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

Manuel Oliveira[1], Carlos Zednik[1], Gunter Bombaerts[1], Bert Sadowski[1], and Rianne Conijn[1]

[1]Department of Industrial Engineering and Innovation Sciences, Eindhoven University of Technology, The Netherlands

Last updated: 16 July 2025

Corresponding author: Manuel Oliveira (m.j.barbosa.de.oliveira@tue.nl), Human Technology Interaction, Department of Industrial Engineering and Innovation Sciences, Eindhoven University of Technology, Den Dolech, Eindhoven, 5200MB, The Netherlands.

Funding

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

CRediT statement

The following contributions are defined according to the CRediT (Contributor Roles Taxonomy): Conceptualization: MO, CZ; Data Curation: MO; Formal Analysis: MO; Funding Acquisition: MO, RC, CZ, GB, BS; Investigation: MO; Methodology: MO; Project Administration: MO; Resources: MO, CZ, RC; Supervision: MO, RC, CZ; Validation: MO, RC, CZ; Visualization: MO; Writing – Original Draft: MO; Writing – Review & Editing: MO, RC, CZ, GB.

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.

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.

Declaration of competing interests

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.

Highlights

- Developed a framework to assess learning through GenAI interactions.
- Created a taxonomy aligning GenAI usage with learning objectives.
- GenAI interaction evaluations positively correlated with traditional essay evaluations.
- High-scoring interactions linked to using GenAI for idea co-development.
- High-scoring essays linked to using GenAI for text refinement.
- Assessment focus emphasizes distinct GenAI use strategies.

Abstract

As generative AI (GenAI) transforms how students learn and work, higher education must rethink its assessment strategies. This paper presents a taxonomy and conceptual framework (DRIVE) to evaluate student learning from GenAI interactions (prompting strategies), focusing on cognitive engagement (Directive Reasoning Interaction) and knowledge infusion (Visible Expertise). Despite extensive research mapping student GenAI writing behaviors, practical tools for assessing domain-specific learning remain underexplored. This paper shows how GenAI interactions inform such learning in authentic classroom contexts, moving beyond technical skills or low-stakes assignments. We conducted multi-methods analysis of GenAI interaction annotations ($n = 1450$) from graded essays ($n = 70$) in STEM writing courses. A strong positive correlation was found between high-quality GenAI interactions and final essay scores, validating the feasibility of this assessment approach. Furthermore, our taxonomy revealed distinct 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 offers educators and researchers a practical tool to design assessments that capture learning in AI-integrated classrooms.

Keywords: Learning, Assessment, Academic Writing, Generative AI

1 Introduction

The emergence of generative artificial intelligence (GenAI) in higher education has fundamentally disrupted traditional methods of assessing student learning, raising questions about adapting learning objectives to this technological shift (e.g., Bower et al., 2024; Xia et al., 2024). This is especially true in text-based assessments, as current GenAI applications produce academic writing increasingly indistinguishable from human work (Casal & Kessler, 2023; Clark et al., 2021; Fleckenstein et al., 2024; Porter & Machery, 2024). With students increasingly engaging in dialogues with systems like ChatGPT (OpenAI, 2022) to develop their academic work (e.g., Ansari et al., 2024), conventional output-focused assessment cannot effectively measure writing skills acquisition (e.g., Swiecki et al., 2022; Yan, 2023) or domain-specific knowledge (i.e., specific to a given academic discipline, e.g., psychology, economics, ethics). This shift in how academic work is produced created a strong incentive to reimagine assessment practices, in light of how the rapid proliferation of GenAI across higher education has rendered traditional evaluation methods ineffective. The rise of AI in education triggered scholarly discussions advocating for a shift from evaluating the final product to analyzing the learning process (e.g., Swiecki et al., 2022). This emphasis on the process provides a more transparent record of student engagement and reasoning.

Despite the proliferation of research on the educational impact of GenAI, a gap persists in understanding how these tools mediate learning. Early research documented student-GenAI interaction patterns in non-classroom contexts (e.g., Cheng et al., 2024; Pigg, 2024), establishing frameworks for categorizing behaviors such as requesting, refining, and evaluating content (Pigg, 2024), or distinguishing knowledge telling from knowledge transformation based on AI suggestion modifications (Cheng et al., 2024). More recently, research has progressed beyond description to examining how interaction patterns serve as learning indicators through experimental academic tasks. For instance, Nguyen et al. (2024) found doctoral students using iterative, interactive AI collaboration (involving research, critical editing, and thoughtful prompting) achieved higher writing performance than those employing linear, uncritical approaches with minimal revision. In a similar vein, Kim et al. (2025) classified student prompts using the well-known Bloom's Taxonomy (Anderson & Krathwohl, 2001) and identified distinct patterns related to AI literacy: high-literacy students employed descriptive, context-rich prompts across Bloom's cognitive levels and viewed AI as an ideational collaborator, whereas low-literacy students used general prompts for lower-order tasks in brief interactions.

The findings from these studies start to shed light on what might be the most effective interactions with GenAI to generate writing evaluated as high-quality. Thus far, this literature consistently suggests that higher levels of cognitive engagement with GenAI as a collaborative "partner" connects with higher-quality writing, while a more reluctant and linear use of the technology connects with lower-quality writing. Existing studies provide valuable insights on student-GenAI interaction but primarily examine controlled settings rather than authentic classrooms where the use of GenAI directly impacts grades. Prior research has documented student-GenAI interaction patterns in experimental contexts, but tools for applying these insights to assess learning in actual courses remain underexplored. Educators who allow GenAI in their classrooms can use additional resources to analyze interaction logs and evaluate writing skill development in these new contexts.

This paper offers such a resource by introducing a conceptual framework to guide the evaluation of learning through GenAI interactions, along with a taxonomy that guides

assessment of how these interactions reveal progress toward learning objectives in authentic classroom contexts. Our framework builds on two core principles: assessing students' active steering of AI dialogue and evaluating how they make their unique knowledge observable within these interactions. To validate this approach, we test whether these interaction patterns correlate with traditional learning outcomes, namely essay scores, providing initial evidence for their use as learning proxies. Our methodology 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 (text generation, basic argumentation) and struggles with (critical evaluation, integrating personal understanding), given its limitations in comprehending its own outputs (West et al., 2023). By systematically analyzing student-GenAI engagement throughout the writing process we can identify the types of interactions that are associated with evidence of learning. Our GenAI interaction evaluation framework and the taxonomy offers educators a practical tool to design assessments that capture learning in AI-integrated classrooms.

2 Background

2.1 The skill of argumentative writing

To contextualize the development of the taxonomic framework, we must first consider the nature of the academic skill it aims to evaluate: argumentative writing. Argumentative writing represents a foundational academic skill that extends beyond mere text composition to also involving critical thinking, evaluation of evidence, and logical reasoning (Andrews, 2015; Newell et al., 2011). Traditional 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 designed by the teacher, which typically 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 worth revisiting in the current discussion. Toulmin (1958)'s 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., Kim et al., 2025) but, to the best of our knowledge,

few studies have systematically analyzed how different patterns of GenAI interaction correlate with learning outcomes in the specific domain of argumentative writing.

2.2 Evidence of learning in the age of GenAI

Several theoretical educational frameworks have been guiding educators' understanding of teaching and learning over the past decades. A well-known perspective is the distinction proposed by Marton and Saljo (1976) between surface learning, focused on rote memorization, and deep learning, which involves actively seeking meaning, integrating new knowledge, and transforming understanding. Complementing this, the widely adopted Bloom's Taxonomy (Anderson & Krathwohl, 2001) offers a hierarchical structure for categorizing cognitive skills in a pedagogical context. This hierarchy ascends from lower-order thinking skills such as Remembering (recalling facts and basic concepts) and Understanding (explaining ideas or concepts), to higher-order thinking skills like Applying (using information in new situations), Analyzing (drawing connections among ideas, breaking material into constituent parts), Evaluating (justifying a stand or decision, critiquing), and Creating (producing new or original work). Educators often use these levels to design learning objectives and assessments for their courses (e.g., Britto & Usman, 2015). Evidence of learning is commonly inferred from a student's ability to demonstrate skills at the higher end of the taxonomy. For instance, an essay that not only recalled information but also analyzed different perspectives and created a new synthesis would be seen as indicative of "deeper" learning and more sophisticated cognitive processing.

These frameworks have historically guided the assessment of student work, while frequently focusing on the final product as the primary evidence of these cognitive processes. However, the advent of GenAI requires a shift in focus. When students use GenAI for their coursework, the final product alone offers an increasingly ambiguous signal of their learning, as it becomes challenging to disentangle the student's contribution from the AI's. One potential approach to circumvent this challenge might involve searching for learning evidence in the interaction process between a writer and AI systems, through the examination of how students steer these systems, how they evaluate their output, or decide to incorporate it in their writing. Existing frameworks are primarily designed to evaluate the output alone, and thus, cannot adequately capture these nuanced interaction strategies or reveal the depth of student agency and critical engagement throughout the AI-assisted writing process.

3 DRIVE framework

With the aim of creating a tool to evaluate the learning process in writing-intensive courses where the use of GenAI is considered acceptable, we propose a conceptual framework composed of two evaluative components: Directive Reasoning and Interaction (DRI), and Visible Expertise (VE), or DRIVE. The primary purpose of this framework is to provide guidance on how to assess evidence of academic writing skill acquisition through the systematic examination of the interaction process between a student and a GenAI system during AI-assisted writing. It specifically seeks to identify behaviors indicative of what the student knows and how their actions produce visible evidence of skill acquisition, such as understanding domain-relevant theories, engaging in critical thinking, and introducing original, user-generated ideas into the dialogue with AI. At its core, this framework posits that a crucial step in assessing AI-assisted writing processes is to observe the extent to

which students actively and purposefully steer the interaction with GenAI, thereby making their learning process, knowledge, and critical thinking visible. Below, we describe in more detail the two evaluative components of DRIVE.

3.1 Directive Reasoning Interaction (DRI)

This component evaluates how actively and purposefully the student steers the interaction with the AI. It echoes the ideas of heutagogy, a framework of self-determined learning (Hase & Kenyon, 2007). Heutagogy is concerned with "learner-centred 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 lens of heutagogy allows us to conceptualize GenAI as a powerful resource that a self-determined learner can direct.

This perspective also aligns with the Interactive, Constructive, Active, and Passive (ICAP) framework by Chi and Wylie (2014) which categorizes learning activities from shallow to deep. While passive engagement involves receiving information without active processing, and active engagement applies existing knowledge for retention, deeper learning arises from constructive and interactive engagement. Constructive engagement involves generating new ideas and outputs beyond learned material, enhancing problem-solving and transversal skills. Interactive engagement, the deepest form, entails collaborative idea generation, leading to novel inferences while fostering communication and collaboration skills. In the context of AI-assisted writing, a student's high Directive Reasoning Interaction (DRI) involves taking a leading role, critically questioning AI outputs, and using one's own reasoning to guide the dialogue. These types of interactions 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 also 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)

This component focuses on the extent to which the student makes their own knowledge, original ideas, and understanding 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. Furthermore, with the rise of GenAI, AI literacy, specifically the skills required to effectively and critically evaluate AI system outputs, is an increasingly important aspect of visible expertise that informs interaction patterns.

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), visible expertise also involves the capacity to critically assess and refine AI outputs, thereby countering risks like automation bias and cognitive offloading.

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

This framework suggests that interaction patterns with high DRI and VE indicate desirable profiles for using GenAI in argumentative writing and other academic tasks. The more these aspects are visible in the interaction logs, the richer the evidence of learning available for assessment in the AI-assisted co-writing process. On the other hand, interactions with limited DRI and VE provide less tangible indicators of active learning and skill acquisition in interaction logs. The DRIVE framework is meant to serve as a conceptual compass and analytical tool for educators, supporting assessment in a manner compatible with GenAI use in the classroom by focusing on the quality of students' intellectual partnership with AI while emphasizing how students can actively steer this interaction and display their learning throughout the process.

4 Overview

4.1 Research Aims

This paper presents the development of a practice-oriented taxonomy for analyzing student-GenAI interactions, which is grounded in the DRIVE framework. Our taxonomy aims to identify strategies of engagement with GenAI technology and explore whether they can provide a meaningful window into student learning during academic writing. The present research is primarily exploratory and descriptive, and is guided by two central questions detailed below.

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

This question seeks to validate our process-focused measure against traditional essay scores. A significant positive association would provide initial evidence that analyzing the interaction process is a valid approach for assessing student learning.

RQ2: What student-GenAI interaction patterns are associated with different levels of 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. 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.

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

This study employs a multi-faceted approach to investigate whether student-GenAI interactions can serve as a meaningful proxy for learning in argumentative writing. Central to our methodology is the development and application of a new taxonomy designed to systematically classify student prompts from real-world classroom settings. 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 analysis proceeded in two phases. For RQ1, we correlated the process- and output-focused performance scores to validate the former. 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 or Master's level courses on philosophy and ethics, covering topics from human-technology interaction to the societal impact of artificial intelligence. In all courses, students were required to individually write a graded argumentative essay. Data were collected across these courses during the 2023-2024 and 2024-2025 academic years. As detailed in Table 1, a total of 445 students were enrolled across these courses. Of these, 103 students (23.2%) chose to use GenAI for their assignments under the condition that they would submit their interaction logs for assessment. The shared interaction log was formally graded using a marking rubric (see Table 2) and contributed to their final course grade. A subset of 70 AI-user essays, along with their corresponding GenAI interaction logs, were annotated using the proposed taxonomy (see Appendix A). A total of 1450 student-GenAI interactions (i.e. prompts) were annotated. The discrepancy between the 103 students who opted to use GenAI and the 70 essays that were annotated results from many interaction logs from AI users being unusable due to issues encountered during data collection and processing. Examples include broken hyperlinks to ChatGPT interaction logs shared by students, or messy screenshots of chat interactions that were difficult to incorporate into the dataset and were ultimately excluded. The "Unknown AI Use" category in Table 1 refers to cases with insufficient information regarding AI tool engagement. Among the 70 annotated AI user essays, ChatGPT was the most predominantly used GenAI tool ($n = 48$, 68.6%). One student (1.4%) used the chatbot Claude, and 21 students (30.0%) did not report their GenAI tool. The prompt-related statistics for the 70 annotated essays, including number of prompts per student (as derived from their interaction logs) and prompt length (as derived from number of characters in prompts), are summarized in Table 1.

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.

5.3 Data Collection Procedure

Over a 10-week period in each course, students completed a graded argumentative essay assignment. 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.

5.5.1 Traditional essay scores

Student essays were evaluated by course instructors using grading rubrics tailored to argumentative writing within the specific course contexts (Data Science Ethics or Philosophy & Ethics of AI). These scores represent an output-focused measure of performance, reflecting the quality of the final written product. Core assessment criteria that were common across these rubrics, independent of any AI tool usage, included: ability to define and contextualize an ethical problem or case relevant to the course; depth of ethical analysis and the construction of well-structured, coherent, and persuasive arguments; demonstration of critical thinking and reflection on complexities and diverse perspectives; effective integration of course concepts and relevant academic literature with proper sourcing; and overall clarity, coherence, structure, and adherence to academic writing style.

5.5.2 GenAI interaction quality scores

The quality of students' interactions with GenAI was assessed experimentally by course teachers and teaching assistants as part of the final course grade for students who chose to use GenAI in their essay assignments, using a set of criteria designed to evaluate GenAI use during argumentative essay writing, with an emphasis on the identification of

learning indicators. These criteria are detailed in Table 2 and cover aspects such as AI for Writing, AI for Argumentation, and AI for Course Content. The criteria are aligned with the proposed taxonomy (see Appendix A). They integrate course learning objectives and teachers' views of interaction quality. 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.

6 Development of the taxonomy

The taxonomy for classifying student-GenAI interactions was developed through an iterative, dual-approach process, carried out by two university teachers (one of whom is a co-author) and teaching assistants, all with expertise in argumentative essay writing. The top-down component of this process was firmly grounded in the intended learning outcomes of the courses under examination, which emphasize the ability to construct well-reasoned arguments, apply ethical reasoning, and critically engage with complex issues. To reflect these goals, the taxonomy was designed to identify how argumentative writing skills manifest in GenAI-supported processes. It organizes interaction patterns into three main categories aligned with core academic competencies: Writing, Content, and Argument. These dimensions were selected because they resonate with established components of argumentative writing. "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 perspectives involved in a given discussion, improving articulation of arguments, strengthening one's thesis).

This top-down, pedagogically-informed structure was complemented and refined by a bottom-up analysis. This involved a hands-on review of actual student GenAI interaction logs, allowing the development team to gain insights into common, real-world user actions and patterns. This iterative process (illustrated in Figure 1) resulted in the final version of the taxonomy, which contains a total of 35 subcategories within three main categories: 13 subcategories under Writing, 10 under Content, and 12 under Argument. The full taxonomy can be found in the Appendix A.

The current taxonomy is informed by an analytical approach to human-AI interaction patterns focused on student agency and knowledge co-construction. The taxonomy development was primarily grounded in the intended learning outcomes of the courses under examination, which emphasize constructing well-reasoned arguments, applying ethical reasoning, and critically engaging with complex issues. The core principles underlying the DRI and VE constructs provided an analytical framework for identifying interaction behaviors that could demonstrate these learning competencies in GenAI-supported

Table 2: Evaluation Criteria for GenAI Interaction Logs in Argumentative Essay Writing.

Criterion	Excellent (10-9)	Good (8-7)	Sufficient (6)	Insufficient (5-0)
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.

argumentative writing processes. To reflect these goals, the taxonomy identifies potential indicators of learning by rethinking how argumentative writing skills manifest in

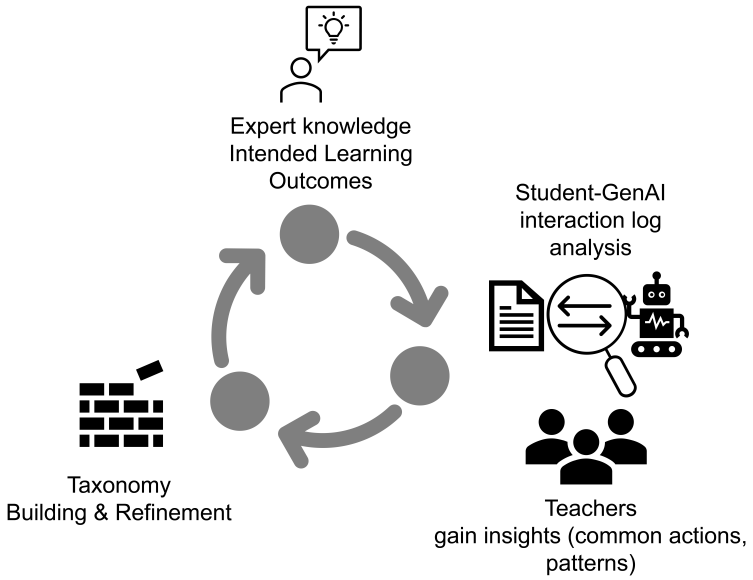

Figure 1: Illustration of the cycle of taxonomy development and refinement.

GenAI-supported processes. It organizes interaction patterns into three main categories aligned with core academic competencies, selected because they resonate with established components of argumentative competence discussed in the literature. The Writing category captures interactions related to textual coherence, structure, and clarity, which are fundamental to conveying an argument effectively, akin to the structural elements often evaluated in traditional rubrics. The Content category addresses how students engage with the substance of their arguments, including the sourcing, evaluation, and integration of information and evidence (echoing the critical use of sources), a process that takes on new dimensions when information is co-constructed with GenAI. Finally, the Argument category directly targets interactions indicative of logical reasoning, the formulation and support of claims, the consideration of counterarguments or rebuttals, and the overall analytical thinking involved in building a persuasive case.

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 $p < .001$). The agreement at the taxonomy subcategory levels (see Figure 2) 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.

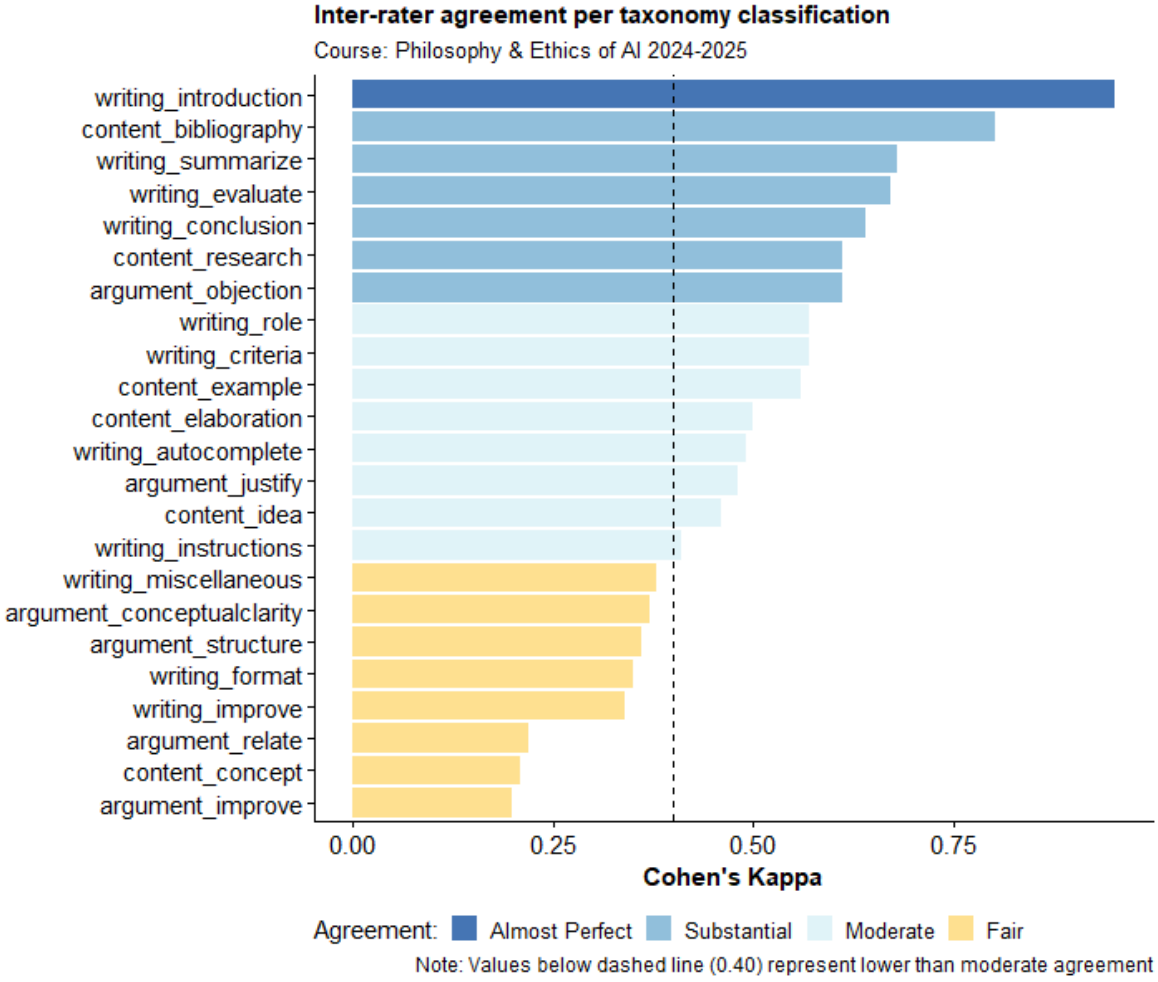

Figure 2: Inter-rater Agreement per Taxonomy Classifications.

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://osf.io/32jg7/?view_only=feed50a5bad04bfab4f5bd60531510e7). 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 (ρ). The additional Spearman's ρ 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:

- **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.

- **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 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 and 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 Figure 3. 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.

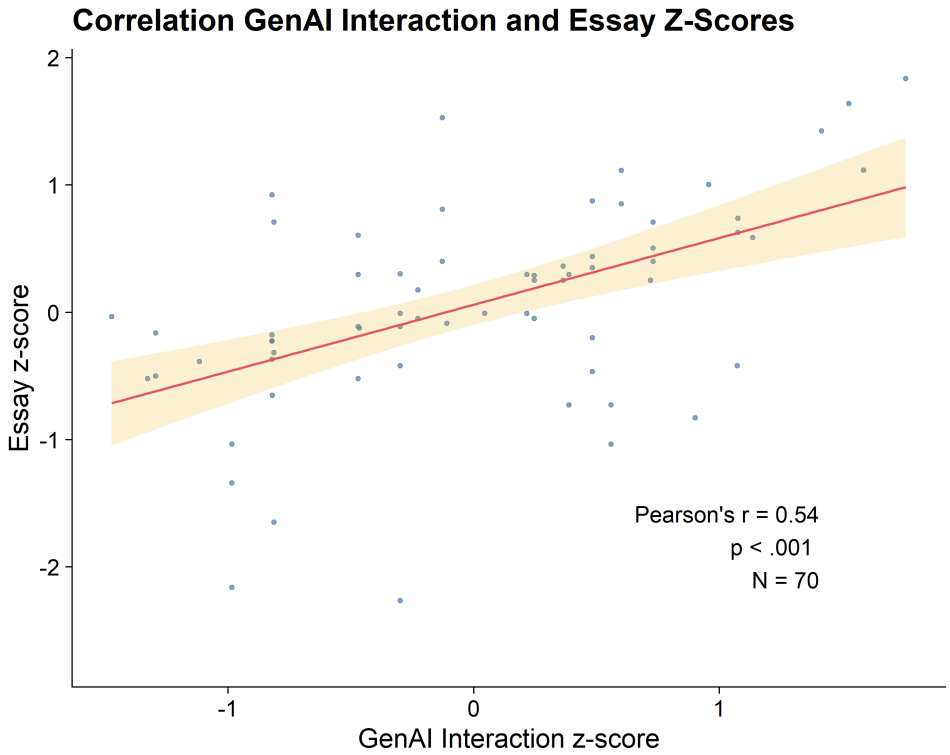

Figure 3: Relationship Between Traditional Essay Scores and GenAI Interaction Evaluation Scores.

9.2.1 Taxonomy descriptives

The frequencies at which taxonomy classifications were observed during the annotation of student-GenAI interaction logs are shown in Figure 4. This figure details the prevalence of the main taxonomy categories (Figure 4-B) and the individual subcategories with over 1% frequency (Figure 4-A). 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%).

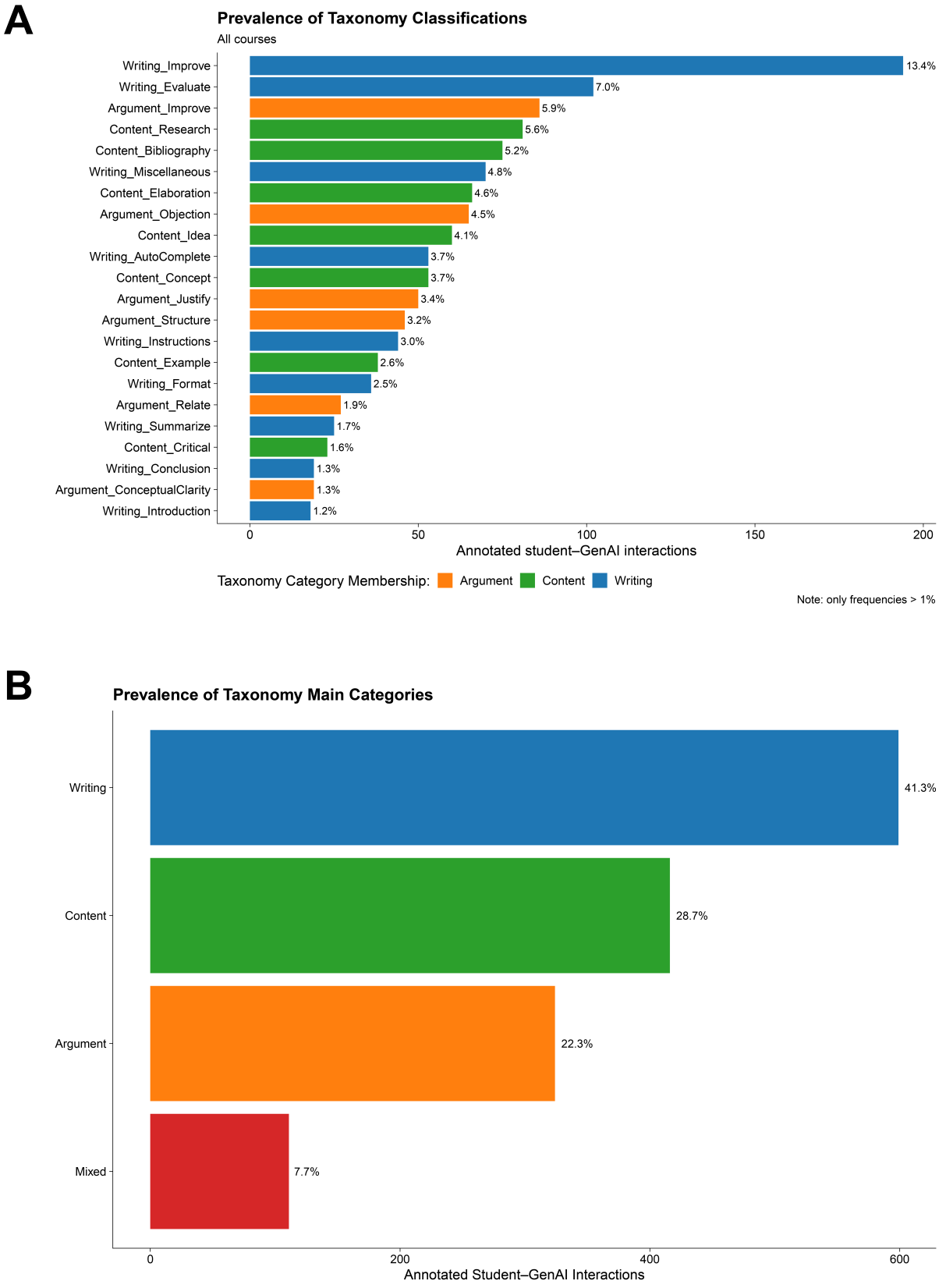

Figure 4: Overall Descriptives of Taxonomy Annotations for All Courses.

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. Figure 5 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.

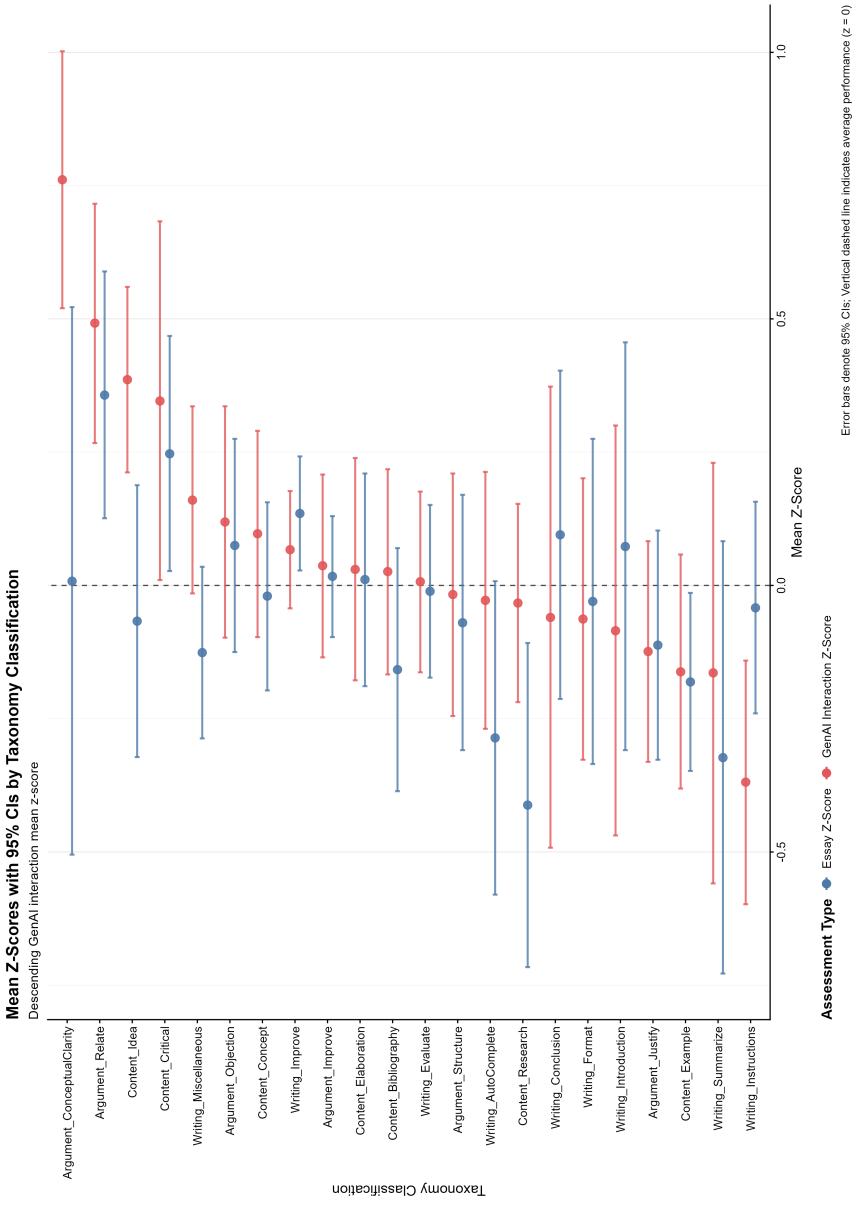

Figure 5: GenAI Interaction Classifications And Mastery Level: Essay and GenAI Interaction Mean Z-Scores + 95% Confidence Intervals Per Taxonomy Classification.

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 ($n = 81$ 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%) 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).

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 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 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 statistical analysis uncovered a more nuanced picture, suggesting additional classifications with significant differences (FDR-corrected). 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, despite 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 investigating whether analyzing student-GenAI interactions could reveal meaningful learning patterns in AI-assisted academic writing, and specifically, argumentative writing. This research shifts the analytical focus from technical skill to the evidence of learning within student-GenAI interactions. Prior work has often focused on the technical aspects of prompting. This includes work on prompt construction (e.g., Chen et al. (2023); Giray (2023); Heston and Khun (2023); Lin (2024); White et al. (2023)) or general interaction patterns (Cheng et al., 2024; Nguyen et al., 2024; Pigg, 2024; Sawalha et al., 2024). Our approach contributes to the emergent body of research by focusing on how interactions with GenAI during writing assignments can be used to assess what a student understands about a subject matter. Using a taxonomy we developed to classify prompting behaviors in light of learning objectives, we examined patterns of student-GenAI engagement associated with mastery levels on both traditional output-focused essay assessments and process-focused evaluations grounded in the

DRIVE (Directive Reasoning Interaction & Visible Expertise) framework.

Our findings support the feasibility of this assessment approach. To our knowledge, this work represents the first formal assessment of student-GenAI interaction logs as graded coursework within authentic classroom contexts. We found a significant positive relationship between traditional essay scores and GenAI interaction quality evaluations, which demonstrates that analyzing the interaction process provides insights that align with established measures such as conventional essay grading. This approach brings transparency into the assessment of written assignments in an environment where there is high uncertainty about the extent to which written content is human-generated. Student-GenAI interactions expose the collaborative process and its influence on final products in ways that traditional assessment of written outputs alone cannot capture.

We applied our taxonomy to reveal distinct interaction profiles associated with different mastery tiers across both assessment approaches. Specifically, we defined mastery in two distinct ways: essay mastery (proficiency demonstrated through final essay quality as evaluated by traditional grading criteria) and GenAI interaction mastery (proficiency in productive engagement with AI tools as assessed through our process-focused evaluation criteria derived from the DRIVE framework). We found that high-performing student-GenAI interactions exhibited sophisticated engagement patterns, with some noticeable differences emerging between the traditional and process-based assessment approaches. Our results revealed distinct patterns in how GenAI interaction strategies connect with learning indicators. Traditional essay evaluations favored systematic text refinement, analytical evaluation of AI outputs, and strategic conceptual integration. These patterns align with Cheng et al. (2024)'s keystroke-level analysis where writers maintaining higher ownership in argumentative contexts engaged in focused self-directed composition with targeted AI modifications, which mirrors the high-scoring essay profile identified in our work. In contrast, our finding that below-average essay scores are connected with basic information retrieval behaviors, correspond with Cheng et al. (2024)'s observation that writers with lower ownership over their writing relied more heavily on directly accepting AI suggestions. This writing genre-specific pattern may explain why, when evaluated through traditional output-focused grading, our argumentative writing context revealed a stronger focus on targeted, purposeful AI collaboration rather than a more explorative collaboration.

Process-focused GenAI interaction evaluations revealed a distinct pattern. High-quality interactions reflected conceptual refinement, development of user-generated ideas, and critical evaluation of AI outputs. This pattern bears resemblance with the high AI literacy behaviors documented by Kim et al. (2025), where descriptive, context-rich prompting led to better writing outcomes. Low-scoring interactions in our process-focused evaluation reflected basic task specification without engaging AI as a thinking partner. This is comparable to Nguyen et al. (2024)'s observation that a more linear and uncritical use of AI related to lower writing performance. While Cheng et al. (2024) found that exploratory behaviors were more prevalent in creative writing than argumentative contexts, our process-focused assessment specifically valued intellectual partnership with AI across both exploratory and targeted refinement activities. Altogether, these patterns may be suggesting how the assessment focus and type of writing genre (or writing goals) can ultimately interact to shape what kinds of (Gen)AI interactions are recognized and rewarded.

The systematic differences between assessment methods, though modest in magnitude, highlight different aspects of student-GenAI engagement. Traditional essay assessment

and process-focused assessment each captured distinct interaction qualities, suggesting that each approach offers a particular lens through which to evaluate student work. This observation connects with ongoing discussions about assessment in technology-enhanced learning environments (e.g., Swiecki et al. (2022)) and how different assessment approaches might emphasize different aspects of student performance.

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.[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. Our research was conducted mainly in philosophy courses at one university, which limits the generalizability of our taxonomy to other disciplines or educational contexts. The main categories of our taxonomy (Writing, Content, Argument) apply broadly to academic writing, but its subcategories require discipline-specific adjustments. For example, an Introduction to Psychology course may focus more on content knowledge (e.g.,

[1] The source document refers to a footnote here, but the text of the footnote was not provided. You can add your note here.

Content_Research) than the ability to persuade through argumentation. The current taxonomy could be further refined by both by removing or redesigning items with low inter-rater agreement, and integrating potentially relevant new items based on emerging GenAI literacy frameworks (e.g., see Jin et al. (2025)).

A limitation of our study is that it 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 as systems gain better access to academic sources) versus which interactions remain relatively stable indicators of learning despite technological change (e.g., Content_Critical, which reflects students' evaluative engagement regardless of interface). Finally, the potential for "meta-prompting" (i.e., fabricating user engagement logs based on AI use evaluation rubrics) represents a threat to the validity of GenAI interaction assessment. Though the technical expertise required to create convincing fabricated logs may deter such malpractice, addressing this risk may require complementary strategies, such as incorporating student reflections on their AI-assisted process (e.g., Nikolic et al. (2023)). 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.

11 Conclusion

The increasing integration of GenAI in higher education presents both opportunities and challenges for assessing student learning. Our work offers a novel perspective by moving beyond evaluating just the final product or general prompting skills. We propose a conceptual framework (DRIVE) and a practical taxonomy that allows educators to discern evidence of domain-specific learning directly from students' interactions with GenAI, particularly within academic writing contexts. For researchers, this contribution means advancing the understanding of human-GenAI interaction in learning contexts. It shifts the focus from merely observing tool use to identifying how students' evolving prompts and interactive strategies reflect their deepening conceptual understanding. This opens new avenues for studying the intricate cognitive processes involved when GenAI assists (or not) in knowledge construction. For teachers, this work provides a concrete approach to assessing learning in GenAI-compatible classrooms. It offers a way to look beyond concerns of GenAI misuse, instead guiding them to interpret student interactions with GenAI as rich indicators of authentic engagement and mastery, thereby promoting more effective and meaningful human-GenAI educational partnerships.

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*, 34–44. 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

Ansari, A. N., Ahmad, S., & Bhutta, S. M. (2024). Mapping the global evidence around the use of ChatGPT in higher education: A systematic scoping review. *Education and Information Technologies*, *29*(9), 11281–11321. https://doi.org/10.1007/s10639-023-12223-4

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*, 610–623. 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. Series B (Methodological)*, *57*(1), 289–300. https://doi.org/10.2307/2346101

Bower, M., Torrington, J., Lai, J. W. M., Petocz, P., & Alfano, M. (2024). How should we change teaching and assessment in response to increasingly powerful generative Artificial Intelligence? Outcomes of the ChatGPT teacher survey. *Education and Information Technologies*, *29*(12), 15403–15439. https://doi.org/10.1007/s10639-023-12405-0

Britto, R., & Usman, M. (2015). Bloom's taxonomy in software engineering education: A systematic mapping study. *2015 IEEE Frontiers in Education Conference (FIE)*, 1–8. https://doi.org/10.1109/FIE.2015.7344084

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), 100068. https://doi.org/10.1016/j.rmal.2023.100068

Chen, B., Zhang, Z., Langrené, N., & Zhu, S. (2023). Unleashing the potential of prompt engineering in Large Language Models: A comprehensive review. https://doi.org/10.48550/arXiv.2310.14735

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*, 178–188. 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

Clark, E., August, T., Serrano, S., Haduong, N., Gururangan, S., & Smith, N. A. (2021). All that's "human" is not gold: Evaluating human evaluation of generated text. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 7282–7296. https://doi.org/10.18653/v1/2021.acl-long.565

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*, 100209. https://doi.org/10.1016/j.caeai.2024.100209

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

Giray, L. (2023). Prompt engineering with ChatGPT: A guide for academic writers. *Annals of Biomedical Engineering*, *51*(12), 2629–2633. https://doi.org/10.1007/s10439-023-03272-4

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

Heston, T. F., & Khun, C. (2023). Prompt engineering in medical education. *International Medical Education*, *2*(3), 3. https://doi.org/10.3390/ime2030019

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

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

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

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

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)*, 133–138. 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

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

OpenAI. (2022). ChatGPT [Version November 30, Large language model].

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

Porter, B., & Machery, E. (2024). AI-generated poetry is indistinguishable from human-written poetry and is rated more favorably. *Scientific Reports*, *14*(1), 26133. https://doi.org/10.1038/s41598-024-76900-1

R Core Team. (2024). *R: A language and environment for statistical computing* [Version 4.3.3]. R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/

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.

Sawalha, G., Taj, I., & Shoufan, A. (2024). Analyzing student prompts and their effect on ChatGPT's performance. *Cogent Education*, *11*(1), 2397200. https://doi.org/10.1080/2331186X.2024.2397200

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.

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*, 100075. https://doi.org/10.1016/j.caeai.2022.100075

Theophilou, E., Koyutürk, C., Yavari, M., Bursic, S., Donabauer, G., Telari, A., Testa, A., Boiano, R., Hernandez-Leo, D., Ruskov, M., Taibi, D., Gabbiadini, A., &

Ognibene, D. (2023). Learning to prompt in the classroom to understand AI limits: A pilot study. In R. Basili, D. Lembo, C. Limongelli, & A. Orlandini (Eds.), *Aixia 2023 – advances in artificial intelligence* (pp. 481–496). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-47546-7_33

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

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), e0286102. https://doi.org/10.1371/journal.pone.0286102

West, P., Lu, X., Dziri, N., Brahman, F., Li, L., Hwang, J. D., Jiang, L., Fisher, J., Ravichander, A., Chandu, K., Newman, B., Koh, P. W., Ettinger, A., & Choi, Y. (2023, October). The Generative AI Paradox: "What It Can Create, It May Not Understand." https://openreview.net/forum?id=CF8H8MS5P8

White, J., Fu, Q., Hays, S., Sandborn, M., Olea, C., Gilbert, H., Elnashar, A., Spencer-Smith, J., & Schmidt, D. C. (2023). A prompt pattern catalog to enhance prompt engineering with ChatGPT. https://doi.org/10.48550/arXiv.2302.11382

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

Xia, Q., Weng, X., Ouyang, F., Lin, T. J., & Chiu, T. K. F. (2024). A scoping review on how generative artificial intelligence transforms assessment in higher education. *International Journal of Educational Technology in Higher Education*, *21*(1), 40. https://doi.org/10.1186/s41239-024-00468-z

Yan, D. (2023). Impact of ChatGPT on learners in a L2 writing practicum: An exploratory investigation. *Education and Information Technologies*, *28*(11), 13943–13967. https://doi.org/10.1007/s10639-023-11742-4

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

A Appendix

Taxonomy to classify 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.

(continued)

Category	Type	Meaning
	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.