



What student prompts reveal about their learning: Introducing the DRIVE framework

MANUEL OLIVEIRA

Webinar 4TU.CEE, 18th November 2025

Department Industrial Engineering & Innovation Sciences (IE&IS)
Human Technology Interaction

4TU. CENTRE FOR
ENGINEERING EDUCATION

10
YEARS

TU/e BOOST!

TU/e
EINDHOVEN
UNIVERSITY OF
TECHNOLOGY




Contents lists available at ScienceDirect

Computers and Education: Artificial Intelligence

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



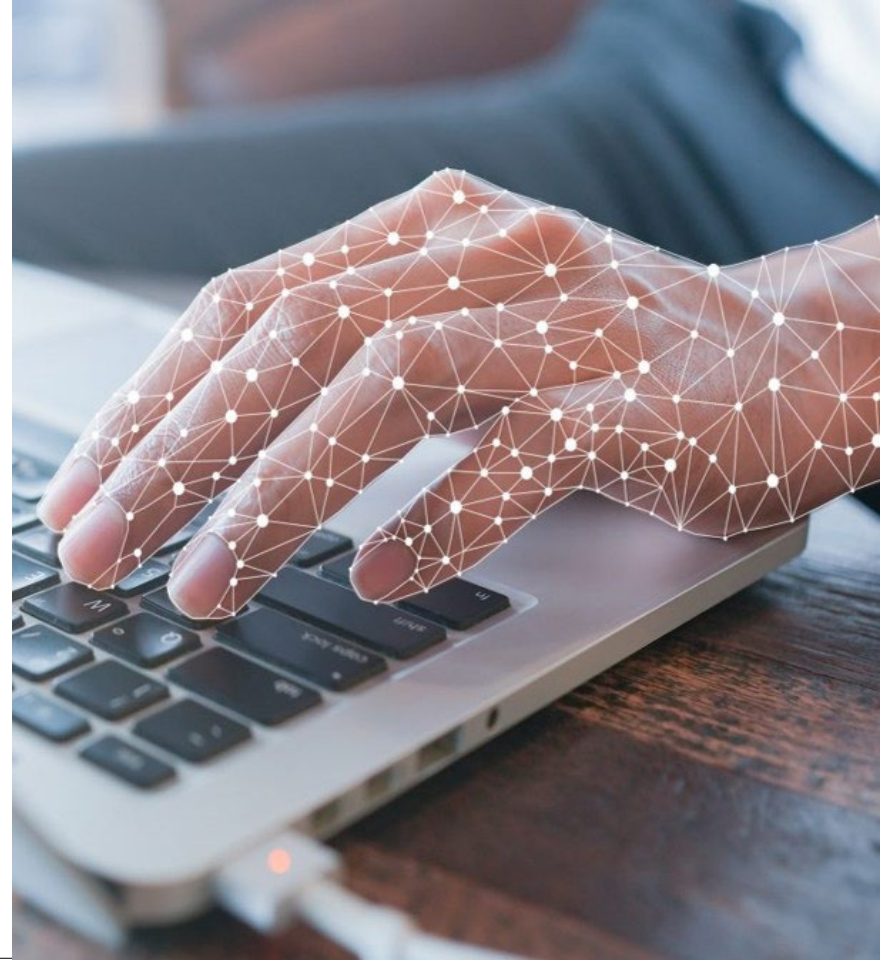
Assessing students' DRIVE: A framework to evaluate learning through interactions with generative AI

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

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

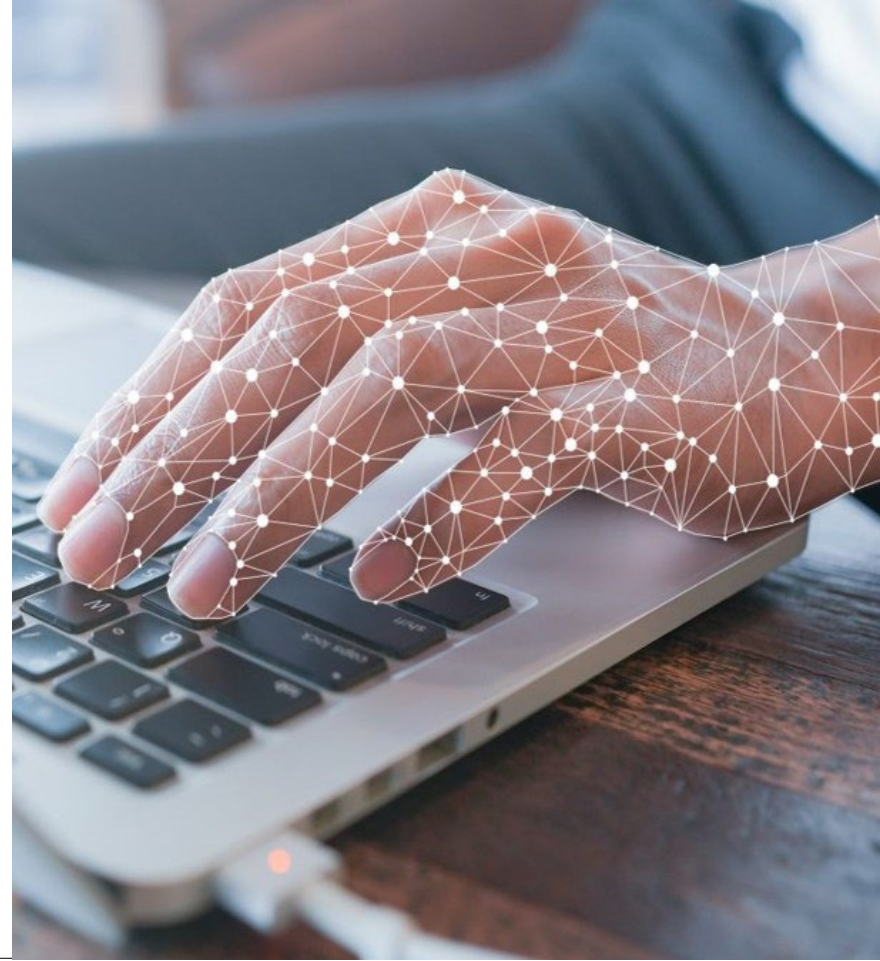
This webinar is NOT about

- Technical prompting (prompt engineering) or AI literacy
- AI adoption ethics
- AI detection (cheating)

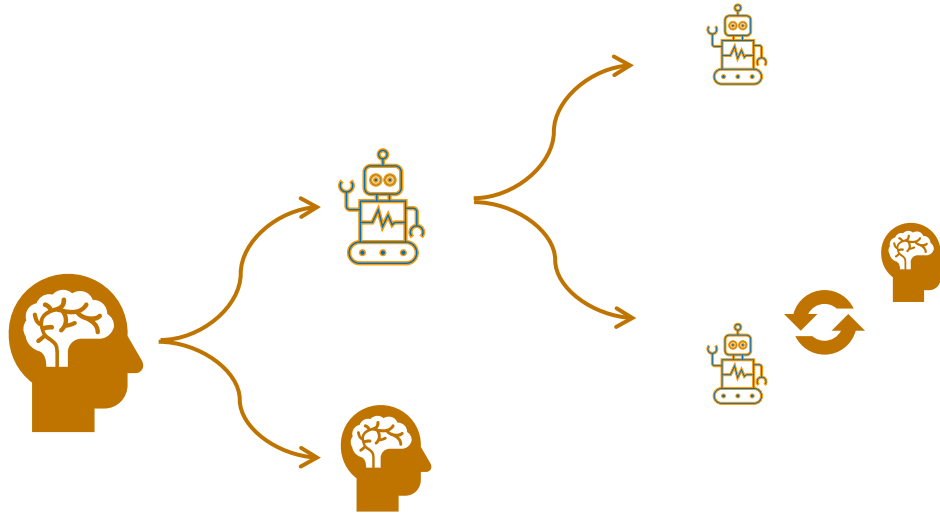


What this webinar is about:

- Assessment in classrooms where GenAI is allowed for graded coursework
- identifying learning **while** using GenAI
- domain-specific learning (course LOs)



Who did what?



Output
=
Learning?



One solution: look at the process



What does the Prompter know?

Activity time!

Feel free to speak up! 😊

MO Act as an expert travel agent. Create a 7-day itinerary for a first-time trip to the Netherlands.

Focus: Must-see tourist spots and culture.

Cities: Include Amsterdam and a day trip for tulips and windmills.

Format: A day-by-day table.

Tone: Enthusiastic and helpful.



Your Ultimate 7-Day Netherlands Adventure! NL

I'm thrilled to help you plan your first trip to the Netherlands! This itinerary combines the vibrant energy of Amsterdam with iconic Dutch experiences. You're in for an absolute treat!

What does the Prompter know?

Activity time!

Feel free to speak up! 😊

MO hey, im planning a 7 day trip to the netherlands and need help with an itenerary. im focused on history and art, so no tulips. My must-see locations are Delft: Vermeer Centre and the New Church. And also Utrecht: The Dom Tower and that Reitveld Schroder House. Amsterdam for the rest of the time. Can you help me build a logical day-to-day itenerary that connects these sights efficiently? also one quick logistics question: for transport in Utrecht, I know Ill have an OV chipkart for trains, but will that work on the trams too? or is it just easier to get one of those OV-fiets at the station?

I'll help you build an efficient itinerary! Given your focus on history and art, and that you're based in the Netherlands, this should work well logistically.

7-Day Itinerary

Where is *learning* in student-AI interactions?



Knowledge and question asking

Rafael Ibáñez Molinero and Juan Antonio García-Madruga
Universidad Nacional de Educación a Distancia

JOURNAL OF VERBAL LEARNING AND VERBAL BEHAVIOR 18, 357–364 (1979)

To Ask a Question, One Must Know Enough to Know What is Not Known

NAOMI MIYAKE AND DONALD A. NORMAN

University of California, San Diego

In this study, we test the notion that a prerequisite for asking questions about new topic matter is some appropriate level of knowledge. Learners should ask the most questions when their knowledge is well matched to the level of presentation. To test this hypothesis, we tested learners with two levels of background knowledge using learning material with two levels of difficulty. The

Questions and Question Asking in Verbal Discourse: A Cross-Disciplinary Review

Greg P. Kearsley^{1,2}

Towards a Theory of Question Asking¹

August Flammer

Department of Psychology, University of Fribourg, 14, Rue St-Mi

Where is *learning* in student-AI interactions?

Self-determined learning (heutatology) (Hase & Kenyon, 2007)

- Learner owns their learning path
- AI acts as mere scaffold



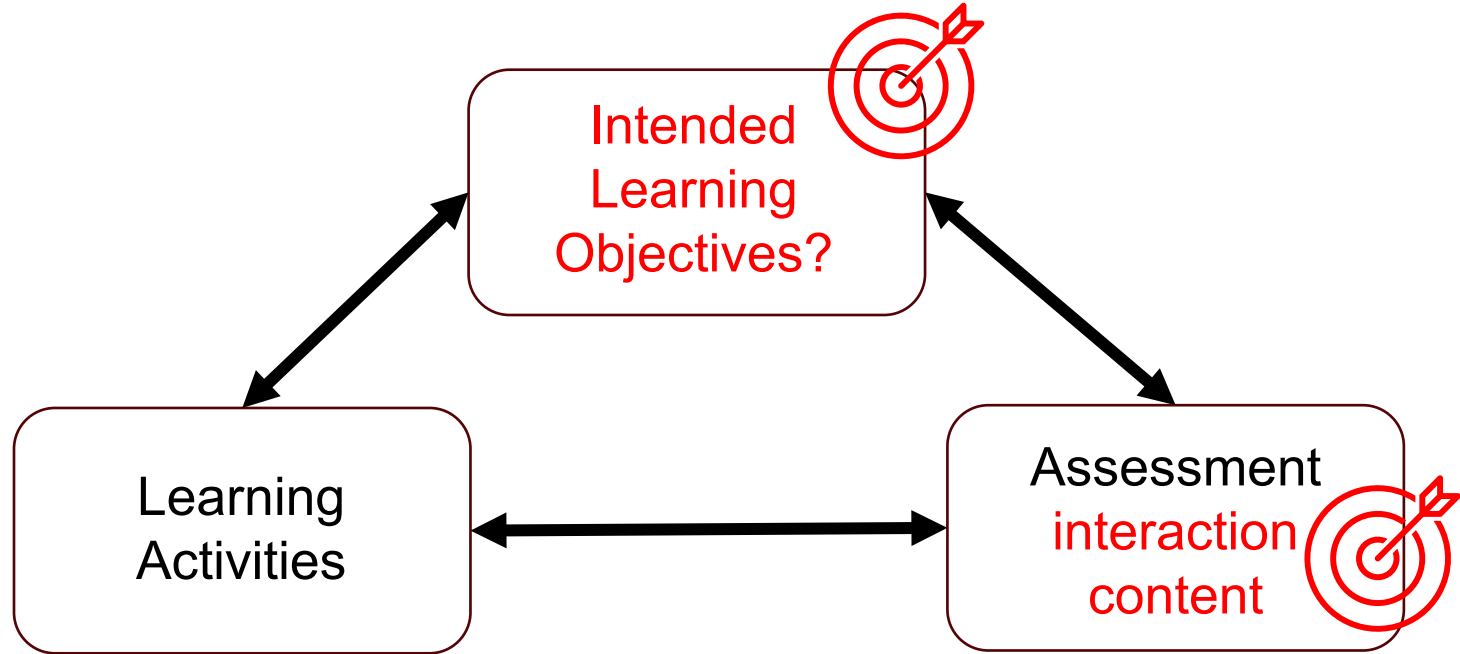
Cognitive engagement & behavior (Chi & Wylie, 2024; Yang et al., 2025)

- Acceptance or simple copying AI output -> **shallow learning**
- Critical feedback, iteration, and steering -> **deeper learning**

Examining process allows assessing:

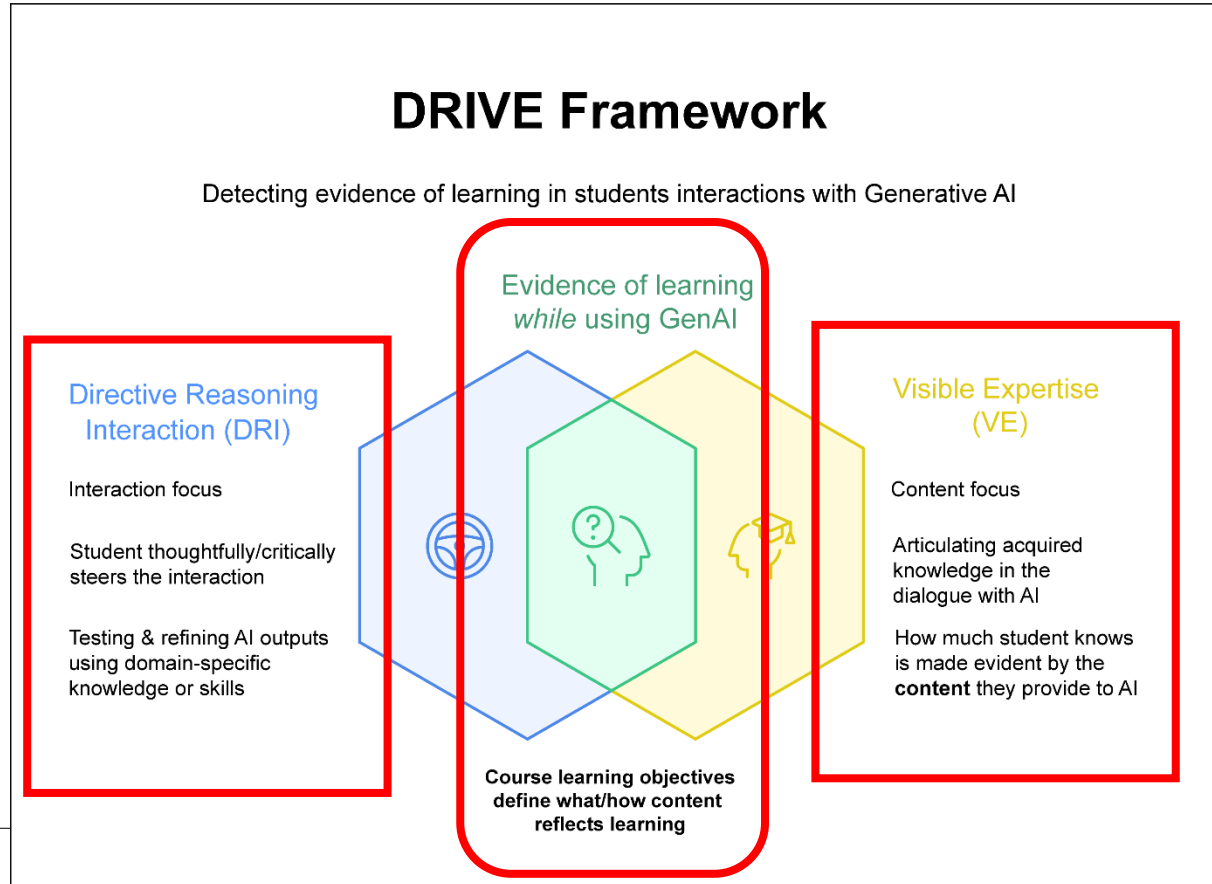
- Degree of cognitive offloading / critical thinking (see Gerlich, 2025)

Where is *learning* in student-AI interactions?



Two criteria to evaluate quality of interaction: DRI & VE

Full paper



Research methodology



CONTEXT

- 2 TU/e courses (3 groups)
- 2023-2025
- N = 445
- AI allowed & graded (23% adoption)

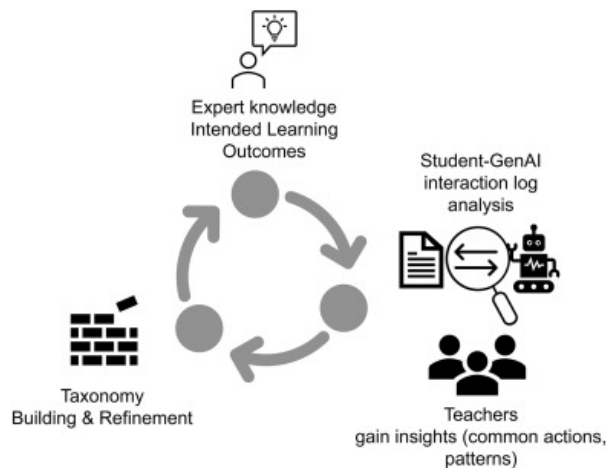
LEARNING OBJECTIVES

- Argumentation
- Critical thinking
- Creativity/originality

ASSESSMENT

- Argumentative essay
- No AI? Only essay graded
- AI? Essay + Interaction graded

To help annotate and evaluate prompts, teachers applied a **rubric** and developed course-specific **prompt taxonomy**.



Appendix A. Taxonomy to evaluate student-GenAI interactions

Category	Type	Meaning
Writing	Instructions	User specifies the task, in terms of the course's assignment description (e.g. copy-paste or upload)
	Criteria	User specifies the task in more detail, by providing the evaluation criteria for the assignment, from the assignment rubric (usually, copy-paste)
	Evaluate	User asks the machine to evaluate a draft against the provided criteria (or without criteria).
	Improve	User provides a phrase, paragraph, or essay to be improved by the machine for e.g. spelling, style or grammar.
	Format	User asks for improved formatting (including e.g. bibliographical formatting)
	Organization	User asks for feedback or improvement of essay structure.
	Introduction	User asks the machine to provide an effective introduction.

Taxonomy categories:

- **Writing** (13)
- **Argument** (10)
- **Content** (12)

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.

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

To help annotate and evaluate prompts, teachers applied a **rubric** and developed course-specific **prompt taxonomy**.

Interrater agreement

Moderate

Cohen's Kappa = **0.44**
(*SD*=0.06)

Appendix A. Taxonomy to evaluate student-GenAI interactions

Category	Type	Meaning
Writing	Instructions	User specifies the task, in terms of the course's assignment description (e.g. copy-paste or upload)
	Criteria	User specifies the task in more detail, by providing the evaluation criteria for the assignment, from the assignment rubric (usually, copy-paste)
	Evaluate	User asks the machine to evaluate a draft against the provided criteria (or without criteria).
	Improve	User provides a phrase, paragraph, or essay to be improved by the machine for e.g. spelling, style or grammar.
	Format	User asks for improved formatting (including e.g. bibliographical formatting)
	Organization	User asks for feedback or improvement of essay structure.
	Introduction	User asks the machine to provide an effective introduction.

Taxonomy categories:

- **Writing** (13)
- **Argument** (10)
- **Content** (12)

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.

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

Essay rubric

Criterion	Excellent	Insufficient
Essay introduction and motivation	Characterization of topic/case is unusually insightful, clear, well-formulated, and well-motivated; <u>demonstrates</u> excellent understanding of context and extensive research.	Characterization of topic/case is unclear, implausible, or absent; shows weak understanding of context or is inadequately supported by sources.
Thesis statement	<u>Thesis</u> statement is highly <u>original / creative</u> , clear, plausible, and well-elaborated.	<u>Thesis</u> statement is unclear, highly implausible, absent, or elaborated in too little depth.
Argument in support of thesis	Argument is highly <u>original / creative</u> , clear, and well-formulated; provides very strong support for position; all crucial parts of argument are well-supported (with sources where appropriate, evidence, and excellent use of relevant course concepts/theories).	Argument is unclear, <u>does not</u> provide support for position, or is missing. Sources, evidence, or use of relevant course concepts/theories to support argument are inadequate or <u>demonstrate significant misunderstanding / error</u> .
(At least one) objection and response	Objection is important, clearly stated, and very well developed. Response to objection is <u>original / creative</u> , well-developed, strong, and well-supported.	Missing, weak, off-topic, or inadequately developed objection; or missing, weak, off-topic, or poorly supported response.
Clarity and organization	Publishable quality.	Sentences and paragraphs <u>hard</u> to comprehend due to imprecision, grammatical or spelling errors, etc. Inappropriate or insufficient structure; <u>hard</u> to follow argument; <u>relationship</u> between some sentences or paragraphs are not obvious.

Overall
Essay
score

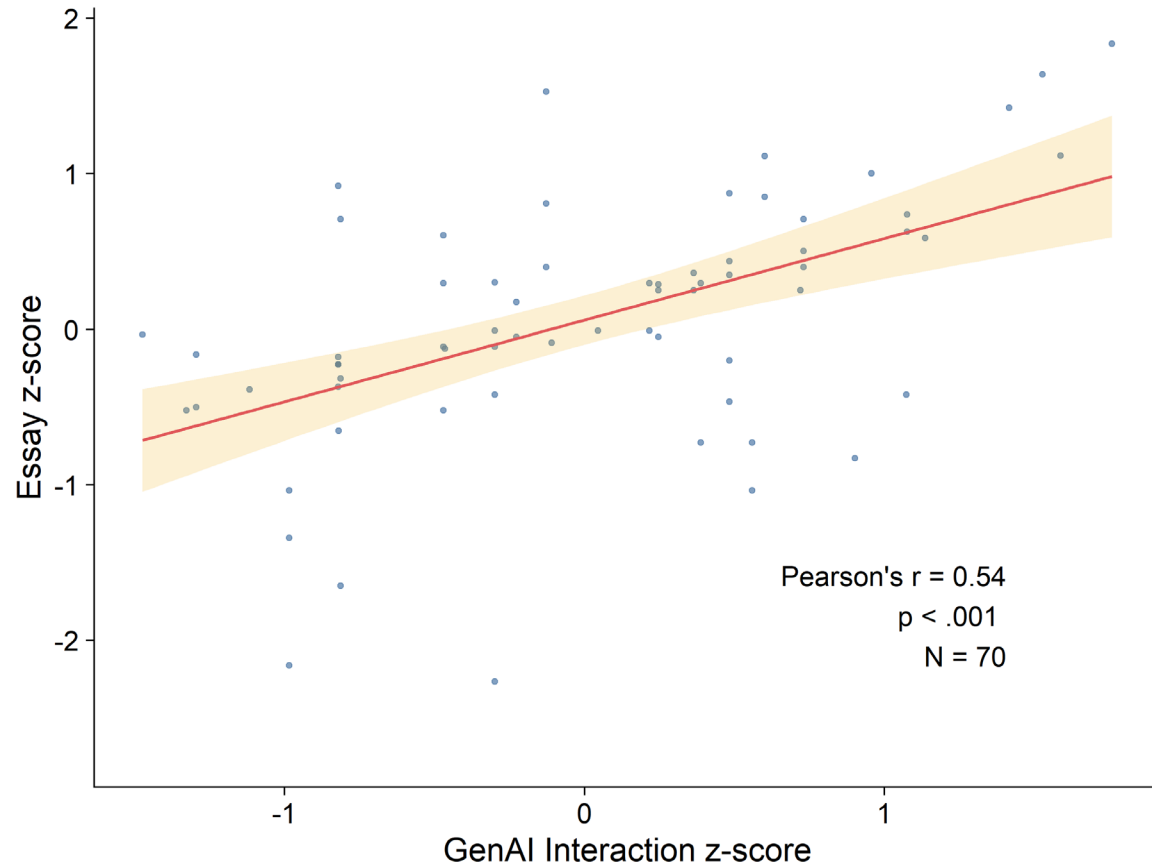
AI interaction rubric

Criterion	Excellent (10-9)	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.	No prompts provided, or prompts <u>unclearly</u> 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.	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 <u>deep</u> and broad understanding of, and engagement with, the course material, at times going beyond that material.	Prompts used <u>insufficiently</u> for content-related research. Prompts reveal no meaningful understanding of, or engagement with, the course material.

DRI
&
VE

Overall
AI interaction
score

Prompt-grading is
no worse a measure
than essay grading



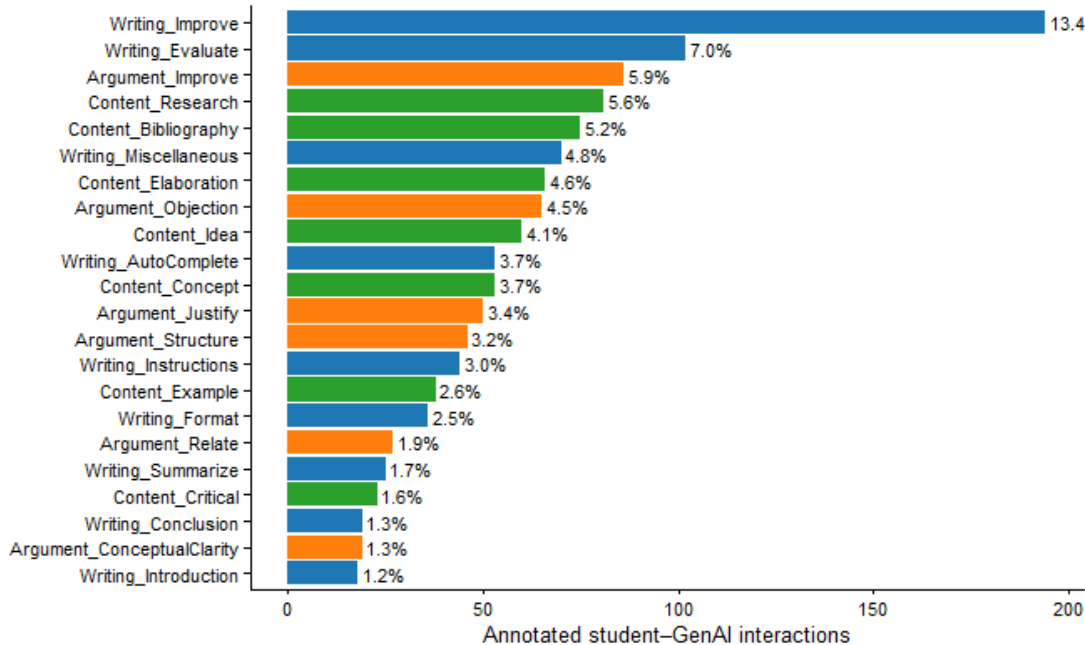


Mapping GenAI interactions

- 1450 taxonomy annotations

Prevalence of Taxonomy Classifications

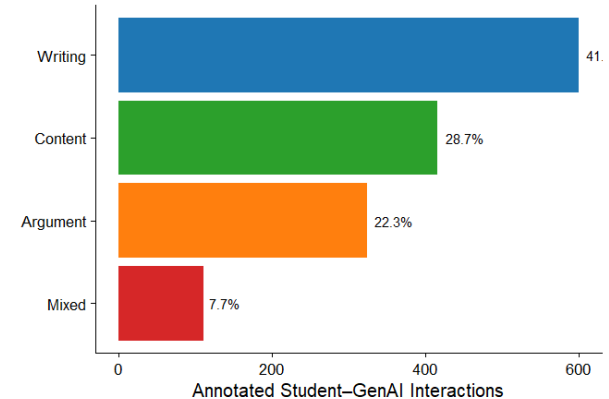
All courses



Taxonomy Category Membership: ■ Argument ■ Content ■ Writing

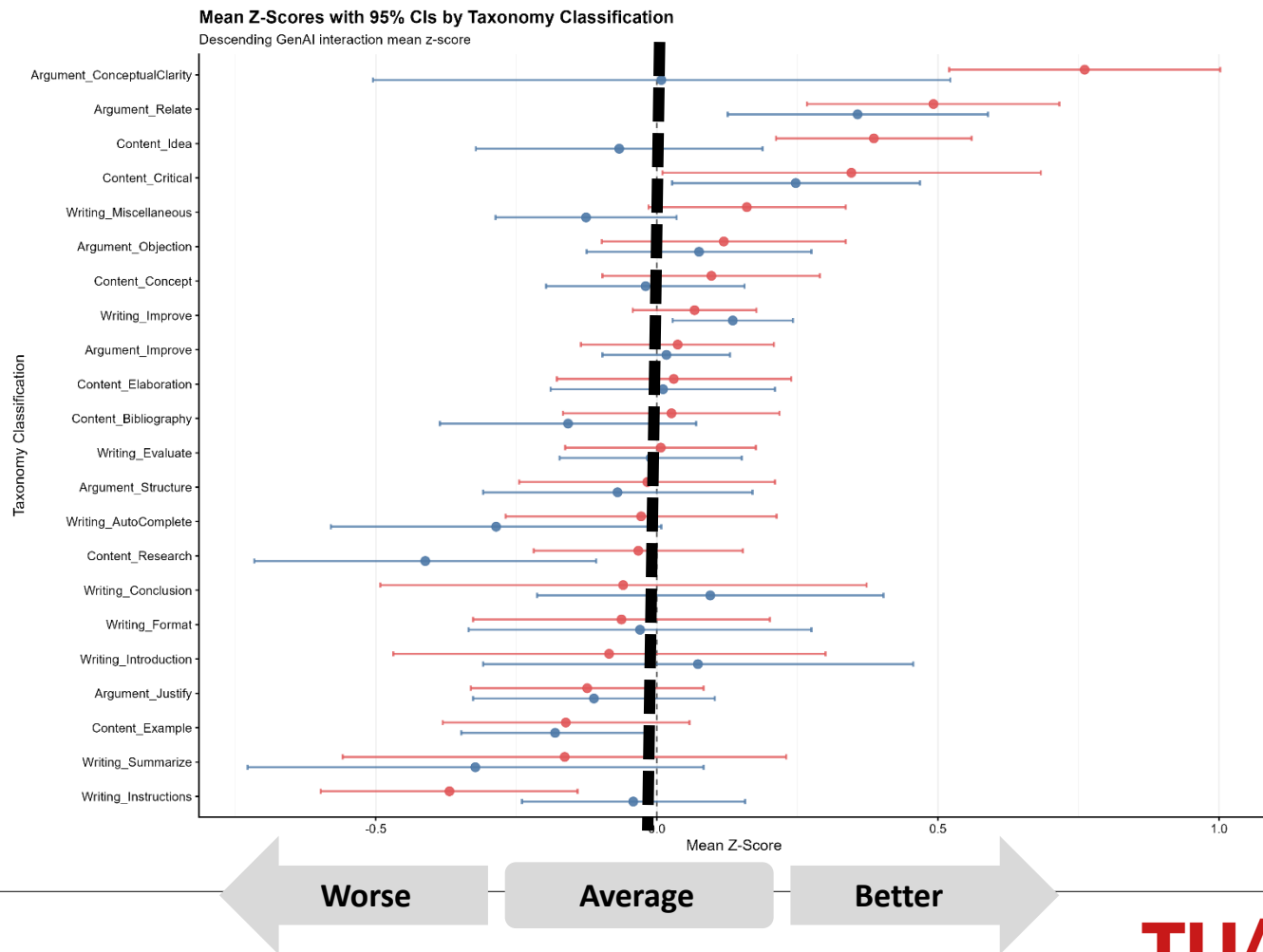
Note: only frequencies > 1%

Prevalence of Taxonomy Main Categories

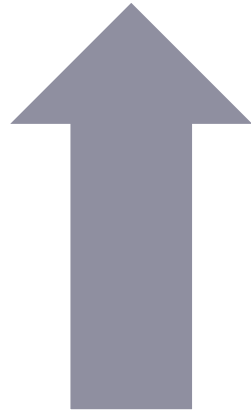


GenAI interactions & performance

● Essay Z-Score
● GenAI Interaction Z-Score

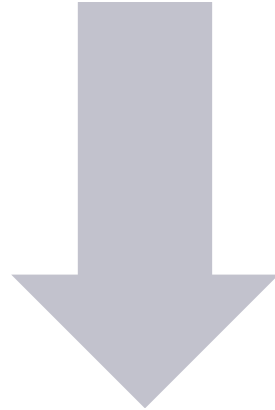


Performance profiles



High performance

- **Collaborative intellectual partnership** (interaction)
- **Targeted improvement partnership** (essay)



Low performance

- **Passive task delegation** (interaction)
- **Basic information retrieval** (essay)

Performance profiles: GenAI interaction score

High GenAI
Interaction Argument
Conceptual
Clarity

Low GenAI Writing
Interaction Instructions

VS.

"Let's revise premise 4.
Here, we still talk about
meaning ... it really looks
like the octopus thought
experiment suggested by
Bender & Koller ... Give me
3 suggestions on how we
can fix this premise ... "

"[file_uploaded] Write an
essay which answers the
essay question by stating a
clear thesis ... "

Performance profiles: Essay score

High Essay Writing
Improve

" ... please leave out anything about language models. that will come in a later paragraph. for this paragraph, really focus on why subjective experience is needed ... "

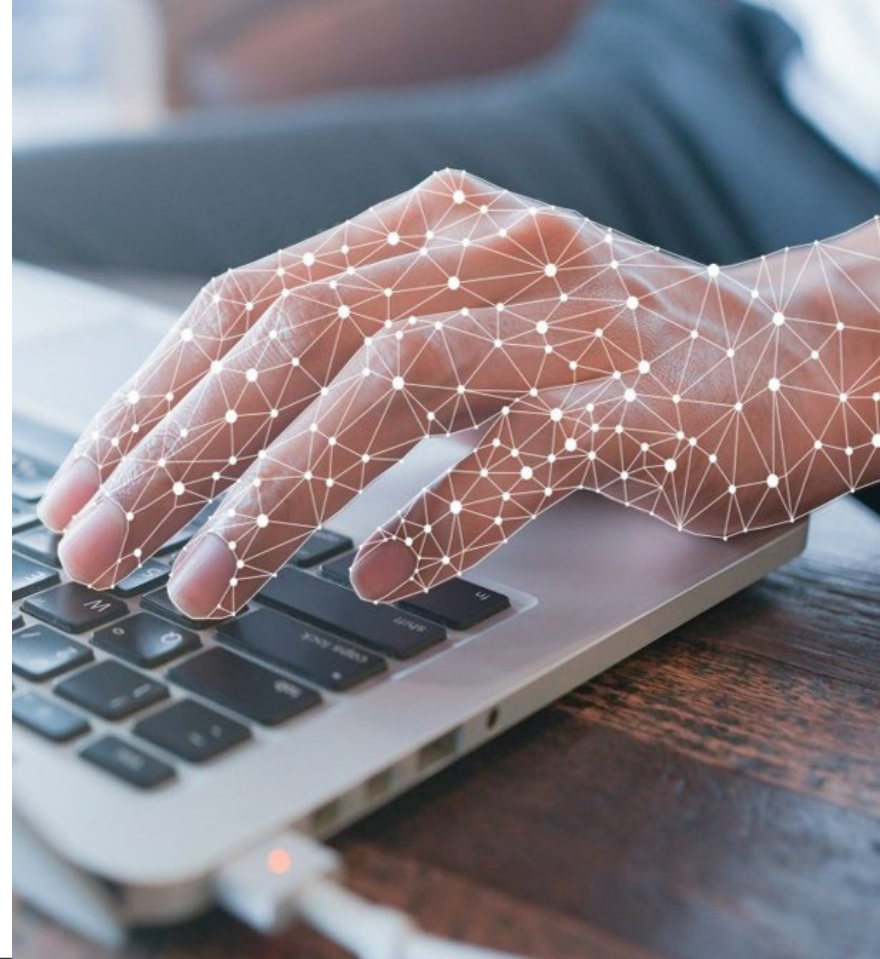
VS.

Low Essay Writing Auto
Score Complete

"Your outline is good, try to write the essay of maximum 1000 words, while respecting the writing guide ... "

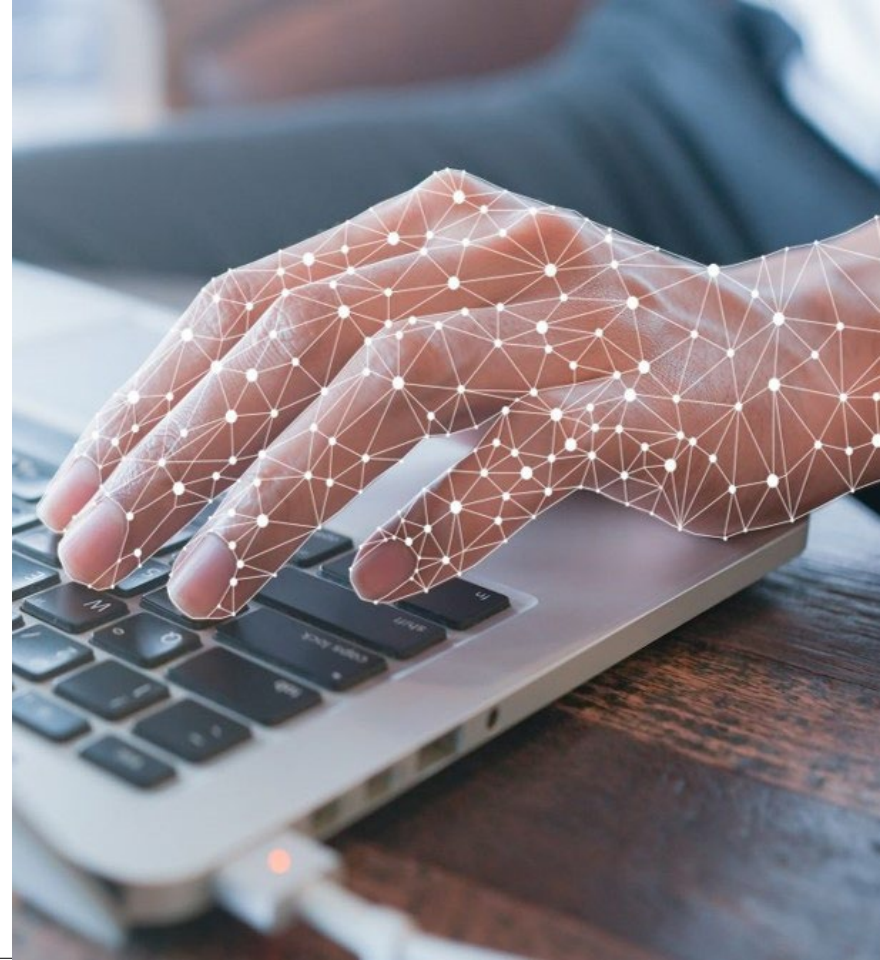
Practical Notes

- Time consuming? Teachers report average ~15 min per log
- Automated prompt classification?
 - “Fair” human-AI agreement (kappa 0.3-0.4)
 - High AI classification consistency (Fleiss kappa = .78)



Limitations

- Meta-prompting
- Unclear generalizability

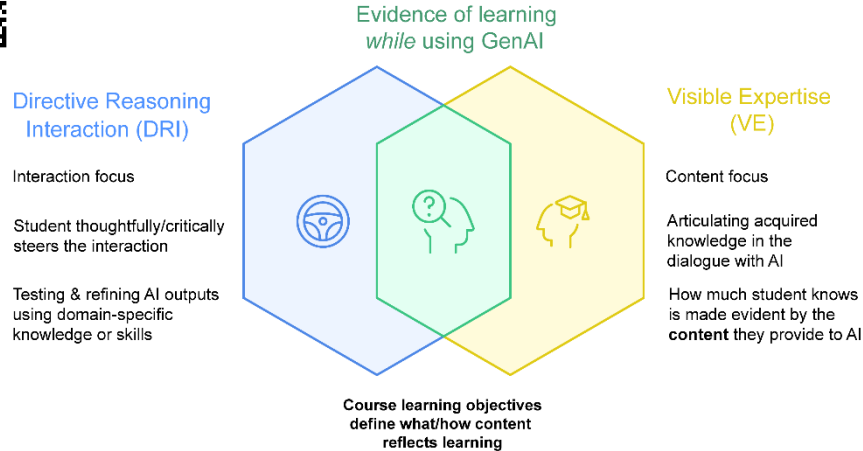


Prompt-grading seems to work to assess student learning in philosophy. **Would it also work in your discipline?**



DRIVE Framework

Detecting evidence of learning in students interactions with Generative AI



Thank you!

Manuel Oliveira

m.j.barbosa.de.oliveira@tue.nl

Extra

Sample

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

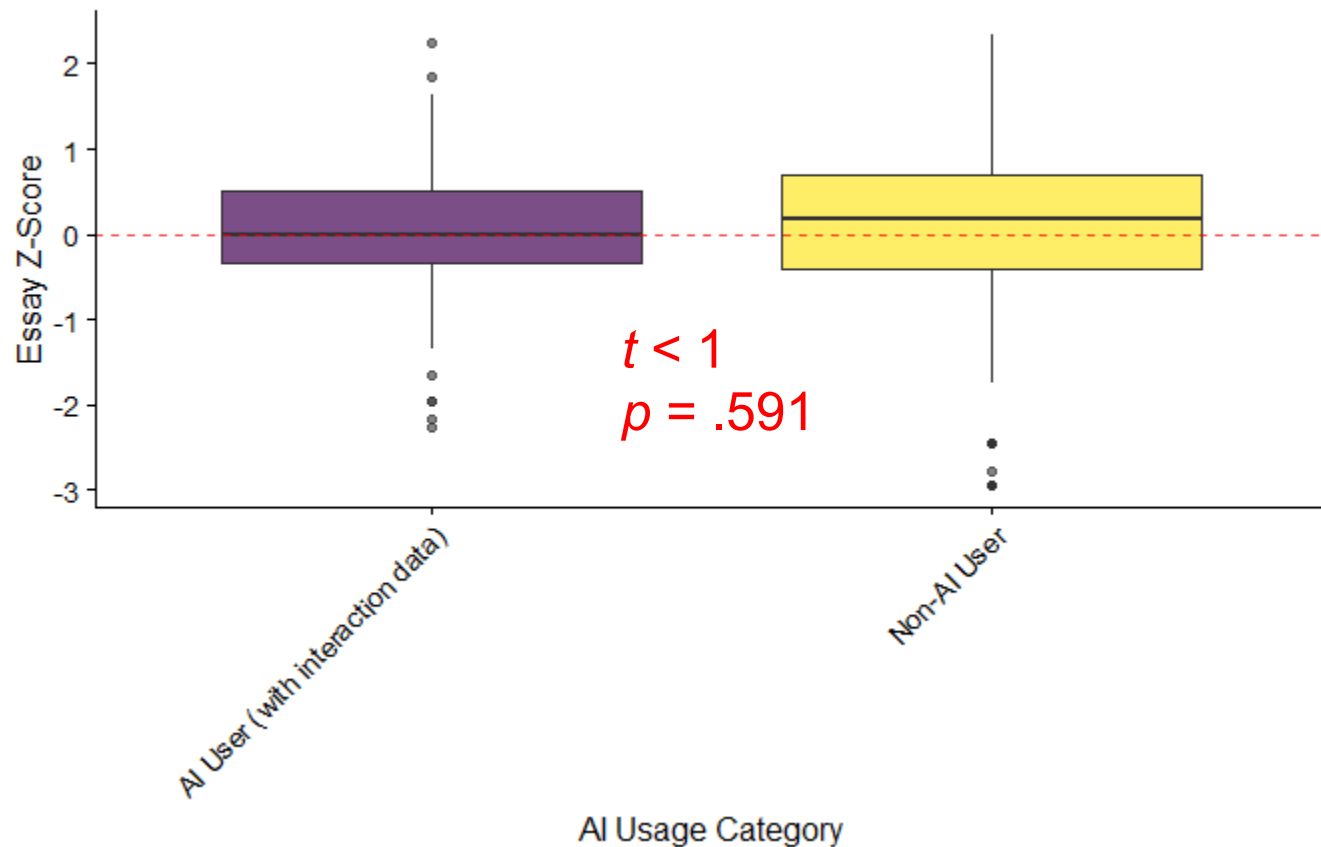
Prompt Statistics
(Annotated Essays Only, N = 70)

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.

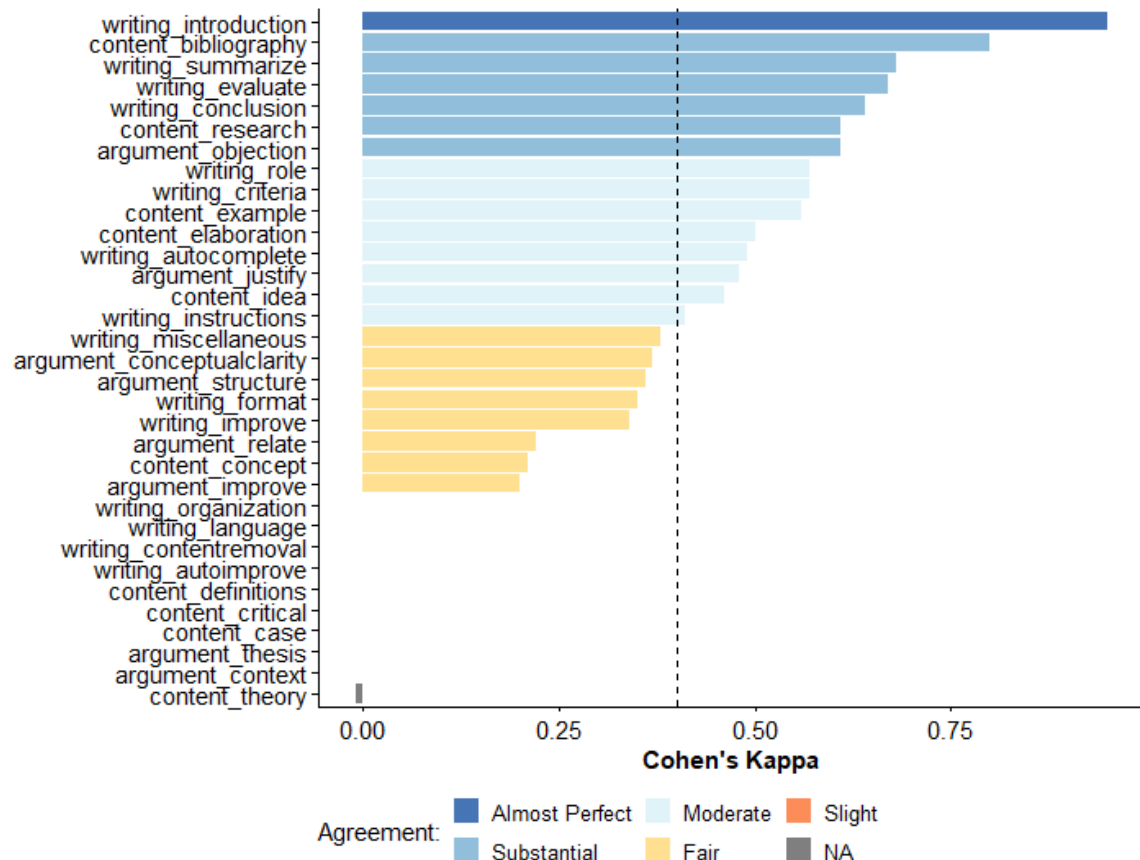
Essay Z-Score Performance by AI Usage Category

Red dashed line indicates average performance ($z = 0$)



Inter-rater agreement per taxonomy classification

Course: Philosophy & Ethics of AI 2024-2025



Note: Values below dashed line (0.40) represent lower than moderate agreement

Human-AI agreement

- **Model:** GPT-4o (March 2024 version) via OpenAI's API
- **Temperature:** 0.1 → for higher classification consistency (less “creativity”)
- **Top-p** (nucleus sampling): 0.1 → limits token selection to only most probable options
- **Prompt engineering instructions:** Tasked with classification using criteria (taxonomy) and allowed to use multiple labels for each input prompt (from student interaction log), similar to how Research Assistant annotator, as well as some other annotators, were labelling

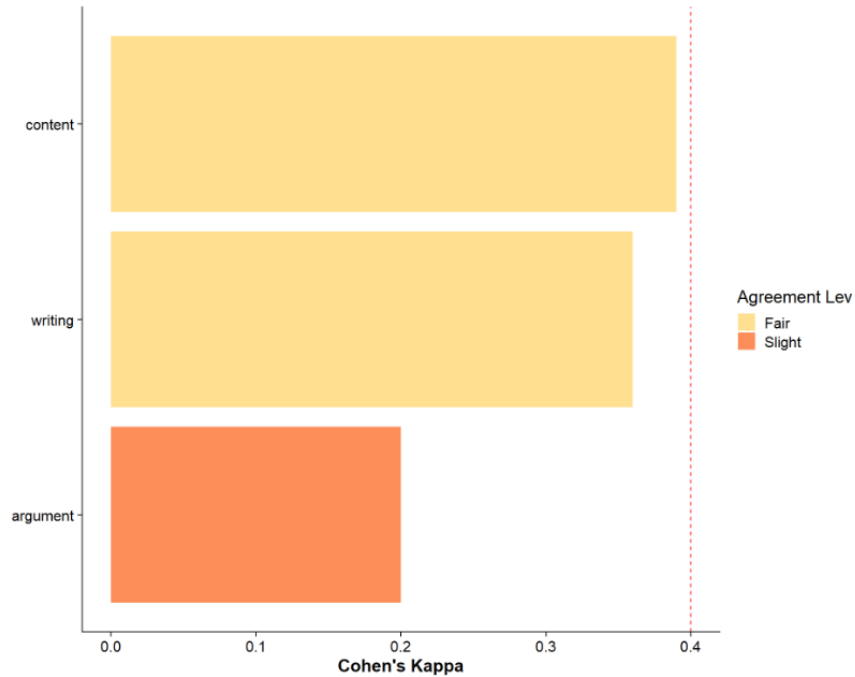
AI Reliability Analysis parameters:

- Sampled classifications for each input prompt: 5

Human-AI agreement

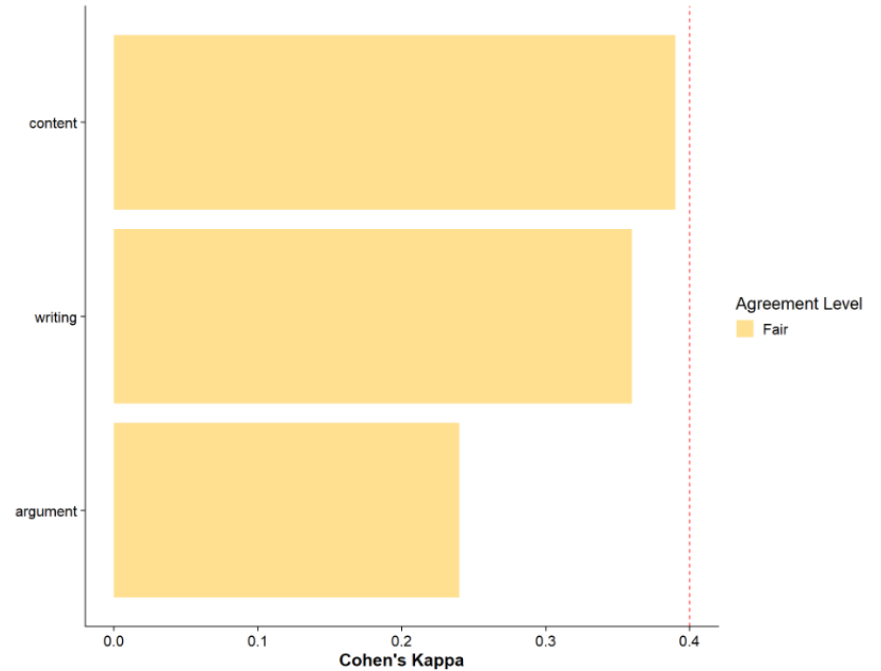
Show

Human 1 (Teachers) vs AI Agreement by Main Category



Show

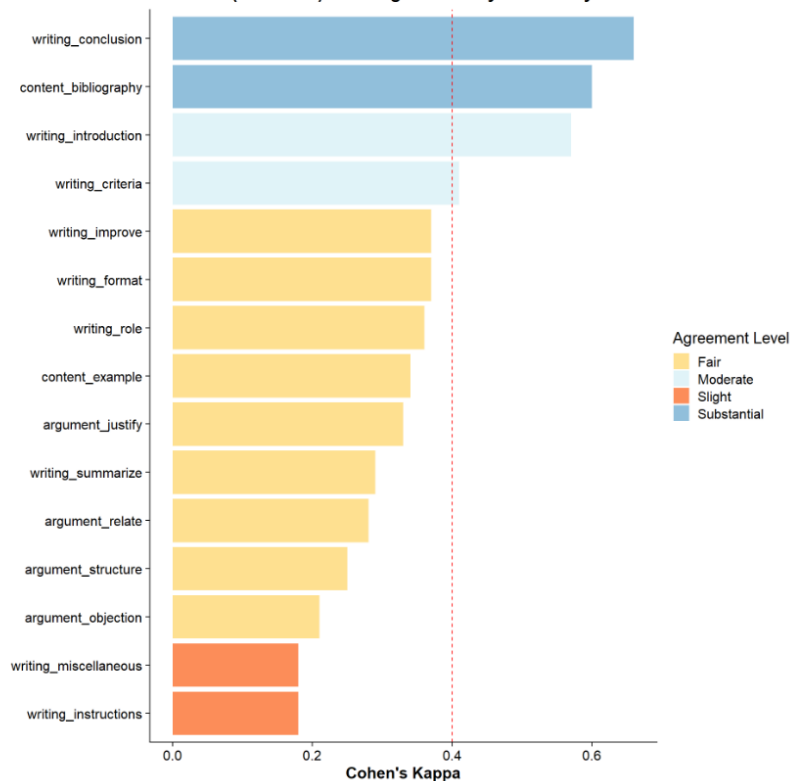
Human 2 vs AI Agreement by Main Category



Human-AI agreement

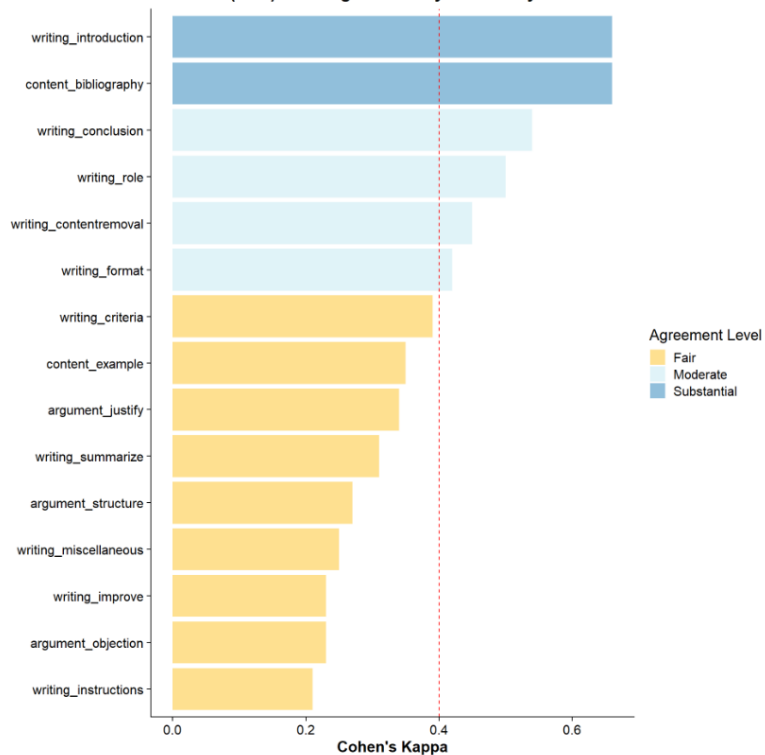
Show

Human 1 (Teachers) vs AI Agreement by Taxonomy Item

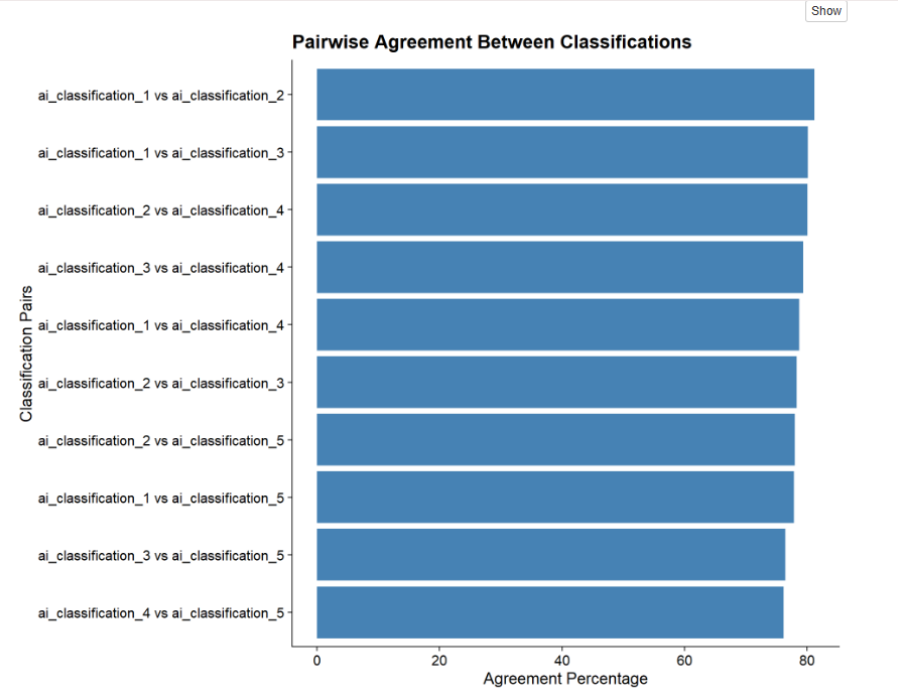
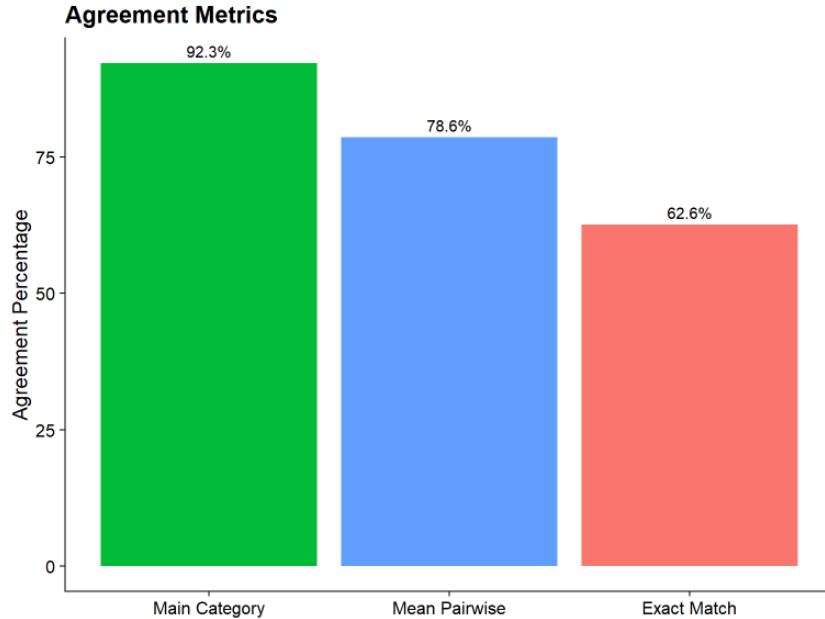


Show

Human 2 (R.A.) vs AI Agreement by Taxonomy Item



AI Classification reliability



AI Classification reliability

Summary of LLM Classification Reliability Metrics

Metric	Value
Sample Size	687
Classifications per Prompt	5
Exact Match Agreement	62.59%
Mean Pairwise Agreement	78.60%
Fleiss' Kappa	0.783
Kappa Interpretation	substantial
Main Category Agreement	92.29%
Mean Response Time	1.80 seconds
Response Time SD	0.60 seconds
API Uptime	99.94%
Error Rate	0.06%