TU/e EINDHOVEN UNIVERSITY OF TECHNOLOGY

4TU.Centre for Engineering Education

TU/e BOOST!

Report 2a

Introducing an assessment framework to evaluate learning through student-GenAI interactions

Manuel Oliveira

Human Technology Interaction Group

Department of Industrial Engineering & Innovation Sciences, TU/e

Project: Scaffolding writing skills using automated essay generation systems

Team: Rianne Conijn, Carlos Zednik, Bert Sadowski, Gunter Bombaerts

Funding: 4TU.CEE & BOOST!

Eindhoven, 22 April, 2025

Contents

Note

The content of this report overlaps with that of a manuscript submitted to a peer-reviewed journal **, that is currently under revise and re-submit status. This work is estimated to include an expanded dataset and will therefore be updated in the future. Some details are omitted to facilitate compliance with confidentiality requirements.

** Oliveira, M.J.B., Zednik, C., Bombaerts, G., Sadowski, B., Conijn, R. (under revision). Assessing learning through students' interactions with generative AI

The need for a new approach to assess learning

The rise of sophisticated Generative AI (GenAI) tools presents a significant challenge to traditional assessment methods in higher education. With user-friendly and freely available AI applications capable of generating high-quality academic text, evaluating student learning based solely on final written outputs is becoming increasingly unreliable. In this GenAI-infused context, teachers will need new approaches to understand the actual skills and knowledge students are developing, especially when GenAI is permitted for academic work.

This report introduces a new taxonomy designed to classify the interactions between students and GenAI chatbots during the process of writing argumentative essays, in a way that provides insight into student learning. By examining student-GenAI interactions, teachers can gain a deeper understanding of how students develop their argumentative writing skills, including their critical thinking, metacognitive engagement, and rhetorical choices. This framework is expected to be particularly valuable in educational settings that allow for the use of GenAI as a learning tool.

Development process

The taxonomy was developed through an iterative process, combining the expertise of teachers and teaching assistants in argumentative writing with observations of real student-GenAI interactions. This involved a process where: first, key learning objective were defined to establish initial taxonomy categories, and second, actual student interaction logs were analyzed to refine and expand the taxonomy content based on observed behaviors through an iterative cycle that combined items generated deductively (learning objectives defined across course iterations) and inductively (based on analyses of student-GenAI interaction logs).

Describing the dialogue between students and Generative AI

The resulting three-tiered taxonomy comprises three main categories:

- **Writing:** Focuses on the mechanical and structural aspects of essay composition, such as providing instructions, specifying evaluation criteria, requesting feedback and improvements on drafts, and seeking assistance with formatting and organization. This category accounted for the largest proportion of interactions (39.7%), highlighting students' reliance on GenAI for writing assistance.
- **Content:** Centers on knowledge construction and understanding, including asking for bibliographic references, examples, definitions, research ideas, elaborations on concepts, and relevant case descriptions. This category represented 30.4% of interactions, indicating a significant use of GenAI for content-related tasks.
- **Argument:** Targets the logical and analytical aspects of writing, such as seeking context for cases, identifying stakeholders, formulating problems, soliciting objections and justifications, improving argumentative structure, and refining the thesis. This was the least prevalent category (22.9%), suggesting less frequent use of GenAI for direct argumentative development.

Applying the taxonomy in TU/e courses

The taxonomy was employed in three writing-intensive courses at TU/e (Technische Universiteit Eindhoven), where students were allowed to use GenAI tools like ChatGPT under the condition that they shared their complete interaction logs. These courses included foundational instruction on argumentative writing and prompt engineering. Details of course, academic year, number of essays, prevalence of AI use, and number of student-GenAI interactions (i.e., input prompt and respective GenAI output) that were annotated using the taxonomy (see **Appendix**) are presented in **Table 1**.

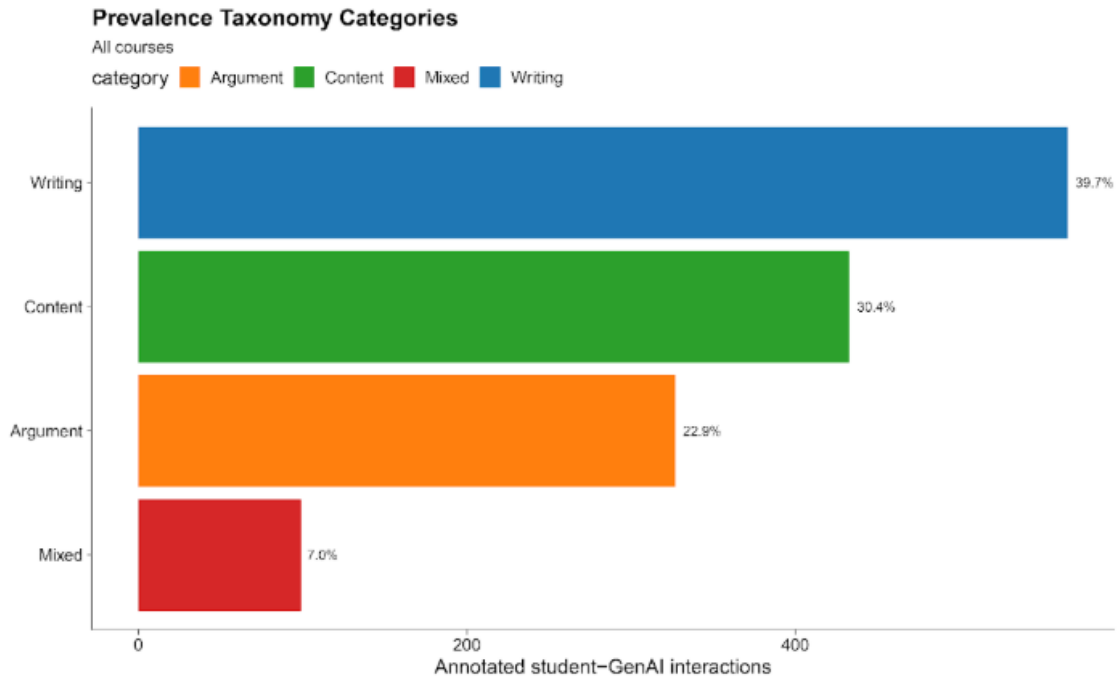
Table 1. *Overall data descriptives.*

Course	Year	Essay Type	Total essays	Used GenAI *n* (%)	Annotated essays *n* (%)	Annotated interactions
Data Science Ethics	2023-2024	Individual	102	34 (25%)	21	377
Rational agents: Robots & AI	2023-2024	Group	20	6 (30%)	6	207
Philosophy & Ethics of AI	2024-2025	Individual	107	54 (33.5%)	33	841
Total			229	94 (41%)	60 (26%)	1425

Prevalence of categories in student-GenAI interactions

Figure 1 show the overall prevalence of categories identified in the annotations of interaction logs shared by students who decided to write their essay with generative AI tools. According to the interaction logs shared by the students, the most used tool was ChatGPT (free version assumed or inferred based on the shared material, as this detail was not a requirement for the submission; less than 1% use other chatbots like Claude or from lesser known platforms who provide access to multiple models, e.g., Blackbox). Annotations were performed by 3 different expert annotators (2 teachers and 1 teaching assistant) and a non-expert research assistant trained by the lead researcher.

Figure 1. Overall descriptives of taxonomy annotations (category only) for all courses.

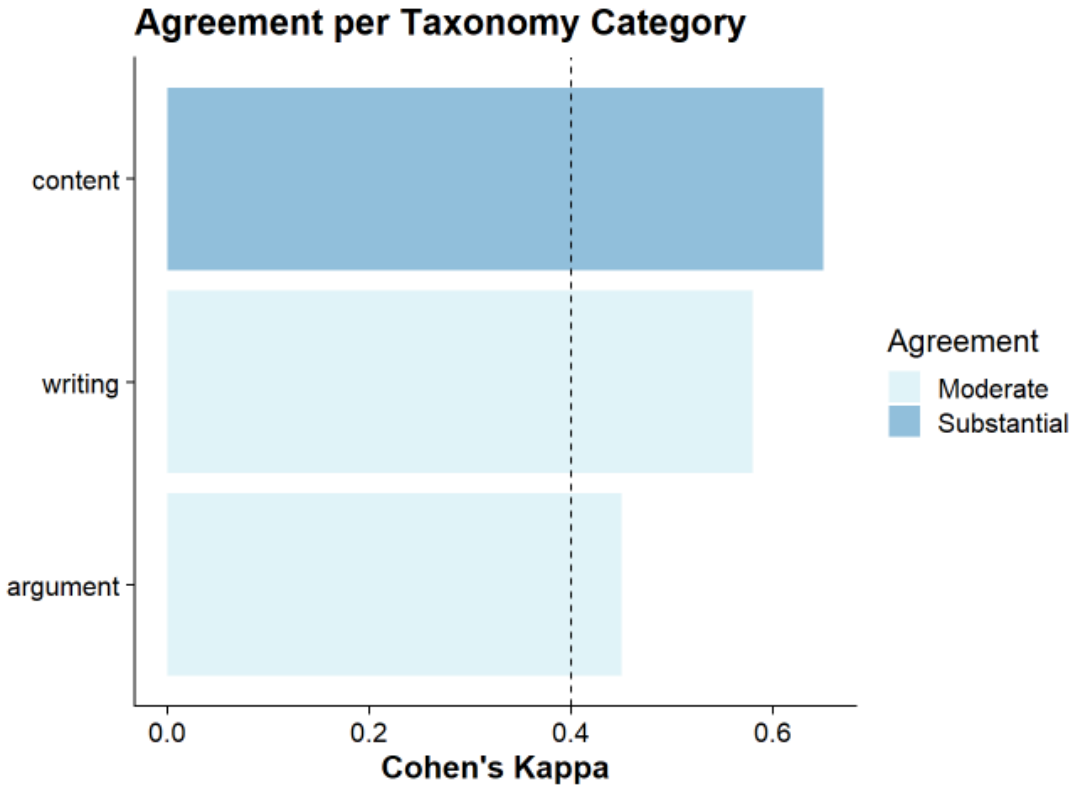


Inter-rater agreement for taxonomy classifications

To gain insight into the consistency of taxonomy classifications across different raters who use the taxonomy to classify student-GenAI interactions (i.e., interactions logs), we computed the inter-rater agreement for annotations that were classified by at least two different raters. These annotations were rated by one of three raters coming from a group of more expert raters (2 course teachers and 1 teaching assistant) and a second rater who rated all the annotations (research assistant trained by the leading researcher on how to use the taxonomy). That is, there was one rater who classified all the interaction logs (n = 779), while the other raters annotated different subsets of the total set of annotations (annotator A = 264, annotator B = 356, annotator C = 159). 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, specifically: research assistant vs. annotator subset A, research assistant

vs. annotator subset B, research assistant vs. annotator subset C. The resulting Kappa values were then averaged, resulting in an overall Cohen's Kappa reflecting the average agreement. The average Kappa was 0.42 (SD = 0.08), 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). Figure 2 illustrates the degree of agreement between annotators (overall, i.e., expert rater group as rater 1 vs research assistant as rater 2) for each taxonomy category.

Figure 2. Inter-rater agreement by taxonomy category (Kappa = 0.40 denotes boundary between Moderate and Fair agreement).



Connection with learning

Analysis of the interaction patterns revealed distinct profiles associated with different levels of student performance, based on both evaluations of their GenAI interactions and traditional essay grades. It is crucial to note that interactions with GenAI in the context of essay writing is fundamentally determined by evaluation criteria defined by teachers who also define the learning objectives in these specific courses. These GenAI interaction criteria are therefore reflecting what an expert (i.e. teacher who designs the learning objectives) subjectively considers to signal quality of learning (see specific GenAI interaction criteria in Table 1).

Table 1. Criteria used to evaluate the logs of students' interactions with GenAI in the context of 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.

Interactions styles and performance on GenAI interaction evaluations

High-performing students tended to use GenAI for higher-order tasks like content ideation (example classification description: user provides well-motivated original idea or question and asks for confirmation/elaboration/discussion) and soliciting counterarguments (example classification description: user asks the machine to provide an objection and/or a response to a given claim.), indicating a collaborative intellectual partnership.

In contrast, lower-performing students (as defined by lower scores on their GenAI interaction logs) often focused on tasks like improving argumentative structure (example: user asks the machine to impose a particular logical structure onto a text) and content research (example: user asks the machine to define an idea, or to identify related ideas to one, given by the user), which suggests a more instrumental approach focused on complying with assignment criteria through content elaboration requests.

To better understand how these results connect with the taxonomy we are proposing please refer to **Table 2**, which reports the taxonomy classifications that were most diagnostic of high and low performance in terms of the perceived quality of GenAI interactions (as rated by the teachers/teacher assistants of the TU/e courses under focus).

Table 2. *Most frequent unique classifications per performance level on GenAI interaction evaluations.*

Taxonomy Classification (see Appendix)	**High Performance** ***n*** **(%)**	**Low Performance** ***n*** **(%)**
Content idea	36 (8.2%)	0 (0%)
Argument Objection	26 (5.9%)	0 (0%)
Writing AutoImprove	26 (5.9%)	0 (0%)
Argument Improve	0 (0%)	41 (8.5%)
Content Research	0 (0%)	39 (8.1%)
Content Elaboration	0 (0%)	28 (5.8%)

It is important to clarify what is meant by high and low performance in this context. A score attributed to an interaction with GenAI by a rater (viz. teacher, teaching assistant) reflects a judgment shaped by the rater's interpretation of the GenAI evaluation rubric (Table 1), the rater's unique expectations about what constitutes evidence of learning and its respective quality, and the experience of the rater analyzing student-GenAI interaction logs. Thus, high or low performance should be best understood as the degree to which a teacher (i.e., rater) judges the interaction to convey evidence of quality of learning. Moreover, this performance is likely to be influenced by unmeasured factors such as a student's degree of GenAI literacy (e.g. prompt engineering, frequency of use) or attitudes developed towards the use of GenAI in light of contextual factors (e.g. clarity of guidelines for AI use, clarity of evaluation criteria for AI use).

Interaction styles and performance on traditional essay evaluations

The most successful academic writing (top scoring essays) demonstrated a collaborative approach with GenAI, where students viewed the technology as an intellectual collaborator. These top-scoring essays integrated AI throughout the writing process, by using it to enhance creative thinking, organize ideas, and provide critical feedback in a dynamic iterative manner. The types of interaction that were exclusive to this group of essays were characterized by: asking the machine to provide an objection and/or a response to a given claim, and providing a well-motivated original idea or question and asking for confirmation/elaboration/discussion.

The least successful essays, based on their lower scores, were characterized by an exploratory approach to GenAI, where students prioritized intellectual discovery and broad knowledge gathering over the structured, targeted writing strategies typically valued in academic evaluations. These lower-scoring papers reflected a more inconsistent engagement with AI, emphasizing conceptual investigation rather than the precise, goal-oriented and structural-focused writing process expected in academic writing assignments. The types of interaction that were exclusive to this group of essays were characterized by: providing a relevant sentence/paragraph and asking the machine to elaborate and provide additional detail, mentioning specific course-related content, or prompting the system in non-specific technical ways (or actions that are too diverse, or not yet captured by the current taxonomy).

To better understand how these results connect with the taxonomy we are proposing please refer to **Table 3**, which reports the taxonomy classifications that were most diagnostic of high and low performance in terms of essay grades. Note that the performance level reported in Table 3 refers to the traditional evaluation of essays, even if these were co-written with GenAI. In this sense, the performance level can be understood as a proxy indicator for the impact of using GenAI on the quality of an essay as it is traditionally assessed (in these TU/e courses).

Table 3. M*ost frequent unique classifications per performance level on traditional essay assessment*.

Taxonomy Classification (see Appendix)	**High Performance** ***n*** **(%)**	**Low Performance** ***n*** **(%)**
Content idea	35 (8.2%)	0 (0%)
Writing AutoImprove	33 (7.7%)	0 (0%)
Argument Objection	22 (5.1%)	0 (0%)
Content Bibliography	0 (0%)	32 (6.5%)
Writing Miscellaneous	0 (0%)	43 (8.7%)
Content Elaboration	0 (0%)	32 (6.5%)

Implications for teaching

This taxonomy equips educators with a robust assessment framework designed to withstand the disruptive impacts of generative AI on the evaluation of argumentative writing. When course learning objectives explicitly incorporate the permissible use of generative AI tools, the taxonomy facilitates deeper insights into students' learning processes by analyzing their interactions with AI during writing tasks. Specifically, the taxonomy enables educators to:

- **Identify patterns of GenAI use:** Understand which aspects of writing and research students are using GenAI for most frequently.
- **Assess evidence of learning**: Differentiate between superficial reliance on AI and substantive, skill-building engagement. The interpretation of this distinction is dependent on clearly defined learning objectives. For example, are skills going to be assessed in a setting where AI is allowed? Is the assessment focused on the quality of student-AI collaboration? These define different learning objectives that determine what constitutes evidence of learning.
- **Provide targeted feedback:** Offer specific guidance on how students can more effectively utilize GenAI for learning and skill development.
- **Inform curriculum design:** Adapt learning objectives, teaching strategies and assignments based on observed student-AI interaction patterns.

Current challenges

Despite the valuable insights yielded by this study, it is important to acknowledge inherent limitations, such as the rapidly evolving characteristics of GenAI and the specific contextual parameters of the collected data. Future research endeavors should prioritize the longitudinal tracking of shifts in student interaction strategies, the expansion of the dataset across a more diverse array of academic disciplines, and the systematic investigation of the influence exerted by students' AI literacy levels on the observed interaction patterns.

Next Steps

Developing practical guides and workshops for educators on how to effectively implement this taxonomy in their teaching practices to further enhance its impact on assessment in GenAI-integrated learning environments.

Disclaimer on AI assistance

The present report was co-written with generative AI assistance. The models used included ChatGPT-4o, Gemini 2.0, and Claude 3.5 Haiku. The AI was used for the following tasks: rephrasing text for increased readability and conciseness. The report sections were initially outlined and iteratively refined through a brainstorming process with AI. All AI suggestions and output were monitored and revised by me (the main author). As the main author, I am therefore, responsible for the content of this report.

References

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

Appendix

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).
	Auto Improve	User asks the machine to improve a draft (against previous feedback from a teacher)
	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.