI into the writing process calls for innovative approaches to both instruction and assessment that consider how students leverage these tools in developing their argumentative competencies.

The literature on argumentative writing assessment has identified several key dimensions. For instance, Toulmin's (1958) model of argumentation, which identifies claims (i.e., statement the writer wants to improve), warrants (i.e. logical/persuasive connection between claim and evidence), backing (i.e., evidence supporting claim), and rebuttals (e.g., acknowledging alternative viewpoints) as essential components, has informed numerous assessment frameworks (Erduran et al., 2004; Sampson & Clark, 2008). More recent approaches have expanded these frameworks to incorporate evaluations of source integration (Wingate, 2012), and the acknowledgement and integration of different perspectives in the argumentative process (Nussbaum & Schraw, 2007; Wolfe et al., 2009). These established assessment criteria provide a theoretical foundation for understanding the quality of argumentative writing, but are not yet able to account for the collaborative process that emerges when students engage with GenAI tools.

Research on technology-enhanced writing instruction has demonstrated that digital tools can support different phases of the writing process (Little et al., 2018; Zhang & Zou, 2022). However, studies examining the specific impact of GenAI on argumentative writing remain limited. Initial investigations have documented students' utilization of GenAI for writing assignments (e.g., J. Kim, Yu, Lee, & Detrick, 2025) but, to the best of our knowledge, few studies have systematically analyzed how different patterns of GenAI interaction associate with learning outcomes in the specific domain of argumentative writing.

2.2. Assessing learning in the age of GenAI

The challenge of assessing student work when GenAI is involved is rooted in the difficulty of observing the learning process. Educational frameworks like Bloom's Taxonomy have historically guided assessment by looking for higher-order thinking in the final student product (Anderson & Krathwohl, 2001; Marton & Saljo, 1976). An essay that analyzes different perspectives, for instance, is seen as evidence of deep learning. When students use GenAI, however, the final product alone offers an ambiguous signal, making it difficult to disentangle the student's contribution from the AI's (e.g., Fleckenstein et al., 2024; Yan et al., 2024, pp. 101–111). This has prompted a shift toward analyzing the learning process, making the student-GenAI interaction log a primary source of evidence for student learning (e.g., Swiecki et al., 2022).

Early research in this area focused on describing and classifying student-GenAI interaction patterns (e.g., Cheng et al., 2024, pp. 178–188; Pigg, 2024). More recently, research has progressed from description to examining how specific interaction patterns may indicate learning. Much of this initial work has been conducted in controlled settings. This is understandable given the current complexity and barriers associated with conducting education research with GenAI in real classrooms (e.g., Batista et al., 2024; Razi et al., 2025). For example, analyzing a large dataset of crowd-sourced writers, Yang, Raković et al. (2025) found that actively editing AI suggestions was associated with producing work of higher lexical sophistication and cohesion. Similarly, a study with doctoral students by Nguyen et al. (2024) identified two distinct collaboration profiles. High-performing writers demonstrated what these authors called a "Structured Adaptivity" profile, characterized by an iterative, interactive engagement with the AI, where they would critically edit and integrate its suggestions. In contrast, lower-performing writers showed an "Unstructured Streamline" profile, through engaging in a more linear and passive workflow of prompting and pasting without refinement. This aligns with the process-focused work by Yang, Cheng, et al. (2024) which used temporal analysis to distinguish between interaction behaviors of passively accepting AI output and actively revising AI outputs, which these authors connect with the concepts of knowledge telling and knowledge transformation (see Bereiter & Scardamalia, 2013; Cheng et al., 2024, pp. 178–188). Research focusing on the influence of AI literacy, as the one by J. Kim, Yu, Lee, and Detrick (2025) found that students with higher AI literacy achieved better writing outcomes by engaging in more collaborative interactions and using context-rich prompts, rather than using the tool for simple information retrieval. Taken together, this emerging body of evidence suggests that observable interaction patterns, particularly those showing critical engagement and iteration, can serve as valuable indicators of learning.

A few studies have also begun to identify links between students' self-regulated learning strategies and their use of GenAI. For instance, Yang, Fan et al. (2025) found that students who consistently integrated GenAI throughout their writing process achieved significantly higher essay scores compared to students who used the tool sparingly, relying instead on more traditional writing strategies or extensive material review. Their findings suggest that effective GenAI integration can lead to high-quality outcomes. Explorations in naturalistic settings are also emerging, highlighting the complexities of student engagement. For example, some studies reveal some limitations of product-focused assessment in activities where GenAI is allowed. In particular, Zheng et al. (2025) found that students could still answer incorrectly on quizzes despite following correct AI advice. Other studies using data from naturalistic settings have focused primarily on developing educational dashboards to facilitate the logging and monitoring of student-GenAI interactions (M. Kim et al., 2024). The focus of these studies suggests an initial consensus that the way students interact with GenAI can serve as an indicator of their learning process.

This body of work brings a core assessment challenge into focus: the need to distinguish between AI literacy and domain-specific learning.

Much of the current discussion around student engagement with GenAI centers on developing technical skills, such as prompt engineering, or broader competencies like AI literacy (e.g., Jin et al., 2025; Lin, 2024; Lintner, 2024; Long & Magerko, 2020; Wang et al., 2023). AI literacy can be understood as "a set of competencies that enables individuals to critically evaluate AI technologies; communicate and collaborate effectively with AI; and use AI as a tool online, at home, and in the workplace" (Long & Magerko, 2020, p. 2). Although these technical skills might be valuable in an increasingly AI-infused society, an unresolved challenge for educators remains: how to design assessments that also evaluate whether the content of a student's prompts reveals their unique, acquired knowledge of the course material. This distinction is critical for authentic assessment. For instance, imagine a course where students are required to read the novel 'Frankenstein' and write an argumentative essay about it. An example of a prompt demonstrating technical AI literacy could be:

"Act as a PhD in literature. Follow these steps:

1. Task: write an analysis of the creature from the novel 'Frankenstein'.
2. Tone: formal and academic.
3. Format: around 500 words, 3 paragraphs.
4. Content: discuss the creature's underlying motivations in a similar way to the attached examples
5. Attachments: documents including other existing discussions"

In contrast, a prompt signaling evidence of domain-specific learning could be:

> "I'm writing an essay where I argue that the creature in the novel 'Frankenstein' by Mary Shelley behaves in a more humane manner than its creator Victor. I want to find quotes that contrast the creature's displays of empathy (as with the De Laceys) with Victor's selfish abandonment of his creation, can you help me with this task?"

The latter prompt is more diagnostic of learning because it requires the student to articulate acquired course-related knowledge. Of course, signs of AI literacy and domain-specific knowledge may be intertwined within the same interaction. A student might use a sophisticated prompting technique (AI literacy) to ask a nuanced, theory-driven question (domain knowledge). Ultimately, as we will soon argue (section 3.3), what constitutes a meaningful signal of learning in an interaction log must be interpreted through the lens of what a student is expected to know and do in a specific course (Biggs, 1996).

2.3. Research gaps

Our literature review reveals three central interconnected gaps. First, there is a scarcity of evidence from more authentic classroom settings where the process of interacting with GenAI is part of the formal, high-stakes assessment. Although naturalistic studies are emerging (e.g., M. Kim et al., 2024; Zheng et al., 2025), it remains an open question whether process-focused assessment can serve as a valid indicator for learning, at least as it is traditionally assessed through product-focused evaluation. Second, while the literature is growing on identifying interaction patterns (e.g., Nguyen et al., 2024; Yang, Cheng, et al., 2024), there remains a need for theoretically informed pedagogical assessment models that move beyond assessing general AI literacy to instead identify and evaluate evidence of domain-specific learning (i.e., course-specific learning objectives) in the interaction between students and GenAI. Third, it is unclear how the assessment method itself (i.e., grading the process versus the product) may shape and reward (through grades) different types of student engagement with GenAI. Our research is designed to address these gaps (see Section 4).

3. DRIVE framework

Traditional assessment methods struggle to evaluate student learning when GenAI is involved, as the final product no longer unambiguously reveals what a student knows or can do independently. We propose the DRIVE framework as a practical tool to assess learning by systematically examining the process of student-GenAI collaboration rather than only the final output. The development of DRIVE aligns with Design-Based Research principles (Reeves et al., 2005; The Design-Based Research Collective, 2003), which emphasize solving authentic educational problems through iterative design in real classroom settings. The initial impetus came from teachers on our research team who, during early ungraded experiments with GenAI in their courses, recognized the need for a structured method to move beyond intuitive judgments when analyzing student learning in chatbot interaction logs. Through this exploratory analysis, teachers learned that they could identify meaningful indicators of student learning by examining two intertwined aspects of the collaborative process: how students steered the dialogue with AI and what domain-specific knowledge they infused into that dialogue. These practical observations were then formalized into DRIVE's two core components: Directive Reasoning Interaction (DRI) and Visible Expertise (VE). These components were subsequently grounded in established concepts from educational psychology, including theories of self-directed learning, cognitive engagement, and the relationship between questioning and knowledge. This development process positions DRIVE as both empirically derived and theoretically informed, as it emerged from authentic assessment challenges while building on existing pedagogical foundations. The framework serves to identify observable behaviors in student-GenAI interactions that indicate skill acquisition and domain-specific learning. DRI captures how students steer the interaction, revealing their strategic thinking and degree of agency in co-creating with AI. VE captures what domain-specific knowledge students articulate within the dialogue, providing evidence of their understanding of course concepts and skills. While DRI emphasizes the quality of the interaction and VE emphasizes the visibility of knowledge, it is their combination that makes a student's learning process and critical thinking truly assessable. We describe these two components in detail below, followed by the taxonomy we developed to operationalize them (Section 6).

3.1. Directive reasoning interaction (DRI)

To address the need to evaluate the process of student-GenAI interaction, Directive Reasoning Interaction (DRI) evaluates how actively and purposefully the student steers the interaction with the AI. The need for such a process-focused evaluation resonates with the nature of human-AI co-creation. Research shows that passive acceptance of AI-generated text is associated with lower-quality writing, whereas students who actively engage by modifying and iterating on AI suggestions produce work with higher lexical sophistication and cohesion (Yang, Raković, et al., 2024). This demonstrates how the interaction process itself can be informative in the context of assessment. DRI also resonates with long-standing concepts in educational theory while adapting them to this new technological context. It echoes the ideas of heutagogy, a framework of self-determined learning (Hase & Kenyon, 2007). Heutagogy is concerned with "learner-centered learning that sees the learner as the major agent in their own learning, which occurs as a result of personal experiences" (Hase & Kenyon, 2007, p. 112). In this model, the teacher (or AI) facilitates learning by providing scaffolding throughout the process, while the learner maintains ownership of their learning path. Framing the student-GenAI interaction through the perspective of heutagogy allows us to conceptualize GenAI as a resource that a self-determined learner can direct. This perspective also aligns with a constructivist view that emphasizes the active role of students in their learning process (von Glasersfeld, 1989). For example, the Interactive, Constructive, Active, and Passive (ICAP) framework by Chi and Wylie (2014) differentiates cognitive engagement based on a student's overt observable behaviors. ICAP proposes a learning hierarchy where deeper learning occurs as students move from passively receiving information

to actively manipulating, and eventually to constructively synthetize their own ideas or interactively discuss them with others. This focus that ICAP places on the learning process is what aligns it with our concept of DRI. In the context of AI-assisted writing, an example that illustrates how these concepts connect with observed behavior is that of a student who merely copies AI output. This represents an "Active" behavior, which signals a shallow depth of learning because it involves only the manipulation of existing information rather than the generation of new ideas. In contrast, a student who receives an output and directs the AI that generated it with feedback (e.g., "that's a good start, but it's missing the ethical dimension, please revise it to include the principle of non--maleficence") would be engaging with AI in a deeper "Constructive" manner. This latter behavior, which reflects high DRI, represents the kind of generative engagement that ICAP identifies as leading to deeper learning and that has been linked to higher-quality writing outcomes in recent studies (Nguyen et al., 2024; Yang et al., 2025). Behaviors signaling a high degree of DRI involve taking a leading role in the human-machine dialogue, critically questioning AI outputs, and using one's own reasoning to guide the dialogue. These types of interactions can serve as tangible evidence of deeper, more purposeful forms of engagement and self-determined agency. Essentially, a high DRI means the student is more in command of the collaboration. More generally, DRI aligns with the principle of "active human agency", or the empowered capacity for a user to critically assess AI output and take steps to adjust it (Fanni et al., 2023; see also Lyons et al., 2021). This directive stance is not only required for maintaining a "human--in-command" approach but also serves as a cognitive safeguard. Through the engagement in reasoning and intentional steering of the interaction, students can counter the negative effects of automation bias (i.e., tendency to uncritically accept AI-generated information) and mitigate the risks of skill atrophy associated with cognitive offloading, through which a person reduces cognitive effort by delegating a task to AI (e.g., Gerlich, 2025; Wahn et al., 2023). A strong DRI profile can thus be understood as an observable proxy for a student's ability to maintain cognitive and ethical control in the collaborative process.

3.2. Visible expertise (VE)

To address the need to assess domain-specific learning beyond technical prompting skills or more general writing skills, Visible Expertise (VE) focuses on the extent to which the student makes their acquired course-related knowledge visible within the interaction log. This concept resonates with earlier research-based pedagogical frameworks such as "Making Thinking Visible" from Harvard's Project Zero, which argues that for thinking to be truly understood, directed, and assessed, it must first be made observable to others (Ritchhart, 2011). In GenAI-assisted writing, visible expertise encompasses the demonstration of declarative and procedural knowledge and skills. This includes the application of domain-specific knowledge and crucial transversal skills, such as critical thinking, problem-solving, and adaptability. Research on the relationship between knowledge and the act of questioning provides strong support for using student prompts as an indicator for their knowledge. This literature suggests that the ability to pose insightful questions is itself an indicator of expertise. One must "know enough to know what is not known" as the title of a paper by Miyake and Norman states (Miyake and Norman, 1979). In line with this idea, there is evidence showing that more knowledgeable individuals are better able to identify gaps in information and formulate the specific questions needed to fill them (Molinero & García-Madruga, 2011). Functionally, a prompt can signal the nature of the asker's knowledge structure. Questions (or prompts) that serve an epistemic function are aimed at acquiring information to fill specific gaps in one's cognitive model of a topic (Kearsley, 1976). Students working with GenAI tools also demonstrate AI literacy through their ability to critically evaluate and strategically direct AI system outputs. When student prompts introduce specific course concepts, apply unique insights, or build upon pre-existing ideas with AI, they make their intellectual contribution and authorial voice evident. This demonstration of expertise is important because, as GenAI transforms learning, the ability to discern, critically engage, and contribute original thinking retains its essential value. Given GenAI's known limitations in reasoning ability and comprehending context, and its potential to produce unverified or biased content (e.g., Amirizaniani et al., 2024; Bender et al., 2021; Maleki et al., 2024; Shojaee et al., 2025), students' domain-specific knowledge and skills are needed to direct the generated outputs to improve their quality and contextual relevance. VE directly addresses the fundamental challenge of evaluating student learning in GenAI-assisted assignments. For fair and effective educational assessment, teachers must clearly discern students' unique intellectual contributions within the interaction. This visibility offers a window into the student's learning process, allowing for an assessment of skill development that would otherwise be obscured in a final product (e.g., essay). In the classroom context, transparency is essential for accountability and trust. Observing how students shape their interaction with GenAI over time allows teachers to more effectively evaluate their growth in light of the intended learning objectives (e.g., Swiecki et al., 2022), especially when these take the technology into account.

3.3. DRI and VE

The DRIVE framework, illustrated in Fig. 1, posits that interaction patterns with high DRI and VE indicate desirable profiles for using GenAI in argumentative writing and other academic tasks. Although they are analytically distinct, DRI and VE are often intertwined. For instance, a student shows VE by infusing domain-specific content into a prompt while simultaneously showing DRI by using that prompt to steer the AI's output. The frequent co-occurrence of these qualities provides the richest signal of learning. Importantly, the framework is flexible by design. What constitutes high-quality DRI and VE is not universal but is defined by an educator's specific learning objectives. For instance, a student might use a sophisticated prompting technique (AI literacy) to ask a nuanced, theory-driven question (domain knowledge). This potential overlap simply reinforces that the educator must define the assessment's focus. An educator could, for example, choose to assess AI literacy as the primary learning outcome, domain-specific learning, or a combination of both. The VE for a philosophy essay (e.g., applying ethical theories) will differ from that of an engineering report (e.g., analyzing technical data), and an instructor can prioritize different forms of DRI, such as critical feedback over creative exploration. By design, our framework positions the educator's pedagogical goals as the ultimate benchmark for what counts as meaningful evidence of learning in a student-GenAI interaction.

The framework thus serve as a conceptual compass and analytical tool for educators, aiming to support GenAI-compatible assessment by focusing on the quality of students' intellectual partnership with AI and how they actively steer this interaction while displaying their learning throughout the process. To operationalize this framework and systematically analyze interactions for evidence of DRI and VE, we developed a detailed interaction taxonomy (see Appendix A). This taxonomy serves as our analytical tool, allowing us to classify the specific prompts that, when viewed holistically, provide evidence of a student's DRI and VE profile. The full methodology detailing the development of this taxonomy is provided in Section 6.

4. Overview of research aims

This paper addresses several underdeveloped areas identified in the literature concerning the need for research in authentic settings, a focus on domain-specific learning, and an understanding of how assessment methods influence interaction strategies in GenAI-assisted writing. To do so, we present the development of a practice-oriented taxonomy grounded in the DRIVE framework. Our research is primarily

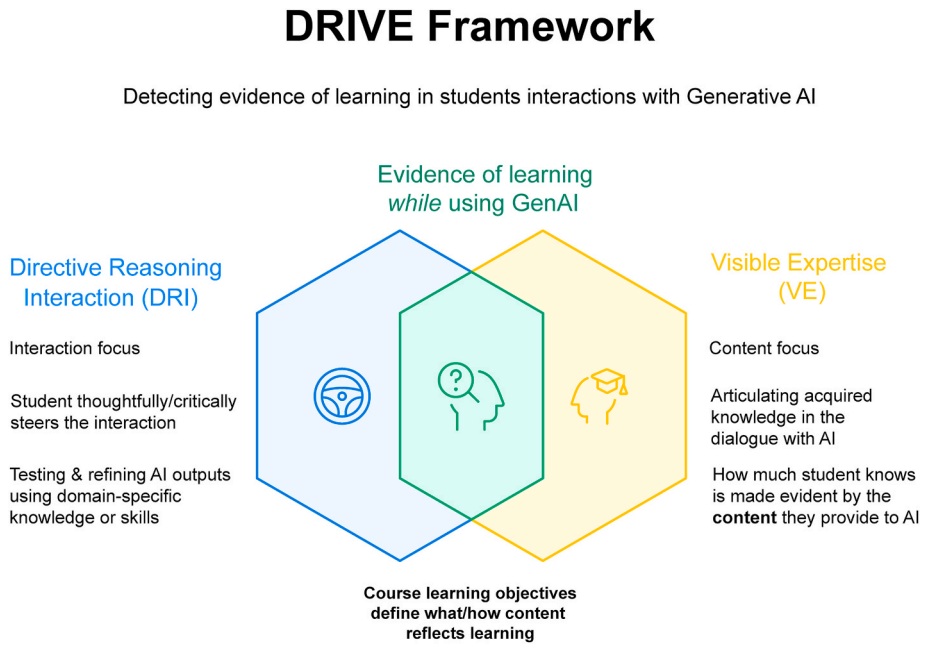

Fig. 1. Overview of the DRIVE framework.

exploratory and descriptive and is guided by the following core questions:

RQ1: How does a process-focused assessment of GenAI interaction quality relate to a traditional, output-focused assessment of essay quality?

This question tests DRIVE's validity by examining whether process-focused measures incorporating DRI and VE indicators correlate with established learning outcomes. A significant positive association would provide initial evidence that analyzing the interaction process is a valid approach for assessing student learning.

RQ2: What types of student-GenAI interaction patterns are associated with different levels of (academic writing) mastery, and do these patterns diverge depending on how mastery is measured?

This question uses our taxonomy to investigate the specific interaction types associated with mastery indicators (see section 8 for the definition of mastery). It is divided into two sub-questions:

RQ2a: How do GenAI interaction strategies connect with different levels of mastery based on traditional essay evaluations and GenAI interaction evaluations?

Here, we aim to identify which taxonomy classifications are associated with above-average versus below-average mastery on each measure. We expect that interaction types associated with higher mastery on both measures will reflect greater student agency over the technology and more visible integration of their own knowledge (core elements of DRIVE).

RQ2b: To what extent do the GenAI interaction patterns associated with different mastery levels overlap between the two assessment methods (traditional essay evaluation vs. GenAI interaction evaluation)?

This is an exploratory follow-up question. We have no specific hypothesis about the outcome. The goal is to investigate the degree to which the two assessment types (grading the final essay vs. grading the interaction process) are sensitive to the same, or different, types of student-GenAI engagement. Differences between assessment methods can inform how the evaluation focus influences the types of interactions with GenAI that are recognized and rewarded (based on DRIVE-based evaluation criteria).

To address these questions, we analyze student-GenAI interaction logs and essay mastery data (i.e., grading scores) from university courses where AI-assisted writing was a graded component. By examining how students use GenAI for real coursework, we aim to provide initial evidence for the utility of the DRIVE framework and its associated taxonomy in understanding learning in AI-integrated settings.

5. Methodology

5.1. Overview

To investigate whether student-GenAI interactions can serve as a meaningful indicator for learning, our study employed a methodology that aligns with Design-Based Research (Reeves et al., 2005; The Design-Based Research Collective, 2003). This approach is well-suited for our context as it focuses on solving practical educational problems in real-world settings through the iterative design of interventions and the generation of new context-bound theory. Central to our approach is the development and application of a taxonomy designed to systematically classify the types of prompts students use to write graded written assignments, in a context where the use of AI tools is allowed by the teachers. We collected both the final written outputs (essays) and the process data (GenAI interaction logs). Student performance was then assessed using two distinct measures: a traditional, output-focused essay score and a novel, process-focused GenAI interaction quality score.

Our research focuses on detecting learning by examining how students interact with GenAI. Therefore, our main analyses necessarily focus on AI users, meaning students who explicitly reported using AI for their graded assignments, since AI use is required for classification and analysis. When we distinguish between output-focused and process-focused assessment, we are comparing two approaches: assessing learning by examining students' AI interactions (specifically, inferring learning from prompt content) versus assessing only the final AI-assisted output using traditional essay evaluation methods. Our analysis proceeded in two phases. For RQ1, we correlated the process- and output-

Table 1
Sample descriptives.

Course/Year (Academic Degree)	AI Users	Non-AI Users	Unknown AI Use	Total Students	Annotated Essays (AI Users)	Total Annotations (AI Users)
Data Science Ethics 2023–2024 (BSc)	32 (21.2 %)	119 (78.8 %)	0 (0 %)	151	21	369
Philosophy & Ethics AI 2023–2024 (MSc)	17 (12.9 %)	106 (80.3 %)	9 (6.8 %)	132	16	309
Philosophy & Ethics AI 2024–2025 (MSc)	54 (33.3 %)	107 (66.0 %)	1 (0.6 %)	162	33	772
Total	103 (23 %)	332 (74.7 %)	10 (2.3 %)	445	70	1450

Prompt Statistics (Annotated Essays Only, N = 70)

Measure	Prompts per Student	Prompt Length (characters)
Mean (SD)	20.71 (18.41)	505 (1026)
Median (IQR)	14.5 (16.75)	168 (390)
Min - Max	2–103	2–9828

Note. Percentages represent proportion within each course. Annotated essays represent the subset of AI user essays that underwent detailed interaction analysis.

Table 2
Evaluation criteria for GenAI interaction logs (process-focused marking rubric).

Criterion	Excellent	Good	Sufficient	Insufficient
AI for Writing	Prompts are clearly formatted and go far beyond the basic parameters of the assignment description, revealing expert-level mastery of using AI as a writing aid.	Prompts are clearly formatted and go considerably beyond the basic parameters of the assignment description, revealing considerable technical ability of using AI as a writing aid.	Prompts are clearly formatted and go beyond the basic parameters of the assignment description, revealing the basic ability of using AI as a writing aid.	No prompts provided, or prompts unclearly formatted. No visible effort to engineer prompts that go beyond the basic parameters of the assignment description.
AI for Argumentation	Extensive critical engagement of AI-generated content. Prompts reveal expert-level use of AI to improve argumentative structure.	Critical engagement of AI-generated content. Prompts reveal considerable efforts to use AI to improve argumentative structure.	Limited critical engagement of AI-generated content. Prompts reveal some effort to use AI to improve argumentative structure.	No critical engagement with AI-generated content. No meaningful effort to use AI to improve argumentative structure.
AI for Course Content	Prompts used to perform extensive content-related research. Prompts reveal deep and broad understanding of, and engagement with, the course material, at times going beyond that material.	Prompts used to perform considerable content-related research. Prompts reveal understanding of and engagement with the course material without going beyond that material.	Prompts used to perform some content-related research. Prompts reveal limited understanding of, or engagement with, the course material.	Prompts used insufficiently for content-related research. Prompts reveal no meaningful understanding of, or engagement with, the course material.

Note. Please note the scores associated with the performance levels (Insufficient to Excellent) were removed to avoid confusion with the standardized scores we use in the main analyses (see Section 8), and due to score range and weighting variations across course cohorts.

focused performance scores to validate the proposed process-focused approach (GenAI interaction evaluation). For RQ2, we identified the interaction patterns characteristic of different mastery tiers on each measure (RQ2a) and then conducted an exploratory comparison to see if both assessment types prioritize the same patterns of GenAI engagement (RQ2b).

5.2. Context and participants

This research was conducted at a STEM university within three Bachelor's or Master's level courses on philosophy and ethics, covering human-technology interaction, the theoretical foundations of artificial intelligence, and the societal impact of Big Data technologies. In all courses, students were required to write a graded argumentative essay. Students were given the option to use GenAI tools for their writing process. This decision was entirely voluntary, as teachers did not enforce AI use but merely offered the option. Data were collected during the 2023–2024 and 2024–2025 academic years.

As detailed in Table 1, 445 students were enrolled across these courses. Of these, 103 (23.2 %) chose to use GenAI under the condition that they would submit their interaction logs for assessment. These logs were formally graded using a marking rubric (Table 2) and contributed to the final course grade .[1] This requirement to have the interaction process formally assessed likely explains the lower-than-expected adoption rate, as students may have perceived the traditional essay-only path as less demanding or lower risk. Final essays (written with or without GenAI assistance) were graded using a product-focused rubric (Table 3). Both rubrics were available to students from the assignment's introduction.

A subset of 70 essays from AI users and their corresponding interaction logs were annotated using the proposed taxonomy (Appendix A), resulting in a total of 1450 annotated prompts. The gap between 103 AI users and 70 annotated essays stems from unusable interaction logs (e.g., broken ChatGPT links or content formatting issues that prevented incorporation into the dataset). The "Unknown AI Use" category in Table 1 refers to cases with insufficient information regarding AI tool engagement. Prompt statistics (number per student, character length) are summarized in Table 1.

Students had free choice in selecting GenAI tools. Among the 70 annotated essays, ChatGPT was most common (n = 48, 68.6 %), one student used Claude (1.4 %), and 21 (30.0 %) did not report their tool. We could not systematically determine which model versions were accessed. The data collection period spans a time when both GPT-3.5 and GPT-4 were available, suggesting these are the most likely models to have been used. This lack of tool- and model-specific data resulted

[1] According to a teacher (also a co-author), students experimented more freely with GenAI before formal assessment of its use began. Student (self-reported) adoption decreased once GenAI use became a graded component.

Table 3
Evaluation criteria for essays (product-focused marking rubric).

Criterion	Excellent	Good	Sufficient	Insufficient
1. Essay introduction and motivation	Characterization of topic/case is unusually insightful, clear, well-formulated, and well-motivated; demonstrates excellent understanding of context and extensive research.	Characterization of topic/case is clear, well formulated-, and well-motivated; demonstrates strong understanding of context; appropriate use of a variety of relevant sources.	Characterization of topic/case is plausible and mostly clear; demonstrates understanding of context; appropriate use of relevant sources.	Characterization of topic/case is unclear, implausible, or absent; shows weak understanding of context or is inadequately supported by sources.
2. Thesis statement	Thesis statement is highly original/creative, clear, plausible, and well-elaborated.	Thesis statement is very clear, plausible, and well-elaborated.	Thesis statement is mostly clear, mostly plausible, and sufficiently elaborated.	Thesis statement is unclear, highly implausible, absent, or elaborated in too little depth.
3. Argument in support of thesis	Argument is highly original/creative, clear, and well-formulated; provides very strong support for position; all crucial parts of argument are well-supported (with sources where appropriate, evidence, and excellent use of relevant course concepts/theories).	Argument is clear and well-formulated; provides strong support for position; nearly all crucial parts of argument are well-supported (with sources where appropriate, evidence, and strong use of relevant course concepts/theories).	Argument is mostly clear; provides support for position; most crucial parts of argument are well-supported (with sources where appropriate, evidence, and appropriate use of relevant course concepts/theories).	Argument is unclear, does not provide support for position, or is missing. Sources, evidence, or use of relevant course concepts/theories to support argument are inadequate or demonstrate significant misunderstanding/error.
4. (At least one) objection and response	Objection is important, clearly stated, and very well developed. Response to objection is original/creative, well-developed, strong, and well-supported.	Objection is significant, clearly stated and developed in detail. Apt, well-developed, well-supported response.	Non-trivial objection developed and discussed. Relevant, plausible response, supported by argument.	Missing, weak, off-topic, or inadequately developed objection; or missing, weak, off-topic, or poorly supported response.
5. Clarity and organization	Publishable quality.	Ideas and claims are well explained; very little repetition, vagueness, ambiguity, or imprecision; sentences are easily understood. Structure of paper is evident and easy to follow; organization is well-suited to the argument. The purpose of each sentence and paragraph is clear.	Most ideas and claims are clear and adequately explained; little repetition, vagueness, ambiguity, or imprecision. Text is understandable. Paper is structured in a mostly clear and logical way. The purpose of most sentence and paragraphs is clear.	Sentences and paragraphs hard to comprehend due to imprecision, grammatical or spelling errors, etc. Inappropriate or insufficient structure; hard to follow argument; relationship between some sentences or paragraphs are not obvious.

Note. Please note the scores associated with the performance levels (Insufficient to Excellent) were removed to avoid confusion with the standardized scores we use in the main analyses (see Section 8), and due to score range and weighting variations across course cohorts.

from our data collection procedure .[2] In the early stages of this research, GenAI models' rapid evolution was not anticipated, so no specific instructions were provided for submitting interaction logs. This led to diverse submission formats (copied text, shared links, screenshots), most of which did not retain the metadata necessary to identify model versions. This resulted in a diversity of formats (e.g., copied text, shared links, screenshots), most of which did not retain the metadata necessary to identify the GenAI model version.

5.3. Data collection procedure

Over a 5-week period in each course, students completed a graded argumentative essay assignment. Students had this entire five-week period to work on their submission. The course had a strict policy that no late assignments would be accepted, except in documented extenuating circumstances. Students were informed that the use of GenAI tools (e.g., ChatGPT) was optional for their essay writing process, encompassing stages such as planning, researching, drafting, or refining arguments. A condition for using GenAI was the submission of complete interaction logs (sequences of input prompts and AI outputs). To mitigate potential disparities in GenAI proficiency, all participating courses included at least one lecture on argumentative writing and basic techniques for using GenAI chatbots effectively, commonly referred to as prompt engineering. Scores reflecting traditional essay grades and experimental overall evaluations of student-GenAI interactions were collected. All data, including interaction logs, essays, and evaluation scores, were collected following informed consent from participating students and ethical approval granted by the Ethical Review Board of [anonymized]. Data were anonymized and stored securely for research purposes.

5.4. Course learning objectives

Across the courses included in this research, students are expected to develop the ability to critically engage with ethical, societal, and philosophical questions related to data science and artificial intelligence. A central learning objective is the capacity to construct well-reasoned, evidence-based arguments in written form. Students learn to identify and evaluate ethical and philosophical arguments, apply major ethical theories to contemporary technological contexts, and analyze value-laden concepts relevant to data-driven practices. They are also trained to read and critically interpret scholarly texts and to use research tools to investigate ongoing societal debates. Argumentative essay writing serves, thus, as a core integrative task through which students demonstrate their ability to synthesize conceptual understanding, ethical reasoning, and domain-specific analysis.

5.5. Measures

Two primary types of measures were used to assess student performance: traditional essay scores and GenAI interaction quality scores.

[2] Although tool diversity and lack of model specification may limit replicability, it benefits ecological validity. Precisely reporting details of tool use and its enforcement may be difficult to implement in such naturalistic settings. It is also conceivable that students can use different chatbots and models concurrently for the same assignment, and it may not always be straightforward to all users to identify detailed specifications of GenAI applications. Please note that our analysis is not fundamentally compromised by this limitation, as our DRIVE framework and taxonomy are designed to be tool-agnostic. Our evaluation focuses on the quality and nature of the students' prompts as indicators of their learning and agency, and these measures are not dependent on the specific capabilities of the AI model.

5.5.1. Traditional essay scores

Student essays from the two courses under study (Data Science Ethics and Philosophy & Ethics of AI) were evaluated by course teachers and teaching assistants using a largely identical grading rubric (see Table 3). The rubric evaluations represent an output-focused measure of performance or mastery of course learning objectives, reflecting the quality of the final written product. The evaluation was based on five core criteria: (1) the clarity and motivation of the introduction; (2) the plausibility and elaboration of the thesis statement; (3) the strength and evidentiary support for the main argument; (4) the development of a significant objection and a persuasive response; and (5) the overall clarity and organization of the essay.

5.5.2. GenAI interaction quality scores

The quality of students' interactions with GenAI was assessed by course teachers and teaching assistants to produce a GenAI interaction quality score, which was part of the final course grade for students who used GenAI. This score was determined holistically (i.e., by assigning a single, overall score based on the entire interaction log rather than each interaction) using a grading rubric (see Table 2), which details criteria for "AI for Writing", "AI for Argumentation", and "AI for Course Content". These criteria are aligned with the taxonomy (i.e., writing, argument, and content; see Appendix A). They integrate course learning objectives and teachers' views of interaction quality .[3] Although these views can be subjective, the criteria link to our DRIVE framework by focusing on agentic cognitive engagement (DRI), seen in students steering prompts and critically revising AI output, and visible knowledge integration (VE), seen in students drawing on and developing their own disciplinary ideas during interaction with the AI. Overall, this score represents a more process-focused measure of performance, compared to the more final output-focused essay scores.

5.5.3. Scoring and reliability

To ensure fairness, a grading policy communicated to students in advance stipulated that the final grade for GenAI users was a weighted average (2/3 essay, 1/3 interaction), while non-AI users were graded on the essay alone. For valid comparison, all scores were standardized as z-scores (see Section 8 for details). Essays and interaction logs were assessed only once, but separately by the same grader (i.e., at different points in time; we also note that sections of essay text were often incorporated in prompts). With this procedure established, baseline analyses confirmed the quality of our measures. Descriptive statistics showed mean z-scores were near the average for both essays ($M = -0.022$, $SD = 0.854$) and GenAI interactions ($M = 0.081$, $SD = 0.798$). Inter-grader reliability was high, with no significant difference between grader means ($F < 1$ for both measures) and low weighted mean absolute deviation (Weighted MADs = 0.035 for essays and 0.209 for interactions). Furthermore, an independent sample Welch's t-test confirmed that there was no significant difference in essay z-scores between GenAI users ($N = 103$) and non-users ($N = 342$) ($t < 1$, $p = .591$, $d = 0.052$, 95% CI [−0.169, 0.273]). These results establish the reliability of our scoring and confirm that simply using GenAI was not associated with essay performance. This allows us to focus our main analysis on how the quality of the student-AI interaction relates to learning outcomes.

[3] It should be noted that, although the rubric is aligned with the DRIVE framework, it was not deductively derived from it. Rather, the rubric emerged from practice, integrating the teachers' expertise and a course's specific learning objectives. As such, the assessment of the GenAI interaction quality based on the rubric involves a degree of expert subjective judgment, as is common in holistic scoring.

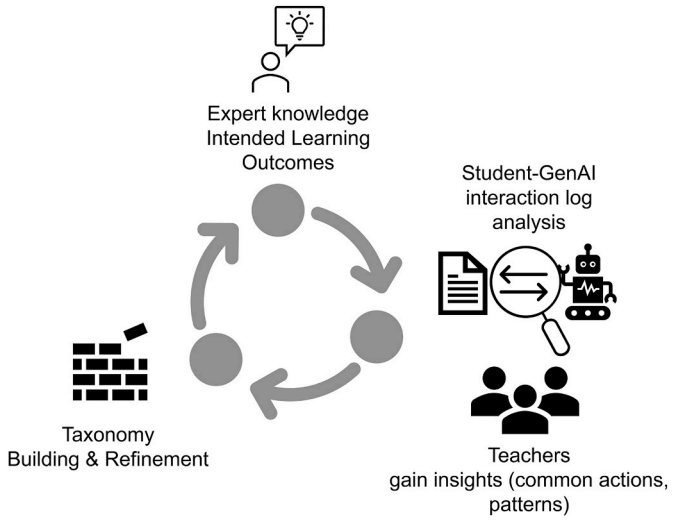

Fig. 2. Illustration of the cycle of taxonomy development and refinement.

6. Development of the taxonomy

This section details the development of our interaction taxonomy, a methodological tool for identifying evidence of DRI (directive reasoning interaction) and VE (visible expertise). Our process aligns with the principles of Design-Based Research in that it stemmed from our teaching team's need for a structured method to move beyond intuitive judgments when analyzing student interaction logs. We developed the taxonomy as a practical tool grounded in course learning objectives, with the DRIVE framework providing the conceptual language (DRI and VE) for the assessment. Its development was also guided by a focus on student agency and knowledge co-construction. To analyze student interactions for evidence of DRI and VE, we developed an interaction taxonomy grounded in course learning objectives (Appendix A). Its development was guided by a focus on student agency and knowledge co-construction. Two university teachers and several expert teaching assistants built the taxonomy through an iterative approach. A top-down component based on pedagogical goals (e.g., see Tables 2 and 3) was refined by a bottom-up analysis of actual student-GenAI interaction logs. The teachers independently reviewed the interaction logs, proposing an initial set of behavioral classifications based on their direct observations. They then met over several consensus-building sessions to collaboratively merge, discuss, and refine these classifications into a consolidated scheme. This practitioner-led process resulted in the three main categories: Writing, Content, and Argument. These main categories align well with established components of argumentative writing theory (e.g., Toulmin, 1958; Wingate, 2012) which identify the core structural elements of an argument (e.g., claims, evidence, rebuttals) and the rhetorical features (e.g., coherence and evidence integration) considered required to present it effectively.

- "Writing" encapsulates interactions focusing on the mechanical and structural aspects of essay composition, including task-oriented actions like providing instruction, requesting content formatting, or requesting assistance to improve and organize specific sections (e.g., introduction, conclusion).
- "Content" captures interactions that center on knowledge construction and understanding, including actions such as requesting definitions, examples, or theoretical explanations, with a particular emphasis on course-specific material and critical engagement with AI-generated output.
- "Argument" encompasses interactions that specifically target logical and analytical aspects of writing such as interactions that develop and further refine argumentative elements (e.g., identifying different

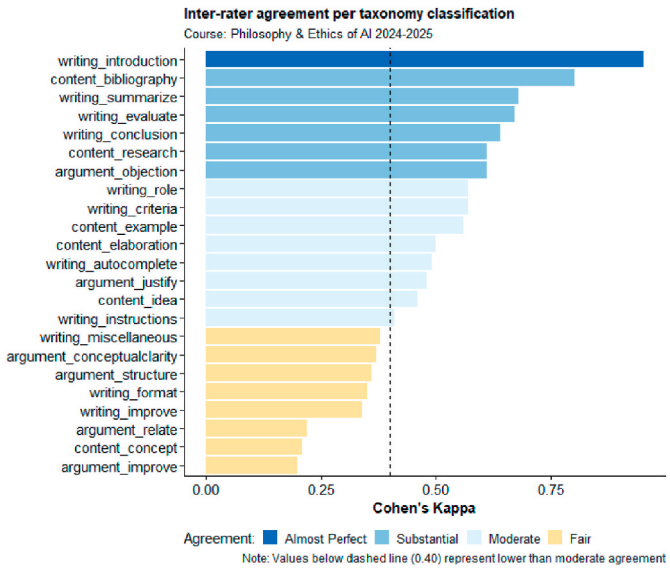

Fig. 3. Inter-rater agreement per taxonomy classifications.

perspectives involved in a given discussion, improving articulation of arguments, strengthening one's thesis).

The final taxonomy contains 35 subcategories within these three main categories: 13 under Writing, 10 under Content, and 12 under Argument. The full taxonomy can be found in Appendix A, and a diagram of the iterative development process is shown in Fig. 2.

7. Annotation of student-GenAI interactions

Across all courses, 103 out of 445 (23 %) essays were (reported to have been) written with GenAI assistance. A total of 70 GenAI interaction logs, associated with the respective amount of graded essays, were annotated. For the remaining 33 cases the interaction logs were sometimes missing (e.g., broken hyperlinks to ChatGPT interaction logs, or missing files), or included a negligible amount of interactions focusing mainly on a few rephrasing requests (e.g., less than five minimally informative interactions). A total of four different annotators annotated the interaction logs. Annotators were instructed to classify each student prompt (i.e., their input) with the best-fitting taxonomy item(s). To accommodate interactions that could be described by more than one item, annotators were free to decide whether to classify an interaction with a single or multiple category-subcategory items (e.g., Writing_Instructions and Content_Research). Interactions classified with multiple taxonomy items are referred to as "Mixed" in our results.

To assess the reliability of the annotation, a subset of interaction logs (n = 33, 772 annotations) of one of the three courses (Philosophy & Ethics 2024–2025) was annotated by three additional raters. Because there was a common second rater to three different raters, we computed the Cohen's Kappa metric of inter-rater agreement for each pair of raters. The average Cohen's Kappa was 0.44 (SD = 0.06), which according to the interpretation guidelines proposed by Landis and Koch (1977), reflect moderate agreement (note this value is at the boundary between the "moderate" and "fair" levels of agreement proposed by these authors). It should be noted that the inter-rater reliability differed between the categories within the taxonomy. At the main category level, agreement was consistently moderate with an indication of higher agreement for Content classifications (ranging from 0.65 for Content and 0.57 for Writing, to 0.46 for Argument, all Kappa values with $ps <$ 0.001). The agreement at the taxonomy subcategory levels (see Fig. 3) was more heterogeneous with some classifications achieving very low agreement (e.g., writing_autoimprove, argument_improve, content_concept) and other very high (e.g., writing_introduction, content_bibliography, argument_objection). In general, however, these data suggest fair and higher agreement for most classifications.

8. Data analysis

Data processing and analyses were conducted using R v.4.3.3 (R Core Team, 2024). Scripts of the analyses are available at the project's repository at Open Science Framework (https://doi.org/10.17605/OSF.IO/32JG7).

Because our primary analyses for RQ1 and RQ2 investigate the relationship between the learning process and the final product, they require data that is, by definition, only available from AI users: their interaction logs and their resulting GenAI-assisted essays. For this reason, our main analyses were conducted exclusively on the data from the 70 students for whom both measures were available. Although we provide baseline data on non-AI users for context (across Section 5), their data is not relevant to these primary analyses.

To allow for the comparability across different course cohorts and grading scales, both traditional essay scores and GenAI interaction evaluation scores were standardized into z-scores within each course subset. A z-score of zero thus corresponded to the average score within the context of a specific course, while negative or positive z-scores quantified how much an individual score was below or above that course average, respectively. This normalization process accounted for the existing heterogeneity in scoring scale ranges across the courses.

To investigate RQ1, which is concerned with the relationship between traditional essay assessment and the experimental assessment of GenAI interaction quality, we calculated the correlation between these measures using the z-scores associated with all the annotated essays (N = 70). Specifically, we calculated both a Pearson product-moment correlation (r) and a Spearman rank correlation (rho). The additional Spearman's rho is particularly useful for classroom-based data as it is less sensitive to common characteristics of real-world educational datasets such as outliers or non-normally distributed observations.

To address RQ2, which investigates whether the developed taxonomy can uncover patterns of student-GenAI interaction associated with different levels of mastery, we analyzed both essay performance and GenAI interaction quality. We define "mastery" as a construct representing skill proficiency in two distinct ways:

1. Essay mastery refers to the demonstrated proficiency in academic writing as reflected in the final essay scores, evaluated based on traditional essay writing quality criteria. They indirectly capture how successfully students incorporated content from GenAI interactions into a coherent academic argument.
2. GenAI interaction mastery refers to demonstrated proficiency in productive engagement with GenAI tools, as assessed by expert graders using interaction quality criteria (Table 2). These criteria capture elements from the proposed DRIVE framework's concepts of Directive Reasoning Interaction and Visible Expertise, which emphasize strategic questioning, critical evaluation of AI outputs, and effective guidance of the AI system toward writing assignment-related goals.

The taxonomy descriptives were calculated to gain a sense of the most prevalent classifications in our sample of annotated interaction logs. Classifications with a prevalence below 1 % were deemed practically irrelevant and were excluded from the analyses of RQ2a and 2b, as their interpretation within the context of the present RQs is less relevant, and these have negligible impact over the results (see online data materials for more details).

For RQ2a, we calculated the mean z-score and 95 % confidence interval (CI) for each taxonomy classification across all annotated interactions with an overall prevalence above 1 %. This allows us to identify which interaction types were associated with different mastery

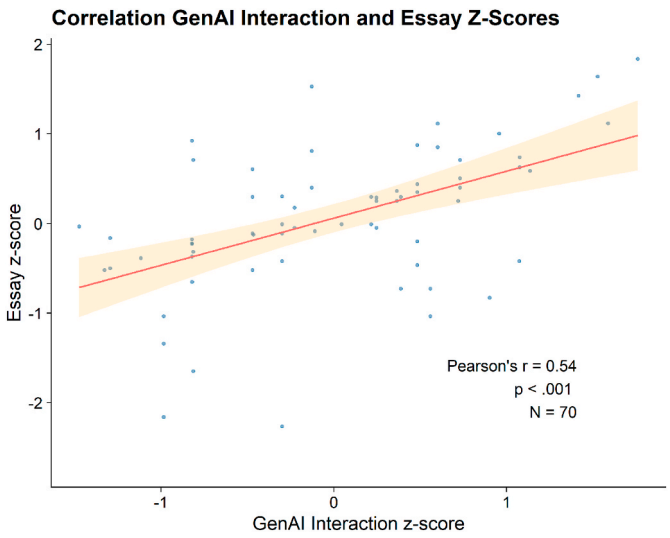

Fig. 4. Relationship between traditional essay scores and GenAI interaction evaluation scores.

levels based on whether the 95 % CI around the mean z-score was entirely above zero (Above Average mastery), entirely below zero (Below Average mastery), or included zero (Average mastery). This approach accounts for the uncertainty in our estimates and ensures that mastery level classifications are supported by sufficient statistical evidence. A z-score of zero represents the average mastery within each course context (as it was calculated within each classroom's sample), thus providing a meaningful reference point for interpreting mastery associations. We then developed qualitative profiles of GenAI interaction patterns by interpreting the taxonomy classifications most strongly associated with each mastery level through analysis of their mean z-scores, 95 % CIs, and theoretical connections to the DRIVE framework.

For RQ2b, we examined whether both assessment methods were sensitive to the same interaction patterns or prioritized different GenAI usage strategies. We employed a dual analytical approach: first examining the degree of overlap between 95 % CIs of mean z-scores for each taxonomy classification as an initial proxy for agreement between assessment approaches. Non-overlapping confidence intervals indicate potential disagreement between methods, while overlapping intervals suggest agreement but do not definitively rule out statistically significant differences. To address this limitation, we conducted exploratory paired t-tests comparing essay and GenAI interaction z-scores for each taxonomy classification, as both measures derive from identical classification observations. We applied a false discovery rate (FDR; Benjamini & Hochberg, 1995) correction across all comparisons to control for multiple testing. Additionally, we calculated Cohen's *d* effect sizes with 95 % CIs to assess the practical significance of any detected differences. This approach allowed us to distinguish between cases where assessment methods truly converge versus those where subtle but meaningful systematic differences exist despite overlapping confidence intervals.

9. Results

9.1. RQ1: relationship between traditional essay evaluations and GenAI interaction evaluations

For the 70 annotated essays, a Pearson correlation indicated a statistically significant, strong positive linear relationship between traditional essay assessment scores (output-focused) and GenAI interaction quality scores (process-focused), $r = 0.54$, 95 % CI [0.34, 0.68], $t(68) = 5.24$, $p < .001$. This suggests that students who demonstrated higher quality interactions with GenAI also tended to achieve higher traditional essay scores. A scatterplot illustrating this relationship is provided in Fig. 4. This alignment between the two types of learning indicators (output-focused essay scores and process-focused GenAI interaction evaluations) lends support to the potential of GenAI interaction evaluations to provide insights into student learning, at least in the same capacity as essay scores allow for.

It should be noted that while the essay z-score distribution met the normality assumption, the GenAI interaction z-score distribution marginally failed the Shapiro-Wilk normality test ($W = 0.965$, $p = .047$). As an additional check, a Spearman's rank correlation was calculated to confirm the relationship remained despite the deviations from normality. This analysis yielded an identical result (rho = 0.54, $p < .001$).

9.2. RQ2: what student-GenAI interaction patterns are prevalent across different levels of mastery, and do these patterns diverge depending on how mastery is measured?

We first describe the overall pattern of taxonomy classifications before focusing on the descriptives per mastery level.

9.2.1. Taxonomy descriptives

The frequencies at which taxonomy classifications were observed during the annotation of student-GenAI interaction logs collected from the three courses are shown in Fig. 5, both at the main taxonomy subcategory level (Fig. 5-A; all above 1 % frequency) and at the category level (Fig. 5-B). The overall pattern for the main categories indicates that the most prevalent category of interactions relate to Writing aspects (41.3 %), followed by Content (28.7 %) and Argument (22.3 %). A total of 7.7 % of the interactions were annotated with more than one category, categorized as "Mixed". Within the Writing category, Writing_Improve (improving spelling, style or grammar of input text) was the most prominent subcategory, accounting for 13.4 % of the total interactions, followed by Writing_Evaluate (requesting evaluation of essay section; 7 %) and Writing_Miscellaneous (prompting system in a non-specific technical way, 4.8 %). It should be noted that the subcategory Writing_Miscellaneous is a "catch-all" classification, and in that sense, its underrepresentation (or overrepresentation) in the results may be interpreted as desirable (or undesirable), as it hints at interactions hard to classify with the current taxonomy content. For Content, the most common subcategory was Content_Research (asking AI to define ideas or find related ideas to user's input; 5.6 %), Content_Bibliography (asking for references, 5.2 %), followed by Content_Elaboration (requesting additional detail incorporating course content, 4.6 %) and Content_Idea (elaborating on existing well-formulated ideas, 4.1 %). Finally, for Argument, Argument_Improve (improving the structure given argument, 5.9 %) was most common, followed by Argument_Objection (providing an objection for a given argument 4.5 %) and Argument_Justify (requesting AI to provide reasons for an input claim, 3.4 %).

9.2.2. RQ2a: how do GenAI interaction strategies connect with different levels of mastery based on traditional essay evaluations and GenAI interaction evaluations?

The following analyses examine how interaction types connect with mastery levels across both traditional essay quality and GenAI interaction quality assessments, revealing how different GenAI usage patterns relate to performance under output-focused versus process-focused evaluation approaches. Fig. 6 shows the mean z-scores (+95 % CIs) by taxonomy classification for both essay scores (in blue) and GenAI interaction scores (in red). Confidence intervals including zero (z-score) reflect average mastery levels, while intervals entirely below or above zero reflect below-average or above-average mastery levels, respectively. This confidence interval approach provides statistically rigorous classification by ensuring that mastery level designations are supported by sufficient evidence rather than chance variation.

To illustrate these findings, Table 4 showcases authentic student

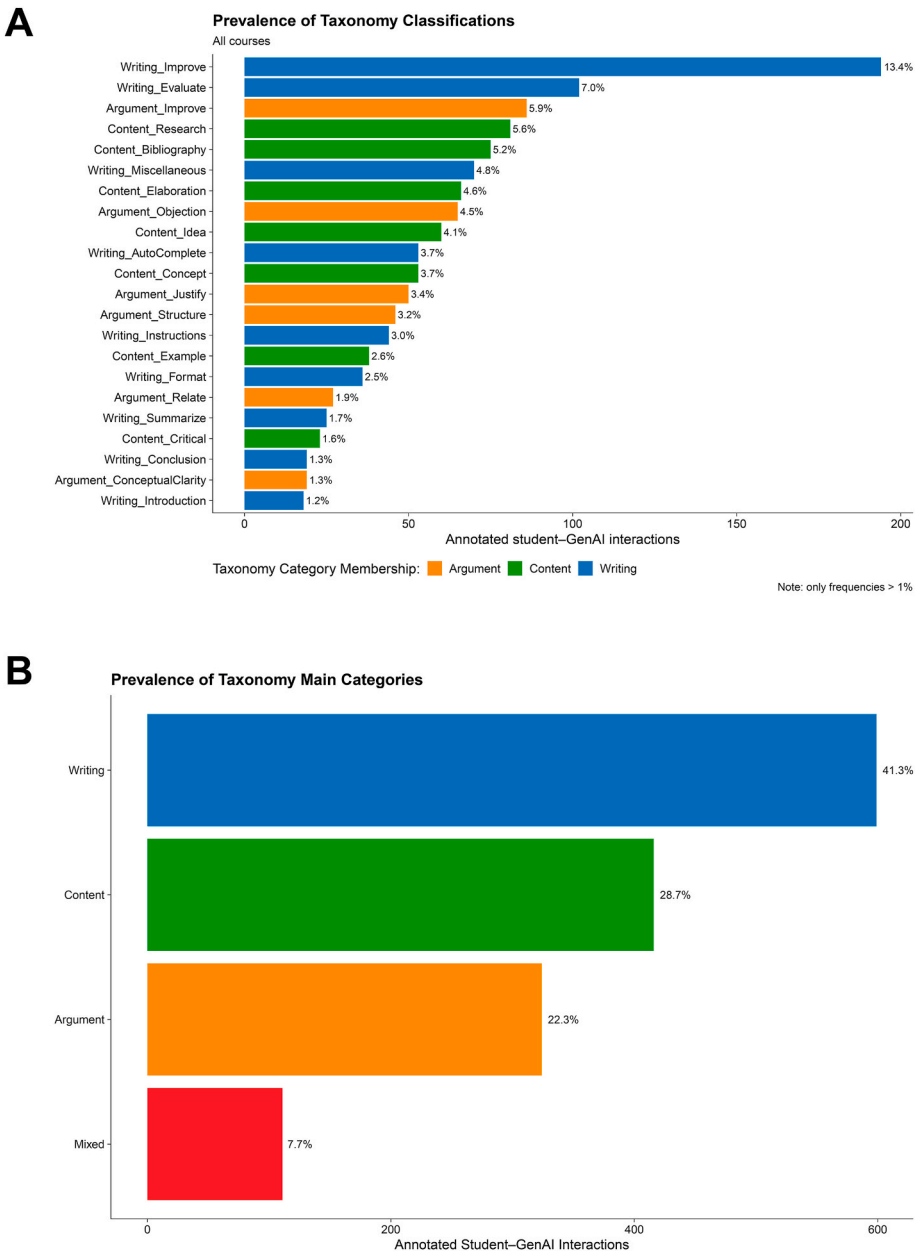

Fig. 5. Overall descriptives of taxonomy annotations for all courses.

prompts from the key classifications associated with above- and below-average mastery. Each prompt is categorized by its associated assessment outcome (Essay or GenAI Interaction score) and includes a rationale analyzing its quality through the criteria of the DRIVE framework.

9.2.2.1. Essay z-scores and taxonomy classifications. Above-average essay mastery was associated with a "targeted improvement partnership" approach, characterized by three distinct but complementary student-GenAI interaction strategies. Writing_Improve dominated this profile ($n = 194$ or 13.4 % of annotations, mean z = 0.13, 95 % CI [0.03, 0.24]), reflecting actions such as the systematic refinement of spelling, style, and grammar in existing text. This was complemented by sophisticated analytical engagement through Content_Critical interactions ($n = 23$ or 1.6 % of annotations, mean z = 0.25, 95 % CI [0.03, 0.47]), where students critically engaged with AI-generated content by asking for clarifications or corrections. This profile was further defined by Argument_Relate interactions ($n = 27$ or 1.9 % of annotations, mean z = 0.36, 95 % CI [0.13, 0.59]), which involved requests to connect or relate two concepts or ideas. This set of strategies suggests that students who achieved higher essay scores engaged GenAI as a targeted text improvement tool, by systematically improving their input work (essay sections) through (inferred) critical evaluation and conceptual integration rather than by seeking comprehensive assistance from GenAI.

Below-average essay mastery was characterized by a "basic information retrieval" prompting strategy, including only two interaction types with z-score confidence intervals entirely below zero (average). Content_Research showed the strongest negative relationship (n = or 5.6 % of annotations, mean z = −0.41, 95 % CI [−0.72, −0.11]), involving requests for AI to define ideas or identify related concepts. Content_Example interactions also demonstrated negative associations (2.6 %, mean z = −0.18, 95 % CI [−0.35, −0.01]), where students asked for specific examples of general cases or issues. This constrained profile suggests that students with lower essay performance primarily used AI for foundational information gathering rather than sophisticated content development or critical engagement.

The predominance of interactions categorized as average (77 %)

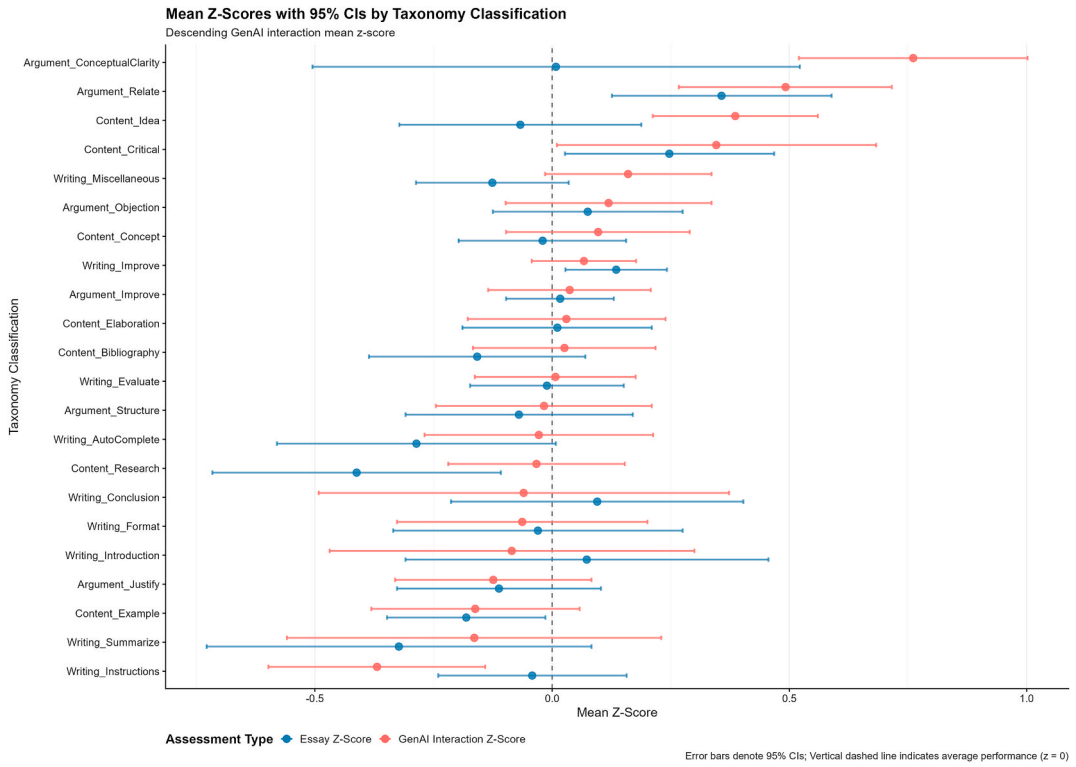

Fig. 6. GenAI interaction classifications and mastery level: Essay and GenAI interaction mean Z-scores +95 % confidence intervals per taxonomy classification.

suggests that most GenAI usage patterns neither significantly enhanced nor detracted from essay writing quality as traditionally assessed (i.e., output focus). This pattern emphasizes the specificity of interaction types that correlate with essay performance and suggests that only a few types of prompting strategies (as identified by the current taxonomy) appear to be connected with very high and very low writing quality as assessed traditionally.

Relating back to the DRIVE framework, the above-average profile demonstrates a moderate display of Directive Reasoning Interaction (DRI) through the apparent targeted steering of the AI toward specific essay improvement tasks. The pattern also suggests an emerging Visible Expertise (VE) as inferred from critical evaluation of AI output, or the requests for assisting with conceptual integration within the essay's narrative. By contrast, the below-average profile shows less evidence of DRI, with interactions focused primarily on information extraction (vs. a more collaborative development of the essay), and minimal VE, as these prompts sought more basic or foundational definitional support (vs. demonstrating original thinking or knowledge synthesis through the usage of GenAI).

9.2.2.2. GenAI interaction z-scores and taxonomy classifications. Above-average GenAI interaction mastery was associated with a "collaborative intellectual partnership" approach, characterized by four interaction strategies that demonstrate an engagement with (Gen)AI as a thinking partner/assistant. Argument_ConceptualClarity emerged as the strongest positive indicator (n = 19 or 1.3 % of annotations, mean z = 0.76, 95 % CI [0.52, 1.00]), involving requests to simplify or improve the definition of concepts. This was complemented by Argument_Relate interactions (n = 27 or 1.9 % of annotations, mean z = 0.49, 95 % CI [0.27, 0.72]), where students asked AI to connect or relate two concepts or ideas in the course of the essay writing process. Content_Idea interactions formed a substantial component of this profile (n = 60 or 4.1 % of annotations, mean z = 0.39, 95 % CI [0.21, 0.56]), where students brought their own well-motivated original ideas or questions to the AI and requested confirmation, elaboration, or discussion of these concepts (assumedly generated outside of the dialogue, likely by the student themself). This profile is further characterized by Content_Critical interactions (n = 23 or 1.6 % of annotations, mean z = 0.35, 95 % CI [0.01, 0.68]), where students critically engage with AI-generated content by asking for clarifications or corrections of the target content (e.g., AI output, student input, or a hybrid content). This combination of strategies suggests that students with higher GenAI interaction scores engaged AI as an intellectual collaborator, leveraging the technology for conceptual refinement, knowledge synthesis, and critical dialogue.

Below-average GenAI interaction mastery was characterized by a "passive task delegation" approach, which included only one interaction type. Writing_Instructions demonstrated the sole negative association (n = 44 or 3.0 % of annotations, mean z = −0.37, 95 % CI [−0.60, −0.14]), involving specifications of tasks in terms of course assignment descriptions, typically through copy-pasting or uploading assignment instructions. This singular profile suggests that students with lower GenAI interaction scores primarily used AI as a direct recipient of student input rather than engaging in collaborative knowledge construction or strategic dialogue. This may be hinting at lower levels of confidence or trust in the capabilities of the AI system, although that remains an open question that cannot be addressed by the current data.

The overwhelming prevalence of average-classified interactions (82 %) indicates that most GenAI usage patterns demonstrated neither exceptional mastery nor deficiency when evaluated against the DRIVE framework's process-focused criteria. This finding highlights the distinctiveness of interaction types that correlate with high or low quality GenAI engagement and suggests that effective collaborative partnership with GenAI requires specific strategic approaches rather than simply general usage competency.

Through the lenses of the DRIVE framework, the above-average profile shows a strong Directive Reasoning Interaction (DRI) demonstrated through the (inferred) strategic steering toward conceptual development and knowledge integration. This was coupled with a clearer display of Visible Expertise (VE) through actions demonstrating original idea contribution and critical evaluation of AI outputs. This

Table 4
Representative prompts illustrating above and below average mastery across essay and GenAI interaction assessments.

Assessment Type by Mastery Level	Interaction Classification	DRIVE-based assessment	Representative Prompt
Above-average examples			
High GenAI Interaction	Argument Conceptual Clarity	High VE: Student brings in specific external knowledge (the "octopus thought experiment"). High DRI: Gives a specific command to revise based on a clear, conceptual critique.	"Let's revise premise 4. Here, we still talk about meaning … it really looks like the octopus thought experiment suggested by Bender & Koller … Give me 3 suggestions on how we can fix this premise … "
High GenAI Interaction	Argument Relate	High VE: Demonstrates deep understanding of what makes a strong argument. High DRI: Provides a nuanced, conceptual command to "create doubt" rather than just giving a simple instruction.	"I think the objection is now a little bit passive … try to create doubt about that it is indeed the correct definition."
High GenAI Interaction	Content Idea	High VE: Student performs external research and brings in a cited academic source. High DRI: Provides a surgical command to use the source for a specific premise and argument.	"I found an article from Redaelli (2024) … Use this source for the third premise, arguing that LLMs lack intentionality … "
High Essay	Argument Relate	High VE: Student introduces a novel, real-world analogy to test a core philosophical concept. High DRI: Asks the AI to make a specific conceptual connection ("How does this relate … ?").	"saw in a discussion in quora that blind people do understand concepts like transparacy … How does this relate to this argument Searle gives … ?"
High Essay	Content Critical	High VE: Student provides a more sophisticated reason to fix the AI's logic. High DRI: The student identifies a specific logical leap (non sequitur) in the AI's reasoning and commands it to expand on the point.	"The last part you mention … Does not follow from the above text necesairly so we will expand on it … the reason it can not prioritize … is because it has not yet been given the same objective as we have … "
High Essay	Writing Improve	High DRI: The student manages the essay's structure across multiple paragraphs, ensuring each has a clear, distinct purpose. This shows high-level strategic control.	" … please leave out anything about language models. that will come in a later paragraph. for this paragraph, really focus on why subjective experience is needed … "
Below-average examples			
Low GenAI Interaction	Writing Instructions	Lack of VE: The student shows no original thought, simply pasting the assignment instructions. Lack of DRI: The student delegates the entire cognitive task of structuring and writing the essay.	"[file_uploaded] Write an essay which answers the essay question by stating a clear thesis … "
Low GenAI Interaction	Writing Summarize	Lack of VE/DRI: The student explicitly outsources the cognitive work of synthesis by telling the AI to "Think about integrating it," rather than doing that thinking themselves.	"Read this lecture about understading in LLMs. Think about integrating it into the essay"
Low Essay Score	Content Research	Low DRI: The student asks a series of disconnected questions, showing curiosity but no clear argumentative direction. This leads to an unfocused final essay.	"Why does it feel like the responses generated demonstrate understanding? What kind of data are LLMs trained on?"
Low Essay Score	Writing Auto Complete	Lack of VE/DRI: The student completely outsources the writing process, asking the AI to compose the entire essay from an outline. This bypasses the learning process.	"Your outline is good, try to write the essay of maximum 1000 words, while respecting the writing guide … "

pattern suggests a behavioral profile where students engage with GenAI as an intellectual collaboration rather than treating it as a mere tool. In contrast, the below-average profile demonstrates minimal DRI, with interactions focused on task specification rather than strategic guidance, and negligible VE, as these actions only show the ability to provide instructions to the system without any signs of user knowledge incorporation, knowledge synthesis, or critical engagement with AI throughout the collaborative process.

9.2.3. RQ2b: to what extent do the GenAI interaction patterns associated with different mastery levels overlap between the two assessment methods (traditional essay evaluation vs. GenAI interaction evaluation)?

Confidence interval overlap analysis revealed substantial convergence between assessment methods, with 21 of 22 taxonomy classifications (95.5 %) demonstrating overlapping CIs. Only Content_Idea showed clear disagreement, with essay evaluation classifying it as below average (95 % CI [−0.32, 0.19]) while GenAI interaction evaluation rated it as above average (95 % CI [0.21, 0.56]). This high level of agreement aligned closely with the strong positive correlation ($r = 0.54$) between assessment methods identified in RQ1. However, this overlap analysis provides a conservative test that may miss statistically meaningful differences when intervals overlap but distributions differ significantly. To explore this possibility, we conducted paired t-tests comparing essay and GenAI interaction z-scores for each taxonomy classification, applying FDR correction across all 22 comparisons to control for multiple testing. This exploratory analysis revealed additional classifications with significant differences. Beyond the already-identified Content_Idea ($p < .001$, $d = -0.50$, 95 % CI [−0.72, −0.28]), four additional disagreements emerged. Argument_ConceptualClarity demonstrated the largest effect ($p = .004$, $d = -0.71$, 95 % CI [−1.09, −0.33]), followed by Writing_Miscellaneous ($p = .001$, $d = -0.40$, 95 % CI [−0.61, −0.20]), Writing_Instructions ($p = .028$, $d = 0.46$, 95 % CI [0.13, 0.80]), and Content_Research ($p = .013$, d

= −0.31, 95 % CI [−0.50, −0.11]). The pattern of disagreements suggests a slight degree of systematic assessment differences in terms of what they may indirectly incentivize through their evaluation focus. Process-focused GenAI interaction evaluation assigned substantially higher scores to conceptualization-related work (Argument_ConceptualClarity, Content_Idea) and flexible AI engagement or diversity of prompts (Writing_Miscellaneous). By contrast, output-focused essay evaluation showed a relative preference for structured task specification (Writing_Instructions) and compensatory information-seeking (Content_Research). Of note, 17 out of 22 classifications (77.3 %) demonstrated negligible effect sizes, indicating that most interaction patterns receive similar evaluations across both methods.

This divergence pattern, although small, suggests that traditional essay assessment may undervalue exploratory behaviors in GenAI interactions that process-focused evaluation rewards as cues to effective student-GenAI collaboration, while simultaneously undervaluing certain foundational interaction patterns that contribute to final product quality. The statistically significant disagreements suggest a small tension between optimizing output quality versus rewarding a more sophisticated engagement with GenAI, which may eventually translate into practical implications for how assessment design shapes student AI usage patterns in educational contexts.

10. Discussion

The present work addressed the challenge of assessing student learning in GenAI-integrated writing environments by shifting the analytical focus from technical skill to the evidence of learning within student-GenAI interactions. While prior work has often focused on prompt construction or general interaction patterns (e.g., Cheng et al., 2024; Nguyen et al., 2024), our approach contributes to this emergent research by focusing specifically on how these interactions can be evaluated to understand what a student knows about a subject matter. To do so, we developed and tested the DRIVE framework and its associated taxonomy, intended as practice-oriented tools for educators working within GenAI-compatible classrooms.

In response to our first research question (RQ1), we sought to validate this process-focused assessment. We found a significant positive relationship between traditional essay scores and our GenAI interaction quality evaluations. This finding provides initial empirical validation for the DRIVE framework's criteria (DRI and VE), demonstrating that a process-focused assessment aligns with established learning outcomes and supports the feasibility of this approach. It also introduces transparency into an environment where the final product can be ambiguous. The study's contribution is further defined by its implementation within an authentic, high-stakes educational setting. By formally assessing student-GenAI interaction logs as graded coursework, our study provides evidence on the real-world application of process-focused evaluation, showing how these interactions expose the collaborative process in a way the final output alone cannot.

Our second research question (RQ2) investigated the specific interaction patterns associated with student mastery, which we defined in two distinct ways: essay mastery, based on the final essay score, and GenAI interaction mastery, based on the quality of the interaction log. When mastery was measured by traditional essay evaluations, the favored patterns were systematic text refinement and analytical evaluation of AI outputs. This finding aligns with multiple studies. For example, it mirrors the "Structured Adaptivity" profile of high-performing writers identified by Nguyen et al. (2024), who engaged in an iterative and critical process. It is also consistent with Cheng et al. (2024) analysis, where writers with higher ownership used targeted AI modifications. In contrast, our finding that below-average essay scores are connected with basic information retrieval and passive workflows corresponds with both Nguyen et al.'s "Unstructured Streamline" profile and Cheng et al.'s observation that writers with lower ownership relied more heavily on directly accepting AI suggestions. Our process-focused GenAI interaction evaluations revealed a distinct pattern, rewarding conceptual refinement and the development of user-generated ideas. This pattern bears resemblance to the high AI literacy behaviors documented by J. Kim et al. (2025), where context-rich, descriptive prompting led to better outcomes. The low-scoring interactions in our process-focused evaluation, which reflected basic task specification without engaging the AI as a thinking partner, are comparable to Nguyen et al.'s (2024) observation that a more linear and uncritical use of AI related to lower performance.This finding is further strengthened by the causal evidence from Yang, Raković, et al. (2025), who demonstrated that actively editing AI suggestions improves essay quality. In contrast, our finding that below-average essay scores are connected with passive GenAI-assisted workflows aligns with the observation across these studies that a linear, uncritical acceptance of AI suggestions relates to lower performance. Our work also provides a new lens through which to interpret the "performance versus learning" paradox highlighted by Yang, Fan, et al. (2025). Their study identified this challenge by analyzing high-level self-regulated learning strategies, finding that high-performing students who heavily integrated GenAI also showed fewer complex metacognitive tactics. This raises the important question of whether high scores always reflect deep learning. Our DRIVE framework offers a complementary, more granular approach to investigate this paradox. The VE component, by design, analyzes the content of the interaction dialogue for evidence of domain-specific learning. Our taxonomy gives teachers a tool to distinguish between interactions that primarily leverage the AI for a high-quality output and those that show the student actively thinking with and through the tool. The "collaborative intellectual partnership" pattern we found associated with high GenAI interaction mastery, characterized by original ideas (Content_Idea) and critical evaluation (Content_Critical), is a tangible example of the kind of engagement our framework can help identify as evidence of this deeper type of learning.

Finally, the systematic, though modest, differences between the two assessment methods (RQ2b) highlight that each approach offers a particular perspective from which to evaluate student work. This observation connects with ongoing discussions about assessment in technology-enhanced learning (e.g., Swiecki et al., 2022). Our findings suggest that the focus of the assessment, specifically whether one grades the process (interaction log) or the product (essay), can interact with the writing task to shape what types of GenAI interactions are ultimately recognized and rewarded.

10.1. Implications for educational practice

Building on our findings, this section provides practical recommendations for teachers in writing-intensive courses where GenAI use is permitted. Our research shows that traditional essay assessment and GenAI interaction evaluation emphasize different aspects of student work. This observation presents teachers with a practical consideration: how to effectively assess both the quality of written outputs and the quality of the collaborative process. We found that combining traditional writing assessment with interaction log evaluation captures complementary aspects of student work. Traditional assessment identified strengths in text refinement and conceptual integration, while interaction log evaluation revealed critical thinking processes and sophisticated AI collaboration strategies not always evident in the final text. For teachers concerned with comprehensive assessment, examining both provides a more complete picture of student competencies. Our data showed specific interaction patterns associated with different types of mastery. In argumentative writing contexts, we observed that targeted, purposeful AI collaboration correlated with higher essay scores, while exploratory, conceptual development correlated with higher interaction quality scores. Teachers may want to consider these patterns when designing assessments for AI-integrated writing assignments. If a decision is made to assess how students use GenAI in a course, AI-related grading rubrics should consider distinguishing between different types

of AI interaction patterns based on course learning goals. Additionally, teachers should consider how writing genre influences GenAI usage strategies, in light of the findings by Cheng et al. (2024) showing how creative and argumentative writing were associated with distinct profiles of GenAI use. During our research, teachers observed a marked decrease in students willingly adopting GenAI after they began formally grading their GenAI interactions (see footnote 1). Understanding student perspectives about process-focused assessment may therefore be valuable before implementation. Finally, while automated classification may eventually assist with log evaluation, human oversight remains essential for accurately assessing sophisticated collaboration. Given GenAI's rapid evolution, teachers should actively engage with educational research to adapt their practices thoughtfully (e.g., Bauer et al., 2025; Theophilou et al., 2023).

10.2. Limitations and future directions

The current work has several limitations that point to opportunities for future research. First, our research was conducted mainly in philosophy courses at one university, which limits the generalizability of our taxonomy to other disciplines or educational contexts.

Second, our findings are subject to self-selection bias as students voluntarily opted into the AI-use condition. This opt-in group may differ from non-users in unmeasured ways (e.g., motivation, risk tolerance, or prior experience), which limits the generalizability of our results. A third limitation concerns our research design. The current study compares the final essays students wrote with GenAI to the interaction logs that produced them. This design, while ecologically valid, does not allow us to compare an individual student's performance with and without AI assistance. Future studies could employ a more controlled experimental design to isolate the specific impact of GenAI on writing quality for the same student, though this may come at the cost of authenticity. Our taxonomy also requires further refinement. The main categories (Writing, Content, Argument) apply broadly, but its subcategories need discipline-specific adjustments. For example, a psychology course may focus more on content knowledge (e.g., Content_Research) than argumentation. Furthermore, there is an imbalance in the inter-rater agreement across different classifications. Our data suggests that the interaction patterns characteristic of high mastery may be less stable (i.e., have lower inter-rater agreement) than other patterns. This highlights the need for further validation and replication studies across diverse contexts to stabilize the taxonomy, refine or remove items with consistently low agreement, and integrate new items based on emerging GenAI literacy frameworks (e.g., see Jin et al., 2025). Another limitation is that our study captures GenAI interaction patterns at a specific technological moment. Future work could identify which interactions from our taxonomy might become obsolete as GenAI technology advances (e.g., Content_Bibliography) versus which remain stable indicators of learning despite technological change (e.g., Content_Critical). Finally, the potential for "meta-prompting" (i.e., fabricating user engagement logs) represents a threat to the validity of our assessment method. A potentially interesting future direction to address could involve tracing the conceptual contributions from the interaction log to the final submitted essay. This analysis would help differentiate performative interactions from genuine engagement. As a complementary method, future studies could investigate how student explanations of their GenAI prompting strategies reveal metacognitive awareness and strategic decision-making that interaction logs alone might not capture (e.g., see also Nikolic et al., 2023). Tool variation across students (ChatGPT versions, Claude, unreported models) also limits replicability, as different AI capabilities may influence interaction patterns. Future studies should systematically document AI tools used while balancing research control with authentic classroom choice.

11. Conclusion

The growing intermix of human and machine input in student writing has made the final essay an ambiguous signal of learning. Our research offers an initial, practice-oriented solution to this challenge with the DRIVE framework and its associated taxonomy. These tools help educators harmonize assessment with the current technological disruption by shifting the focus from the ambiguous final product to the observable learning process. The long-term relevance of any specific assessment method will evolve with technology. Therefore, the more enduring contribution of this work is the insight it provides for the future development of how learning is assessed in the age of AI. The current work helps inform the ongoing design of assessment practices that can effectively and meaningfully evaluate evidence of learning in the emergent reality of human-AI educational partnerships.

CRediT authorship contribution statement

Manuel Oliveira: Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Carlos Zednik:** Writing – review & editing, Validation, Supervision, Resources, Funding acquisition, Conceptualization. **Gunter Bombaerts:** Writing – review & editing, Funding acquisition. **Bert Sadowski:** Funding acquisition. **Rianne Conijn:** Writing – review & editing, Validation, Supervision, Resources, Funding acquisition.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work the author(s) used several Generative AI chatbots (Microsoft Copilot, Gemini, Claude) in order to structure and improve the readability of the text throughout. These tools were also used to assist with R coding (e.g., improved visualization and structuring, efficient refactoring). After using this tool/service, the author(s) critically reviewed and edited the content as needed and take (s) full responsibility for the content of the publication.

Funding

This research was co-funded by the 4TU.Centre for Engineering Education and BOOST! Education Innovation Program of Eindhoven University of Technology.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We are immensely grateful to all the dedicated research assistants (Nelke Engels), teachers (Gijs van Maanen, Patrik Hummel, Tijn Borghuis), and teaching assistants (Céline Budding, Kaush Kalidindi) who collected, scored, and annotated the data for this project. We also thank our colleagues from the Human Technology Interaction and Philosophy & Ethics groups for their sharp and insightful comments at several stages of this project.

Appendix A. Taxonomy to evaluate student-GenAI interactions

Category	Type	Meaning
Writing	Instructions	User specifies the task, in terms of the course's assignment description (e.g. copy-paste or upload)
	Criteria	User specifies the task in more detail, by providing the evaluation criteria for the assignment, from the assignment rubric (usually, copy-paste)
	Evaluate	User asks the machine to evaluate a draft against the provided criteria (or without criteria).
	Improve	User provides a phrase, paragraph, or essay to be improved by the machine for e.g. spelling, style or grammar.
	Format	User asks for improved formatting (including e.g. bibliographical formatting)
	Organization	User asks for feedback or improvement of essay structure.
	Introduction	User asks the machine to provide an effective introduction.
	Conclusion	User asks the machine to provide an effective conclusion.
	Role	User specifies the role/character/expertise the language model should take.
	AutoComplete	User asks machine to append or expand on text, without providing specific guidance about the content.
	Summarize	User asks machine to summarize text (e.g. an uploaded article).
	Content Removal	User ask machine to delete existing text (e.g., deleting a specific paragraph or sentence)
	Miscellaneous	User prompting system in a non-specific technical way.
Content	Bibliography	User asks for bibliographic references on a specific topic.
	Example	User asks the machine to provide specific example for a general case or issue.
	Research	User asks the machine to define an idea, or to identify related ideas to one, given by the user.
	Definitions	User provides the machine with definitions to/elaborations of key technical terms discussed in the course (e.g. "data activism").
	Case	User describes a relevant case from class/their own research.
	Idea	User provides well-motivated original idea or question and asks for confirmation/elaboration/discussion.
	Concept	User introduces a keyword concept from the course material and asks the machine to define it or apply it to a case.
	Elaboration	User provides a relevant sentence/paragraph and asks the machine to elaborate and provide additional detail, mentioning specific course-related content.
	Theory	User asks the machine to appeal to a philosophical or ethical theory (e.g. consequentialism), named or not.
	Critical	User critically engages with AI-generated content, asking for clarification or correction
Argument	Context	User asks the machine to describe or analyze the context of a real world case, technology, or news story. E.g. setting the case into a broader debate.
	Case Research	User asks the machine to describe or analyze the details of a given case.
	Stakeholders	User asks the machine to identify the stakeholders for a case or technology.
	Values	User asks the machine to specify the values of the stakeholders in a case.
	Moral Problem	User asks the machine to formulate a moral problem or identify an ethical issue with a particular case or technology
	Objection	User asks the machine to provide an objection and/or a response to a given claim.
	Justify	User asks the machine to provide reasons for a given claim
	Structure	User asks the machine to impose a particular logical structure onto a text.
	Improve	User asks the machine to improve the argumentative structure (according to given criteria).
	Relate	User asks the machine to relate or connect two concepts or ideas.
	Conceptual Clarity	User asks the machine to simplify or otherwise improve the definition of concepts.
	Thesis	User asks the machine to make a thesis/conclusion more precise, concise, or clear.

References

Amirizaniani, M., Martin, E., Sivachenko, M., Mashhadi, A., & Shah, C. (2024). Can LLMs reason like humans? Assessing theory of mind reasoning in LLMs for open-ended questions. *Proceedings of the 33rd ACM international conference on information and knowledge management*. https://doi.org/10.1145/3627673.3679832

Anderson, L. W., & Krathwohl, D. R. (2001). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives*. Addison Wesley Longman.

Andrews, R. (2015). Critical thinking and/or argumentation in higher education. In M. Davies, & R. Barnett (Eds.), *The palgrave handbook of critical thinking in higher education* (pp. 49–62). Palgrave Macmillan US. https://doi.org/10.1057/9781137378057_3.

Batista, J., Mesquita, A., & Carnaz, G. (2024). Generative AI and higher education: Trends, challenges, and future directions from a systematic literature review. *Information, 15*(11), Article 11. https://doi.org/10.3390/info15110676

Bauer, E., Greiff, S., Graesser, A. C., Scheiter, K., & Sailer, M. (2025). Looking beyond the hype: Understanding the effects of AI on learning. *Educational Psychology Review, 37*(2), 45. https://doi.org/10.1007/s10648-025-10020-8

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big?. *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*. https://doi.org/10.1145/3442188.3445922

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, 57*(1), 289–300. https://doi.org/10.2307/2346101

Bereiter, C., & Scardamalia, M. (2013). *The psychology of written composition*. Routledge.

Biggs, J. (1996). Enhancing teaching through constructive alignment. *Higher Education, 32*(3), 347–364. https://doi.org/10.1007/BF00138871

Casal, J. E., & Kessler, M. (2023). Can linguists distinguish between ChatGPT/AI and human writing?: A study of research ethics and academic publishing. *Research Methods in Applied Linguistics, 2*(3), Article 100068. https://doi.org/10.1016/j.rmal.2023.100068

Cash, T. N., & Oppenheimer, D. M. (2024). Generative Chatbots ain't experts: Exploring cognitive and metacognitive limitations that hinder expertise in generative Chatbots. *Journal of Applied Research in Memory and Cognition, 13*(4), 490–494. https://doi.org/10.1037/mac0000202

Cheng, Y., Lyons, K., Chen, G., Gašević, D., & Swiecki, Z. (2024). Evidence-centered assessment for writing with generative AI. *Proceedings of the 14th learning analytics and knowledge conference*. https://doi.org/10.1145/3636555.3636866

Chi, M. T. H., & Wylie, R. (2014). The ICAP framework: Linking cognitive engagement to active learning outcomes. *Educational Psychologist, 49*(4), 219–243. https://doi.org/10.1080/00461520.2014.965823

Erduran, S., Simon, S., & Osborne, J. (2004). TAPping into argumentation: Developments in the application of Toulmin's argument pattern for studying science discourse. *Science Education, 88*(6), 915–933.

Fanni, R., Steinkogler, V. E., Zampedri, G., & Pierson, J. (2023). Enhancing human agency through redress in Artificial Intelligence systems. *AI & Society, 38*(2), 537–547. https://doi.org/10.1007/s00146-022-01454-7

Ferretti, R. P., & Graham, S. (2019). Argumentative writing: Theory, assessment, and instruction. *Reading and Writing, 32*(6), 1345–1357. https://doi.org/10.1007/s11145-019-09950-x

Fleckenstein, J., Meyer, J., Jansen, T., Keller, S. D., Köller, O., & Möller, J. (2024). Do teachers spot AI? Evaluating the detectability of AI-generated texts among student essays. *Computers and Education: Artificial Intelligence, 6*, Article 100209. https://doi.org/10.1016/j.caeai.2024.100209

Gao, C. A., Howard, F. M., Markov, N. S., Dyer, E. C., Ramesh, S., Luo, Y., & Pearson, A. T. (2023). Comparing scientific abstracts generated by ChatGPT to real abstracts with detectors and blinded human reviewers. *npj Digital Medicine, 6*(1), 1–5. https://doi.org/10.1038/s41746-023-00819-6

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

Hase, S., & Kenyon, C. (2007). Heutagogy: A child of complexity theory. *Complicity: An International Journal of Complexity and Education, 4*(1). https://doi.org/10.29173/cmplct8766

Jin, Y., Martinez-Maldonado, R., Gašević, D., & Yan, L. (2025). GLAT: The generative AI literacy assessment test. *Computers and Education: Artificial Intelligence, 9*, Article 100436. https://doi.org/10.1016/j.caeai.2025.100436

Kearsley, G. P. (1976). Questions and question asking in verbal discourse: A cross-disciplinary review. *Journal of Psycholinguistic Research, 5*(4), 355–375. https://doi.org/10.1007/BF01079934

Kim, M., Kim, S., Lee, S., Yoon, Y., Myung, J., Yoo, H., Lim, H., Han, J., Kim, Y., Ahn, S.-Y., Kim, J., Oh, A., Hong, H., & Lee, T. Y. (2024). Designing prompt analytics dashboards to analyze student-ChatGPT interactions in EFL writing (No. arXiv: 2405.19691). *arXiv*. https://doi.org/10.48550/arXiv.2405.19691

Kim, J., Yu, S., Detrick, R., & Li, N. (2025). Exploring students' perspectives on Generative AI-assisted academic writing. *Education and Information Technologies, 30*(1), 1265–1300. https://doi.org/10.1007/s10639-024-12878-7

Kim, J., Yu, S., Lee, S.-S., & Detrick, R. (2025). Students' prompt patterns and its effects in AI-assisted academic writing: Focusing on students' level of AI literacy. *Journal of Research on Technology in Education, 0*(0), 1–18. https://doi.org/10.1080/15391523.2025.2456043

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*(1), 159. https://doi.org/10.2307/2529310

Lin, Z. (2024). How to write effective prompts for large language models. *Nature Human Behaviour, 8*(4), 611–615. https://doi.org/10.1038/s41562-024-01847-2

Lintner, T. (2024). A systematic review of AI literacy scales. *Npj Science of Learning, 9*(1), 50. https://doi.org/10.1038/s41539-024-00264-4

Little, C. W., Clark, J. C., Tani, N. E., & Connor, C. M. (2018). Improving writing skills through technology-based instruction: A meta-analysis. *The Review of Education, 6*(2), 183–201. https://doi.org/10.1002/rev3.3114

Long, D., & Magerko, B. (2020). What is AI literacy? Competencies and design considerations. *Proceedings of the 2020 CHI conference on human factors in computing systems*. https://doi.org/10.1145/3313831.3376727

Lyons, H., Velloso, E., & Miller, T. (2021). Conceptualising contestability: Perspectives on contesting algorithmic decisions. *Proceedings of the ACM on Human-Computer Interaction, 5*(CSCW1), 106:1–106:25. https://doi.org/10.1145/3449180

Maleki, N., Padmanabhan, B., & Dutta, K. (2024). AI hallucinations: A misnomer worth clarifying. *2024 IEEE conference on artificial intelligence (CAI)*. https://doi.org/10.1109/CAI59869.2024.00033

Marton, F., & Saljo, R. (1976). On qualitative differences in learning: I. Outcome and process. *British Journal of Educational Psychology, 46*(1), 4–11. https://doi.org/10.1111/j.2044-8279.1976.tb02980.x

Miyake, N., & Norman, D. A. (1979). To ask a question, one must know enough to know what is not known. *Journal of Verbal Learning and Verbal Behavior, 18*(3), 357–364. https://doi.org/10.1016/S0022-5371(79)90200-7

Molinero, R. I., & García-Madruga, J. A. (2011). Knowledge and question asking. *Psicothema*, 26–30.

Newell, G. E., Beach, R., Smith, J., & Van Der Heide, J. (2011). Teaching and learning argumentative reading and writing: A review of research. *Reading Research Quarterly, 46*(3), 273–304. https://doi.org/10.1598/RRQ.46.3.4

Nguyen, A., Hong, Y., Dang, B., & Huang, X. (2024). Human-AI collaboration patterns in AI-assisted academic writing. *Studies in Higher Education, 49*(5), 847–864. https://doi.org/10.1080/03075079.2024.2323593

Nikolic, S., Daniel, S., Haque, R., Belkina, M., Hassan, G. M., Grundy, S., Lyden, S., Neal, P., & Sandison, C. (2023). ChatGPT versus engineering education assessment: A multidisciplinary and multi-institutional benchmarking and analysis of this generative artificial intelligence tool to investigate assessment integrity. *European Journal of Engineering Education, 48*(4), 559–614. https://doi.org/10.1080/03043797.2023.2213169

Nussbaum, E. M., & Schraw, G. (2007). Promoting argument-counterargument integration in students' writing. *The Journal of Experimental Education*. https://doi.org/10.3200/JEXE.76.1.59-92

Pigg, S. (2024). Research writing with ChatGPT: A descriptive embodied practice framework. *Computers and Composition, 71*, Article 102830. https://doi.org/10.1016/j.compcom.2024.102830

R Core Team. (2024). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing (Version 4.3.3) [Computer software].

Razi, A., Bouzoubaa, L., Pessianzadeh, A., Seberger, J. S., & Rezapour, R. (2025). *Not a Swiss army knife: Academics' perceptions of trade-offs around generative artificial intelligence use* (No. arXiv:2405.00995). *arXiv*. https://doi.org/10.48550/arXiv.2405.00995

Reeves, T. C., Herrington, J., & Oliver, R. (2005). Design research: A socially responsible approach to instructional technology research in higher education. *Journal of Computing in Higher Education, 16*(2), 96–115. https://doi.org/10.1007/BF02961476

Ritchhart, R. (2011). *Making thinking visible: How to promote engagement, understanding, and independence for all learners*. Jossey-Bass.

Sampson, V., & Clark, D. B. (2008). Assessment of the ways students generate arguments in science education: Current perspectives and recommendations for future directions. *Science Education, 92*(3), 447–472.

Shojaee, P., Mirzadeh, I., Alizadeh, K., Horton, M., Bengio, S., & Farajtabar, M. (2025). The illusion of thinking: Understanding the strengths and limitations of reasoning models via the lens of problem complexity. *arXiv Preprint arXiv:2506.06941*, 1–37.

Swiecki, Z., Khosravi, H., Chen, G., Martinez-Maldonado, R., Lodge, J. M., Milligan, S., Selwyn, N., & Gašević, D. (2022). Assessment in the age of artificial intelligence. *Computers and Education: Artificial Intelligence, 3*, Article 100075. https://doi.org/10.1016/j.caeai.2022.100075

The Design-Based Research Collective. (2003). Design-Based research: An emerging paradigm for educational inquiry. *Educational Researcher, 32*(1), 5–8. https://doi.org/10.3102/0013189X032001005

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). Nature Switzerland: Springer. https://doi.org/10.1007/978-3-031-47546-7_33.

Toulmin, S. E. (1958). *The uses of argument (Repr. of updated ed)*. Cambridge University Press.

von Glasersfeld, E. (1989). Cognition, construction of knowledge, and teaching. *Synthese, 80*(1), 121–140. https://doi.org/10.1007/BF00869951

Wahn, B., Schmitz, L., Gerster, F. N., & Weiss, M. (2023). Offloading under cognitive load: Humans are willing to offload parts of an attentionally demanding task to an algorithm. *PLoS One, 18*(5), Article e0286102. https://doi.org/10.1371/journal.pone.0286102

Wang, B., Rau, P.-L. P., & Yuan, T. (2023). Measuring user competence in using artificial intelligence: Validity and reliability of artificial intelligence literacy scale. *Behaviour & Information Technology, 42*(9), 1324–1337. https://doi.org/10.1080/0144929X.2022.2072768

Wingate, U. (2012). 'Argument!'helping students understand what essay writing is about. *Journal of English for Academic Purposes, 11*(2), 145–154.

Wolfe, C. R., Britt, M. A., & Butler, J. A. (2009). Argumentation schema and the myside bias in written argumentation. *Written Communication, 26*(2), 183–209. https://doi.org/10.1177/0741088309333019

Yan, L., Martinez-Maldonado, R., & Gasevic, D. (2024). Generative artificial intelligence in learning analytics: Contextualising opportunities and challenges through the learning analytics cycle. *Proceedings of the 14th learning analytics and knowledge conference*. https://doi.org/10.1145/3636555.3636856

Yang, K., Cheng, Y., Zhao, L., Raković, M., Swiecki, Z., Gašević, D., & Chen, G. (2024). *Ink and algorithm: Exploring temporal dynamics in human-AI collaborative writing* (Version 1). *arXiv*. https://doi.org/10.48550/ARXIV.2406.14885

Yang, K., Fan, Y., Tang, L., Raković, M., Li, X., Gašević, D., & Chen, G. (2025). *Beyond self-regulated learning processes: Unveiling hidden tactics in generative AI-assisted writing* (No. arXiv:2508.10310). *arXiv*. https://doi.org/10.48550/arXiv.2508.10310

Yang, K., Raković, M., Liang, Z., Yan, L., Zeng, Z., Fan, Y., Gašević, D., & Chen, G. (2025). Modifying AI, enhancing essays: How active engagement with generative AI boosts writing quality. *Proceedings of the 15th international learning analytics and knowledge conference* (pp. 568–578). https://doi.org/10.1145/3706468.3706544

Zhang, R., & Zou, D. (2022). Types, features, and effectiveness of technologies in collaborative writing for second language learning. *Computer Assisted Language Learning, 35*(9), 2391–2422. https://doi.org/10.1080/09588221.2021.1880441

Zheng, J., Hao, L., Lu, K., Garg, A., Reese, M., Yap, M.-J., Wang, I.-J., Wu, X., Huang, W., Hoffman, J., Kelly, A., Le, M., Zhang, R., Lin, Y., Faayez, M., & Liu, A. (2025). *Do students rely on AI? Analysis of student-ChatGPT conversations from a field study* (Version 1). *arXiv*. https://doi.org/10.48550/ARXIV.2508.20244