Report 2b

A taxonomy for assessing learning in argumentative essays co-written with Generative AI: Users' guide

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 partially 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

A taxonomy to assess learning despite GenAI use

This report introduces a novel taxonomy designed to analyze the interactions between students and Generative AI (GenAI) chatbots, such as ChatGPT or similar applications, during the process of writing argumentative essays, where they are expected to critically engage with topics that are societally relevant. The construction of this taxonomy framework aims to provide university teachers with a method to assess student engagement and learning in classroom contexts where GenAI tools are permitted for writing graded assignments.

Our taxonomy is structured into three primary categories, encompassing a total of 36 subcategories (see full taxonomy in **Appendix**) that capture both the nuances of student-AI interactions, as well as the types of interactions that more specifically relate with the goal of writing argumentative essays:

1. Writing: This category includes 14 subcategories and focuses on interactions related to the mechanical and structural aspects of essay composition. Subcategories describe interactions such as:

- Providing instructions to the AI
- Specifying evaluation criteria
- Asking for evaluations of drafts
- Requesting automated improvements (grammar, spelling, style)
- Seeking assistance with formatting, introductions, and conclusions

- Asking for text completion or summarization

2. Content: This category branches into 10 subcategories, and centers around knowledge construction and understanding. Examples of these subcategories include:

- Asking for bibliographic references
- Seeking examples or definitions
- Requesting related research ideas
- Asking for elaborations of course-specific concepts
- Seeking descriptions of relevant cases
- Asking for well-motivated original ideas

3. Argument: This category encompasses 12 subcategories and targets the logical and analytical aspects of argumentative writing. Examples include:

- Seeking context for real-world cases
- Researching case details
- Identifying stakeholders and their values
- Formulating moral problems
- Soliciting objections to claims
- Asking for justifications for claims
- Seeking help with logical and argumentative structure
- Refining the thesis statement

Taxonomy as an assessment tool

This taxonomy can be employed as an assessment tool in educational contexts where GenAI is permitted for co-writing argumentative essays. To implement this approach, the following conditions are crucial:

- **Allowed use of GenAI tools:** Students should be explicitly allowed to use GenAI tools as part of their essay writing process.
- **Mandatory submission of interaction logs:** Students must be required to submit their complete interaction logs with the GenAI chatbot alongside their final essay assignments. This provides the target data for analysis using the taxonomy.
- **Well-defined evaluation criteria for the interaction logs:** A well-defined rubric is essential for evaluating the submitted GenAI interaction logs. This rubric should

outline the criteria for assessing the quality and nature of the student's engagement with the AI, potentially focusing on aspects such as:

- **Strategic use of GenAI:** How effectively does the student leverage GenAI for different aspects of the writing process (writing, content, argument)?
- **Critical engagement:** Does the student critically evaluate the AI's output, refine prompts, and demonstrate independent thinking?
- **Depth of inquiry:** Do the interactions indicate a surface-level use of AI or a deeper engagement with the subject matter and argumentation?
- **Alignment with Learning Objectives:** Do the interactions demonstrate alignment with course learning objectives? Example: if the goal is to learn how to critically evaluate AI output, is the interaction showing evidence of that (evidence as defined by teacher or matching a relevant taxonomy item)?

How does the taxonomy connect with evidence of learning?

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 focus on philosophical argumentation and ethics of human technology interaction, with an emphasis on AI-related topics. A study was conducted to investigate the feasibility of using the taxonomy to assess evidence of learning through the analysis of the interaction logs themselves. In our investigation, we focused on cases where a student used GenAI to co-write their essay and who successfully submitted both the essay and the interaction log. Each user prompt in these interaction logs was classified using the taxonomy (see **Appendix**) by at least two different raters (an expert teacher/teaching assistant and a trained research assistant), with moderate agreement overall (Cohen's Kappa = 0.42, SD = 0.08) (Landis & Koch, 1977). Subsequently, we computed the most frequent taxonomy classifications for two grops of students defined by the cluster of highest scores on either GenAI interaction quality (as subjectively assessed by an expert teacher/teaching assistant), or traditional essay grade (as assessed using traditional essay quality criteria in these TU/e courses). In this sense, a distinction should be made between assessing the final product (the essay) and assessing the student's interaction with GenAI (the prompts). **Tables 1** and **2** report the most diagnostic taxonomy items for high and low performance on GenAI interaction evaluation (**Table 1**) and traditional essay grade (**Table 2**).

More details can be found in a previous project report (Oliveira, 2025) or the article under preparation (Oliveira et al., 2025, in prep).

What is a good interaction with GenAI when co-writing essays?

Table 1. *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 Auto Improve	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%)

In regard to the perceived quality of GenAI interactions, teachers and teaching assistants perceived higher-quality interactions between students and GenAI when students engaged in a collaborative intellectual partnership, focusing on tasks such as content ideation (providing original ideas and seeking confirmation, elaboration, or discussion) and soliciting counterarguments (asking for objections and responses to given claims). These interactions aligned more closely with the course learning objectives. On the other hand, interactions perceived as lower in quality by educators often involved students taking a more instrumental approach, emphasizing tasks like improving argumentative structure (requesting the imposition of specific logical structures) and content research (seeking definitions or related ideas), suggesting a focus on complying with assignment criteria through content elaboration requests, which may not fully address the intended learning outcomes (in the eyes of teachers/teaching assistants in these TU/e courses).

In practical terms, these are the specific interactions with GenAI that we found to be associated with the best outcomes in perceived quality of GenAI interaction (extracted from taxonomy in **Appendix**):

- **Content Idea:** User provides well-motivated original idea or question and asks for confirmation/elaboration/discussion.
- **Argument Objection:** User asks the machine to provide an objection and/or a response to a given claim.
- **Writing Auto Improve:** User asks the machine to improve a draft (against previous feedback from a teacher)

For teachers of courses that emphasize philosophical thinking and writing, such as those from which the current results are derived, directly applying this taxonomy to assess learning through student-GenAI interaction logs may be feasible. This approach can provide insights beyond the mere assessment of essays alone, provided the educators are interested in evaluating student learning through these interactions in addition to traditional essay grading.

Teachers must consider the learning outcomes they wish to achieve and how student-GenAI interactions align with those goals. If certain types of interactions, such as content ideation and soliciting counterarguments, are deemed valuable, they should be explicitly defined as learning objectives and supported by learning activities that guide students to interact with

GenAI in these desirable ways. On the other hand, if the primary focus is on developing writing skills independently of GenAI, teachers should communicate clear expectations about how GenAI should not be used. When teachers grasp the consequences of the various ways through which students interact with GenAI, they can more consciously determine how to incorporate GenAI into their courses and adjust their instructional approaches as needed.

Which interactions with GenAI are related with good essay grades?

Table 2. 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 Auto Improve	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%)

High-performing essays demonstrated a collaborative approach with GenAI, integrating AI throughout the writing process to enhance creative thinking, organize ideas, and receive critical feedback. Interactions exclusive to this group included asking for objections to claims and providing original ideas for elaboration. On the other hand, lower-performing essays prioritized intellectual discovery and broad knowledge gathering over structured writing strategies, showing inconsistent AI engagement. These essays focused on conceptual investigation rather than the precise, goal-oriented writing expected in academic assignments, with interactions such as asking for elaboration on relevant content or prompting the system in non-specific technical ways.

In practical terms, the specific interactions with GenAI that were found to be associated with the best outcomes in perceived quality of essays are the same as the ones associated with perceived quality of interaction with GenAI, namely:

- **Content Idea:** User provides well-motivated original idea or question and asks for confirmation/elaboration/discussion.
- **Writing Auto Improve:** User asks the machine to improve a draft (against previous feedback from a teacher)
- **Argument Objection:** User asks the machine to provide an objection and/or a response to a given claim.

The current results also provide initial insights into how the use of Generative AI tools can impact the traditional assessment of essays. Notably, the findings suggest that certain specific ways of interacting with GenAI, such as engaging in a collaborative and iterative writing process, are associated with higher grades. Students who view GenAI as an intellectual partner and leverage its capabilities to enhance creative thinking, organize ideas, and receive critical feedback tend to produce essays that are perceived as high-quality by readers (in this case teachers).

While these results may not be universally applicable to all contexts and should be interpreted with caution, they can serve as a guide for students looking to use GenAI more effectively in their essay writing. When students understand how certain ways of using GenAI lead to better results (e.g., through rubrics informed by the current insights), they can tweak their approach to get the most out of the technology in circumstances where it is allowed. This could mean deliberately using GenAI to brainstorm ideas together, letting it help organize and polish their arguments, and repeatedly weaving its suggestions into their work to end up with a stronger final product. Again, teachers should carefully assess where to place the boundaries for GenAI use, in alignment with the learning objectives they set, and make sure these guidelines are very clearly understood by students.

Feasibility of assessing learning through student-GenAI interactions

The overlap in the types of interactions with GenAI for high performance in the two different metrics (essay grade vs. interaction evaluation), strongly suggests that essays rated as high-quality are positively related to the perceived quality of interactions with GenAI (through analysis of interaction logs using the proposed taxonomy). This can also be understood as evidence that the analysis of student-GenAI interactions can provide a window into the same types of learning indicators used in traditional assessments.

Practical uses for teachers

By utilizing this taxonomy and the associated interaction logs, teachers can:

- **Analyze patterns of GenAI use:** Identify which categories and subcategories of interactions are most prevalent among students.
- **Gain insights into student processes:** Understand how students are approaching argumentative writing with the assistance of GenAI, revealing their strengths and weaknesses along the timeline of interactions. ***Attention***: Please note that many cognitive processes and decisions cannot yet be captured by this method, and care must be taken when inferring the absence of skills, as these may simply be unobservable through reported interactions alone (e.g., no changes to AI output does not necessarily imply absence of critical assessment of AU output).
- **Inform personalized feedback:** Provide targeted feedback to students based on their interaction patterns, guiding them on how to use GenAI more effectively for learning and skill development. In this sense, the evaluations scores should strive to be aligned with the learning activities promoting effective uses of GenAI before assessment, and make sure to fairly assess GenAI users and non-users.

How can these results promote constructive alignment in a writing-intensive course?

The current insights on the connection between use of GenAI for co-writing and learning performance indicators can inform the (re-)design of learning objectives that better align with the reality of GenAI's presence in the writing process. As a first step towards designing courses that are more resistant to the disruptive impact of GenAI, teachers must carefully consider the learning outcomes they wish to achieve while being aware of how GenAI influences the outcomes they assess. As mentioned above, teachers should consider how their learning objectives align (or not) with the use of GenAI. Some potentially useful questions you may want to ask yourself as a teacher at this stage:

- Can the student evaluate the output of AI if they never developed the skill required to produce that same type of output in an AI-free environment?
- Is the goal to learn how to interact with AI?

If the primary goal is to develop strong academic writing skills, teachers should focus on guiding students to use GenAI collaboratively, integrating it throughout the writing process to aid with creative thinking, organize ideas, and receive critical feedback. Through activities that encourage interactions such as asking for objections to claims and providing original ideas for elaboration, teachers can help students leverage GenAI effectively to produce high-quality essays.

However, if the learning objectives prioritize the development of research skills, critical thinking, or conceptual understanding, teachers may need to adapt their assessment criteria to account for the exploratory nature of some student-GenAI interactions. In such cases, assessing the student's interaction with GenAI, rather than exclusively focusing on the final essay, can provide valuable insights into their learning process and growth. This would rflect an assessment approach that is more focused on the learning process, which some authors advocate as a more promising focus in the age of AI (Swiecki et al., 2022).

Ultimately, by understanding the impact of GenAI on student writing and the different types of interactions that lead to successful outcomes, teachers can make informed decisions about aligning their learning objectives, instructional strategies, and assessment methods with the reality of GenAI-assisted writing.

Concluding remarks

In summary, this taxonomy offers a structured approach to assessing evidence of learning through the examination of student-GenAI interactions, in the specific context of argumentative essay writing. When implemented with clear guidelines for GenAI use and a robust evaluation rubric for interaction logs, this taxonomy can serve as a valuable tool for assessing the outcome of student learning processes in the dynamic landscape of AI advancements, by providing an assessment approach that is relatively compatible with an increasingly GenAI-infused higher education environment.

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

Oliveira, M. J. B. (2025). *Report 2a: Introducing an assessment framework to evaluate learning through student-GenAI interactions* (pp. 1–13). Eindhoven University of Technology.

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

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.