Large Language Models for the Assessment of Students' Authentic Tasks. A Replication Study in Higher Education

Daniele Agostini[1,*,†], Federica Picasso[1,†] and Helga Ballardini[1,†]

[1]University of Trento, Department of Psychology and Cognitive Sciences, Corso Bettini, 84, 38068 Rovereto, Italy

Abstract

After the public release of ChatGPT (November 30th, 2022) and consequently, that of all its competitors, the use of Large Language Models (LLMs) has become widespread among the public. The most significant impact was perceived from the very beginning in the field of Education and Instruction [1, 2, 3, 4, 5, 6, 7]. Of particular interest for this paper is its use both by teachers and students in particular in the context of higher education [8, 4, 9]. The immediacy with which Large Language Models (LLMs) have been integrated into higher education practices, both by teachers and students, leads to questions of fundamental importance relating to their effectiveness and reliability. In this field, LLMs become the means through which teachers have the opportunity to revolutionise the interaction with students, the management of workload and the personalisation of each learning experience [2]. Although these technologies are recognised as having advantages and potential for improving learning in terms of accessibility and personalisation [7], a crucial question concerns their application in assessment practices, especially the ability to objectively and impartially evaluate students' performance. The possibilities of using these tools in the field of learning evaluation is relatively little known, which implies the need to delve deeper into the topic for its application both in pedagogical theory and in educational practice. A previous study has been already published [10] which explored the use of the main LLM in the specific context of assessing students' papers, and this is a replication study based on it. The purpose of the current study is to explore the possible use of the main LLMs in the specific context of evaluating students' written productions, with a focus on the aspects of accuracy that are evaluated with the help of a rubric proposed by the teacher. This article is part of a series of contributions that focus on this topic, in light of the principles and application of the AI-Mediated Assessment for Academics and Students (AI-MAAS) model [11].

Keywords

Large Language Models (LLMs), AI-Assisted Assessment, Rubrics, Authentic Tasks, Academic Assessment

1. The Context: AI Assessment in Higher Education

In the last two years, Large Language Models (LLM) have taken on a very significant role in the technological landscape thanks to the launch of ChatGPT, followed subsequently by the release of competing models. The impact of LLMs remained relatively limited over time until increasingly simple and intuitive user interface functions were introduced, firstly the "chat" level, which brought the general public closer to these tools. This phenomenon of "democratisation" boosted the commercial and large-scale use of LLMs, which led institutions, companies and individuals to increase investment in this sector [12, 13, 14, 15]. In addition to OpenAI's ChatGPT, Anthropic's Claude, Microsoft's Copilot and Google's Gemini are just some of the most used LLMs, in addition to the much more numerous open-source models to which Meta's LLAMA has given a notable boost At the same time, however, this has led to a crisis in search engines since LLMs, without requiring advanced research skills, offer new ways of querying and analysing data, more natural interaction and sufficiently precise

AIxEDU 2024: 2nd International Workshop on Artificial Intelligence Systems in Education, November 25–28, 2024, Bolzano, IT

[*] Corresponding author.

[†] D. Agostini: Conceptualisation, Methodology, Investigation, Formal Analysis, Writing - original draft, Writing - review & editing, Resources, Supervision. F. Picasso: Investigation, Writing - original draft, Data curation. H.Ballardini: Investigation, Writing - review & editing.

daniele.agostini@unitn.it (D. Agostini); federica.picasso@unitn.it (F. Picasso); helga.ballardini@unitn.it (H. Ballardini)
https://webapps.unitn.it/du/en/Persona/PER0247709 (D. Agostini);
https:https://webapps.unitn.it/du/en/Persona/PER0242228 (F. Picasso); https://webapps.unitn.it/du/en/Persona/PER0033179 (H. Ballardini)
0000-0002-9919-5391 (D. Agostini); 0000-0002-8381-6456 (F. Picasso); 0000-0002-3603-9551 (H. Ballardini)

and exhaustive answers. For example, LLMs allow users to avoid various typical inconvenient steps that characterise the standard use of search engines, such as the selection of long lists of websites, the acceptance of cookies and the appearance of advertising banners. As a result, educational institutions and agencies have begun to incorporate LLMs and generative AI into their curricula at various levels, developing courses to harness the potential of these innovative technologies. There is currently a strong emphasis on AI Literacy [16, 17], which allows professionals from different sectors, including educational institutions, to deepen their understanding of the fundamental elements of AI generative, the availability of tools, the functionalities and methods of use that make LLMs effective tools in all fields [18, 19, 20, 21, 22, 23]. However, one of the most critical issues concerns information management: LLMs possess enormous potential due to their ability to analyse and generate data; this raises numerous questions about accuracy, privacy and ethics in information management and ownership of output. The challenge in this continuously and rapidly evolving field becomes the ability to pay constant attention and critically evaluate so that end users always use LLMs responsibly [24, 25, 26, 27]. In relation to this issue, higher education institutions have reacted by placing themselves on the defensive, so much so that some universities, in order to counteract the possible use of LLMs by students during exam tests, they have reintroduced the obligation to write by hand and also take oral tests [8, 28]. At the same time, pieces of software created specifically to detect the productions generated by LLMs were introduced on the market. However, these turned out to be ineffective, causing management and legal problems for institutions because students could be unfairly accused of sending texts generated by artificial intelligence [29, 21, 30]. To avoid such inconveniences, national and international institutions and universities promptly provided themselves with guidelines that promote ethical behaviour towards the use of LLMs while maintaining a certain caution, allowing students and teachers to use them effectively to carry out tasks and benefit institutions. Important international bodies and universities moved in this direction, such as UNESCO [31, 32], the JISC National Center for AI [33], the Russell Group [34], the French National Ministry of Education [35], the US Department of Education [36] and University College London [37]. Assessment tasks have proven to be arguably the ones that can profit most from the AI technology, especially in terms of sustainability. However, caution is needed as LLMs without specific task adaptations have proven incapable and unreliable in managing students' assessments independently [38, 33], while LLMs supported by assessment tools have been shown to produce satisfactory results [39]. Above all, the use of artificial intelligence by students requires that teachers know how to take ethical aspects into consideration and act with responsibility when evaluating tasks and tests which results could have a great impact on students' careers (for example, motivation, grades, scholarships, acceptance into master's or doctoral programs).

2. Theoretical Framework

Since the 1980s the idea of being able to use computerised systems (and now also artificial intelligence) to assist educators in their assessment tasks and to be able to make precise, impartial and informed decisions has been present in much literature [40, 41]. The possibility of using LLMs for learning assessment had already been explored in the period immediately preceding the release of ChatGPT, while the use of transformer models, including OpenAI's GPT-3, were already well established. Tamkin et al. [42] emphasised their educational application, which included:

- Summary: LLMs are able to summarise even very long texts. This use can help students submit concise summaries. Furthermore, various parameters can be considered for the synthesis, and this supports educators in providing precise information on the elements of the text that will be evaluated.
- Questions and Answers: LLMs can "understand" various portions of text, answer questions, and ask questions when required. These features are useful for providing interactive feedback and learning experiences.
- Classification: LLMs can classify the text into predefined categories: this allows you to introduce assisted assessment or classify students' feedback.

- Plagiarism detection: by comparing the similarity between different texts, LLMs are very useful to detect potential cases of plagiarism among students or to identify the misuse of original materials by students.
- Assessment of knowledge: LLMs can assess students' understanding of a topic based on their written productions, especially if the information is generated from correct homework and with the help of an assessment rubric to refer to.

These five applications are fundamental to using LLMs in learning assessment. Following the introduction of ChatGPT and other universally accessible LLMs, UNESCO published the guidelines "AI and education: Guidance for policy-makers" [31], which suggest the following recommendations for learning assessment:

1. Test and implement AI technologies to support the assessment of various dimensions of skills and outcomes.
2. Use caution when using an automated assessment with closed-ended, rule-based questions.
3. Use AI-generated formative assessment as an integrated feature of learning management systems (LMS) to analyse learning outcomes more accurately and efficiently and reduce the risk of human bias.
4. Use the ability to provide AI-powered progressive assessments to regularly update students and parents.
5. Examine and evaluate the use of facial recognition and other artificial intelligence capabilities for users' recognition and their tracking in remote online assessments.

Based on these different theoretical approaches, recommendations and guidelines, the AI-Mediated Assessment for Academics and Students (AI-MAAS) model was developed which is currently under validation; it proposes two potential implementations of LLMs for the assessment of learning: the first one for formative evaluation and the second one for summative evaluation [11]. In both cases, the selected LLM must be able to evaluate using an assessment rubric provided by the teachers or by the students. Given the novelty of the tool, so far, there are not many experiments in this field. Martin et al. [39] worked on this opportunity, starting from the need to be able to assign students even complex tasks that involve a certain degree of reasoning, abstract conceptualisation and reprocessing of information; while the correction of these types of tasks (with a large quantity of long open answers) often proves to be an unsustainable task for teachers. Some researchers, working on this aspect, have demonstrated that in the evaluation of a chemistry task, for example, it is possible to use LLMs: in this case, an almost perfect match was obtained between the scores assigned by human raters and the scores generated by the LLMs. It should be highlighted, however, that Martin and colleagues did not simply use an LLM to achieve this result. The researchers used a complex procedure that involved, among other operations, the unsupervised machine learning technique HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise), a cluster mapping and training of a deep neural network classifier. The aim of this study was to test an operational model and demonstrate its feasibility. This excellent solution represents the result of models trained on specific tasks and populations; therefore, it cannot be assumed that the procedure applied can be replicated by any teacher not specialised in Machine Learning. Other studies have instead used LLMs for assessment purposes without comparing the performance of the AI with that of a teacher. These studies applied for example in the evaluation of L2 English tasks [43] and in supporting self-assessment came to satisfactory results [44]. Machine learning has also been applied in the evaluation of tasks related to STEM disciplines, but without using LLM [45].

Finally, a previous study [10] explored the use of the main LLMs in the specific context of assessing students' papers, with a focus on their accuracy in assessing according to a rubric developed by the teacher. The idea was that employing LLMs for assessment in higher education can enable the adoption of teaching and assessment approaches that were previously unsustainable and unscalable. This should help to ensure constructive alignment [46] and thereby improve the quality and effectiveness of university teaching. The study, aimed at selecting the most human-like evaluation amongst LLMs, highlighted that while some AI models, like ChatGPT-4 and Claude 2, performed well in most of the assessment criteria, others, such as Microsoft Copilot and Google Bard, were far from human-like assessment. The article recommends further research on ChatGPT and Claude, with potential inclusion of open-source models as well as involving multi-shot prompting, expanding the student sample, involving more evaluators, and refining and redesigning the rubrics.

3. Methodology and Tools

This is a replication study of "Are Large Language Models Capable of Assessing Students' Written Products? A Pilot Study in Higher Education" published in "Research Trends in Humanities Education & Philosophy, 11" [10] that follows the same methodology with updated LLMs employed, a greater sample of students and of human evaluators. It explores the use of leading Large Language Models (LLMs) in the specific context of assessing student written products, focusing on their precision and ability to evaluate according to a grading rubric developed by the teacher. The goal is to understand whether and which models can be used by university and non-university educators who are not experts in Machine Learning to assess students' written products, even in the presence of open-ended tasks and questions, thanks to grading rubrics. The pilot study was conducted at the University of Trento within the context of a university habilitation course for secondary school teachers, during the module concerning learning methodologies. One-hundred-fourty-two students participated anonymously, divided into 35 groups, along with 3 evaluating teachers, experts in experimental pedagogy and assessment. No data regarding the students' demographics was collected. The groups were tasked with carrying out an authentic task, namely to re-designing a past educational intervention that proved to be unsuccessful, targeted at a specific class (which could range from 1st grade of lower secondary school to 5th grade of higher secondary school, depending on the group composition). They were instructed to identify the past teaching approaches and strategies and to now think of different ones more suited to reach the intended learning outcomes. Furthermore, students' reflection and redesign ability was evaluated through the rubric of reference. To complete this task, groups were given two hours and thirty minutes, and a template for the educational design consisting of the following sections was provided: Involved Disciplines, Class and Grade Level, Intervention Title, Teacher, Programme and Learning Objectives, Context and Environment (formal, informal, type of setting, etc.). Moreover, a description of the reflection process applied for renewing the formative design is required, and it is considered under evaluation. In the part of the schedule with details, they were asked to explain the programming with concise descriptions of the various educational activities, the teacher's tasks, and those of the students. Within this framework, groups had the freedom to propose their original programming. The final product of each group is thus an MS Word file containing the programming of the educational intervention according to the described template. For the evaluation of the products, the following grading rubric (Table 1) was prepared, consisting of five evaluation criteria with four levels for each criterion.

Three expert human evaluators and seven LLMs (plus one that merged the feedback of all 7 models in a single one) evaluated all the student groups' products. The LLMs selected for this study were the most popular competing models at the time, and they are applied in the assessment process through the use of big-AGI (https://big-agi.com/). Big-AGI is an AI suite created to make advanced artificial intelligence accessible and was chosen for ease of adding several models through API, the possibility of imparting system prompts and the function (called "beam") for sending the same prompt to several LLMs at the same time. Human results were then compared with the results produced by the LLMs with various statistical analysis (see Method of Analysis section). The models used are:

1. GPT-4o: released in May 2024, GPT-4o is a multilingual, multimodal generative pretrained transformer developed by OpenAI. The model is capable of processing and generating text, images, and audio, making it a versatile tool for a wide range of tasks. Its multimodal capabilities

Table 1
Rubric for the Assessment of the Educational Intervention

Assessment Criteria	Insufficient Level (1 point)	Sufficient Level (2 points)	Good Level (3 points)	Excellent Level (4 points)
Understanding and applying educational architectures	Demonstrates a limited understanding of educational architectures, with applications not always appropriate or consistent.	Shows a basic understanding of educational architectures, applying them generally correctly but with some uncertainties.	Applies educational architectures correctly, with a good understanding of their use in the specific context.	Demonstrates a thorough understanding of educational architectures, applying them in an innovative and contextually relevant manner. Clearly justifies the choices made.
Selection and implementation of teaching and learning strategies	The teaching strategies chosen are limited or not always appropriate for the objectives of the intervention.	Uses some relevant teaching strategies, but their implementation could be more targeted or diversified in relation to the intervention goals.	Selects and implements appropriate teaching strategies with a good correlation to the intervention goals.	Selects and implements highly effective and diversified teaching strategies, perfectly adapted to the objectives and context of the intervention.
Definition of the intended learning outcomes	The objectives are vague, not measurable or not aligned with the chosen teaching architectures and strategies.	The objectives are present but could be more specific or better aligned with the teaching architectures and strategies.	The objectives are well-defined and generally aligned with the chosen teaching architectures and strategies.	The objectives are clear, specific, measurable and perfectly aligned with the chosen teaching architectures and strategies.
Detailed scanning of the intervention	The scan is incomplete, unclear, or lacks a logical progression of activities.	The scan is present but could be more detailed or better structured in some parts.	The scan is clear and generally well-structured, with a good progression of activities.	The scan is detailed, logical and well-structured, with a clear progression of activities and realistic timeframes.
Critical reflection on the redesign process	There is a lack of critical reflection on the changes made or the justifications are superficial.	Includes some reflection on the changes, but the analysis could be more thorough.	Provides good reflection on the changes, with clear links to the learning objectives.	Provides deep and critical reflection on the changes made, clearly justifying each choice in relation to the learning objectives.

enable a deeper integration of different data formats, enhancing its utility in complex applications. Link: https://openai.com/index/hello-gpt-4o/

2. Gemini 1.5 Pro Latest: a large language model developed by DeepMind (Google), is natively multimodal and supports an extended context window of up to two million tokens, which is currently the longest of any large-scale foundation model. This expansion in token capacity allows for processing more extensive sequences of data, thereby increasing its utility in tasks that require long-term contextual understanding. Link: https://deepmind.google/technologies/gemini/pro/.
3. Claude 3.5 Sonnet: developed by Anthropic, excels in the ability to understand nuanced language, humour, and complex instructions. It is designed to generate high-quality content in a relatable, natural tone, showing marked improvements in areas such as writing and human-centric communication. Link: https://www.anthropic.com/news/claude-3-5-sonnet.
4. Mistral Large (2402): is designed to excel in complex reasoning tasks, particularly in multilingual contexts. The model is highly effective in text understanding, transformation, and code generation.

It demonstrates top-tier performance in handling sophisticated reasoning challenges, making it a robust tool for both natural language processing and technical tasks. Link: https://mistral.ai/news/mistral-large/.

5. Open Mixtral 8x22B (2404): is one of the latest model developed by Mistral, featuring a sparse Mixture-of-Experts (SMoE) architecture. Despite its large size, with 141 billion parameters, only 39 billion parameters are actively engaged during processing, optimising both performance and cost efficiency. This approach sets new standards in the AI community for balancing model complexity with computational resource usage. Link: https://mistral.ai/news/mixtral-8x22b/.
6. Llama 3.1 70B Instruct Turbo: developed by Meta, is a 70-billion parameter language model designed for instruction-following tasks. The model is optimised to improve interactions where clear guidance or step-by-step reasoning is required, positioning it as an effective tool for applications in both academic and practical domains. Link: https://ai.meta.com/blog/meta-llama-3-1/
7. Qwen2 72B Instruct: developed by Alibaba Cloud, is a 72-billion parameter language model optimised for instruction-based tasks. It integrates the latest advancements in generative AI, offering improved efficiency in tasks ranging from conversational AI to complex text generation and reasoning. Its design caters specifically to high-performance needs in both commercial and research applications. Link:https://www.alibabacloud.com/en/solutions/generative-ai/qwen?_p_lc=1

All these LLMs can "understand" and write in Italian, but it cannot be ruled out that performance in English may be different (presumably better, since most of the training is done in that language). Mistral's models were added for their specific training with European languages that renders them "natively fluent in English, French, Spanish, German, and Italian, with a nuanced understanding of grammar and cultural context". Privacy shouldn't be a concern since there is no data saved to LLM provider servers due to our use of API on a local instance of big-AGI. All the conversations are saved only locally.

4. Prompting

This study aimed to understand which models could be used by university educators (and, potentially, other educators) to assess students' products. For this reason, overly sophisticated prompting techniques were not used; instead, what an educator might do by providing clear instructions and giving the necessary context data for evaluation was employed. The LLMs systems were promoted through the following instruction (originally written in Italian). The first one is the System Prompt:

```
You are an experienced and impartial university lecturer. Your job is
to assess the quality of student assignments according to a specific
assessment rubric.

**How to respond to requests:**

* Do not express personal opinions or subjective judgements. *
Focus exclusively on the criteria provided in the rubric. * Provide
a fair and impartial assessment based on the task's adherence to
the criteria. * Carefully review the student's entire paper before
beginning the assessment. * Offer constructive suggestions as to
how the student might improve. * Uses clear and concise language.
* Justify the marks awarded with specific references to the paper
and the rubric. * In your assessment, take into account that the
students only had 2 hours for planning.

**Request format

Each request will include:
```

> * **The student’s assignment:** The text of the assignment you are to assess. * **The grading rubric:** A list of criteria with descriptions for each grade level.
>
> **Response format:**
>
> Your answer should follow this format:
>
> **Title of the paper (also called title of the paper) as it appears in the document: [insert title here]**.
>
> **Total score:** [Insert total score here].
>
> **Scoring breakdown:**
>
> | Criterion | Score | Comments |–|–|–| | [Criterion 1] | [Score] | [Comments with specific examples from the task] | | [Criterion 2] | [Score] | [Comments with specific examples from the task] | | [Criterion 3] | [Score] | [Comments with specific examples from the task] | | ... | ... | ... |
>
> **Suggestions for improvement
>
> * [Suggestion 1] * [Suggestion 2] * ... **Answer following the answer format provided above.**

The second is the prompt that were given to the LLMs to assess the products (originally written in Italian):

> Evaluate the attached teaching design (student task) that was created by a group of students from the secondary school teaching qualification course. The key competence of this assignment lay in being able to design a teaching intervention that makes effective use of teaching architectures and strategies. In particular, the group’s competence in terms of redesign and depth of reflection is taken into account with respect to previous instructional design. At the same time, the instructional design had to prove effective in achieving the goals they set themselves. Take into account that the students only had 2 hours to design. Use the evaluation rubric below to assess:
>
> <Starting teaching design evaluation rubric>
>
> Evaluation rubric:
>
> Criterion 1 - Understanding and application of teaching architectures:
>
> - Insufficient (award 1 point): Limited understanding and applications not always appropriate. - Sufficient (award 2 points): Basic understanding with some uncertainties in application. - Good (awarded 3 points): Correct application and good understanding. - Excellent (awarded 4 points): Thorough understanding and innovative and relevant application.
>
> Criterion 2 - Selection and implementation of teaching strategies:
>
> - Insufficient (award 1 point): Limited or not always adequate strategies. - Sufficient (award 2 points): Relevant strategies but implementation can be improved. - Good (award 3 points): Strategies appropriate and related to the objectives. - Excellent (award 4 points): Highly effective, diverse and well adapted strategies.

```
Criterion 3 - Definition of learning objectives:

- Insufficient (award 1 point): Vague or non-measurable objectives.
- Sufficient (award 2 points): Objectives present but not very
specific. - Good (award 3 points): Well-defined and generally
aligned objectives. - Excellent (award 4 points): Clear, specific,
measurable and perfectly aligned objectives. Criterion 4 - Detailed
scanning of the intervention

- Insufficient (award 1 point): Incomplete or unclear scan. -
Sufficient (award 2 points): Scan present but can be improved in
structure. - Good (awarded 3 points): Clear and well-structured scan.
- Excellent (awarded 4 points): Detailed, logical and well-structured
scanning.

Criterion 5 - Critical reflection on redesign:

- Insufficient (award 1 point): Lack of critical reflection
or superficial justifications. - Sufficient (award 2 points):
Reflection present but not very thorough. - Good (awarded 3 points):
Good reflection with clear connections. - Excellent (award 4 points):
Deep and critical reflection, clear justifications.

<end of assessment rubric>

**Total score:**

**Scoring distribution:**
```

The prompt was sent simultaneously to all the LLMs involved. Through big-AGI the authentic task document was attached in PDF format. A zero-shot prompting procedure was used for all LLMs, meaning that no examples of human task assessments were given to the models. It is possible for a university instructor to provide an example that can enhance the quality of LLM assessments, however, the goal in this instance was to choose the most suitable models for this type of evaluation, not to find methods for optimising the results. Finally, an "eighth" LLM evaluator has been added, which uses the "Beam" function of big.AGI software. All the nuanced answers from the 7 LLM have been sent for consideration and synthesis to GPT-4o, resulting in an eighth assessment that considers the feedback from all seven LLMs.

5. Models' Settings

In order to set a common limit of length for every model's answers, all of them have been set through big.AGI API controls to 8128 tokens maximum. Also, the temperature was set to 0.2, that should ensure quite strict adherence to the instructions yet leave some room for creativity in answers.

5.1. Attention to Tokens and Context

Understanding tokens and context is crucial when using a Large Language Model (LLM). Tokens can be simplified as units of text that might consist of a word, part of a word, or even a single character. The characteristics of tokens can vary between models. However, it is generally safe to assume that, on average, English might require one to one and a half tokens per word, and Italian might need one and a half to two tokens per word. The context window, another essential concept, represents the number of tokens a language model can consider simultaneously when generating responses. This context depends on the model used and the available memory. Exceeding a model's context window could cause errors if it happens in a single prompt or, in a more extended conversation, the model might start ignoring the earlier parts of the dialogue to make room for more recent inputs. Therefore, preserving context is vital for generating coherent and relevant responses. It is important to note that not only the user's prompts consume context, but the model's responses do as well. To preserve the context window, some LLMs platforms impose a character limit on the prompts that can be sent and on the length of the generated responses, which are shorter than the maximum context window. Contrasting this replication study with the original one, it can be noted that context windows are decidedly wider than the ones that were found in LLMs one year ago, posing less of threat to the coherence of the assessment. Below Table 2 illustrates the maximum context window size for each of the models used:

Table 2
Context Windows of the used LLMs. The context windows refer to the APIs. Note that this feature may change with updates.

Large Language Model (versions available in Italy, September 2024)	Context Window (in tokens)
GPT-4o	128,000
Claude 3.5 Sonnet	200,000
Gemini 1.5 Pro Latest	2,000,000
Mistral Large (2402)	32,000
Mixtral 8x22B (2404)	64,000
Meta Llama 3.1 70B Instruct Turbo	131,072
Qwen2 72B Instruct	32,768

6. Method of Analysis

The analysis method for evaluating the data involved examines the levels assigned by each evaluator (both LLMs and humans) to the various criteria of the rubric for each of the 35 group products. Each of the seven evaluators assigned a level to each of the five criteria for every product, resulting in each evaluator assigning a level to a total of 175 criteria. Several statistical techniques were employed to extract insights from the data, including Principal Component Analysis (PCA), analysis of standard deviation, and the creation of a disagreement index among evaluators. Microsoft Excel and JASP (based on R) were used for the statistical analyses.

7. Results

The consistency of the assessment for different models has been tested over the evaluation of three random tasks from the sample for three times each by each one of the models. For this test, each LLM assessed a total of 45 criteria.

From these tests, the following behaviours were observed:

- GPT-4o, Gemini 1.5 Pro Latest and Claude 3.5 Sonnet were extremely consistent, with only one instance of a different assessment for one criterion, by just one point.
- Mistral Large 2402 was perfectly consistent with zero instances of different assessments.
- Open Mixtral 8x22B (2404) and Qwen2 72B Instruct were quite consistent, with five instances of different assessments of single criterion by one point.
- Llama 3.1 70B Instruct Turbo: was the fairly inconsistent, with 19 instances of different criteria assessment by one point.

7.1. Principal Component Analysis

The first analysis conducted, in addition to descriptive data, was the PCA, a dimensionality reduction technique that allows the identification of latent variables within the data and that can represent a general model of the data. Three principal components were identified from the PCA conducted on the assessment data (Table 3).

Table 3
PCA Component Loadings

Evaluator	RC1	RC2	Uniqueness
e1 (GPT-4o)	0.579	0.327	0.472
e2 (Gemini 1.5 Pro)		0.421	0.832
e3 (Claude 3.5 Sonnet)	0.695	0.312	0.321
e4 (Mistral Large 2402)	0.775		0.430
e5 (Mixtral 8x22B 2404)	0.651		0.592
e6 (Llama 3.1 70B Instruct)	0.743		0.405
e7 (Qwen2 72B Instruct)	0.825	-0.326	0.336
e8 (Merge of 7 LLMs by GPT-4o)	0.802		0.301
e9 (Human Evaluator 1)		0.515	0.708
e10 (Human Evaluator 2)		0.753	0.420
e11 (Human Evaluator 3)		0.644	0.604

Note. Applied rotation method is promax.

The first component (RC1) is formed by evaluators e1, e3, e4, e5, e6, e7, e8 loadings, which correspond respectively to the LLMs GPT-4o, Claude 3.5 Sonnet, Mistral Large, Mixtral 8x22B, Llama 3.1 70B, Qwen2 72B and the merge of LLMs opinion. The second component (RC2) comprises those of e2, e9, e10 and e11 corresponding to Gemini 1.5 Pro and human evaluators 1, 2 and 3. As can be appreciated in Figure 1, Both GPT-4o and Claude 3.5 Sonnet contributes mainly to RC1 component but also to RC2. Gemini Pro 1.5 on the other hand, contributes only to RC2 component (the tiny loading to RC1 is negative). Trying to name the identified components, RC1 could be called "LLM Evaluation Pattern" and RC2 "Human Evaluation Pattern".

7.2. Analysis of Standard Deviation of Grades by Product and Assessment Criterion

To understand how assessments differed from criterion to criterion and from evaluator to evaluator, an analysis was conducted on the standard deviation (SD) of the different variables of the study. The criteria, numbered or abbreviated in some of the graphs, are those listed in Table 4. Firstly, an effort was made to identify which assessment criteria had the slightest and the most SD (Table 4) to understand which were assessed more consistently by all evaluators. The criteria with the minimum SD across all products is Criterion 4 and 1 (“Detailed scanning of the intervention” and “Understanding and application of teaching architectures”), with an average of about 0.5. This suggests a high level of agreement among evaluators in assessing the quality and details of the detailed activities envisaged in the educational design and the understanding and correct application of the teaching architectures at their bases. On the other hand, the criterion with the maximum SD among all activities is Criterion 5 (Critical reflection on redesign), with an average of about 0.8. This indicates a higher level of disagreement or inconsistency in how evaluators assessed the quality of teacher’s critical reflection about their past activities and the way in which they tried to improve them.

7.3. Agreement Index

An "Agreement Index" (AIdx) was developed to obtain a more robust metric and better understand which evaluators assigned more similar scores for the various criteria. This index combines the average difference between the scores assigned to a criterion and the variability of this difference. It was calculated to understand which evaluators are most similar to the human ones for each criterion. While LLMs evaluators are treated individually, the human benchmark is an average of the human evaluators’ (e9, e10 and e11) assessments. It is constructed as follows:

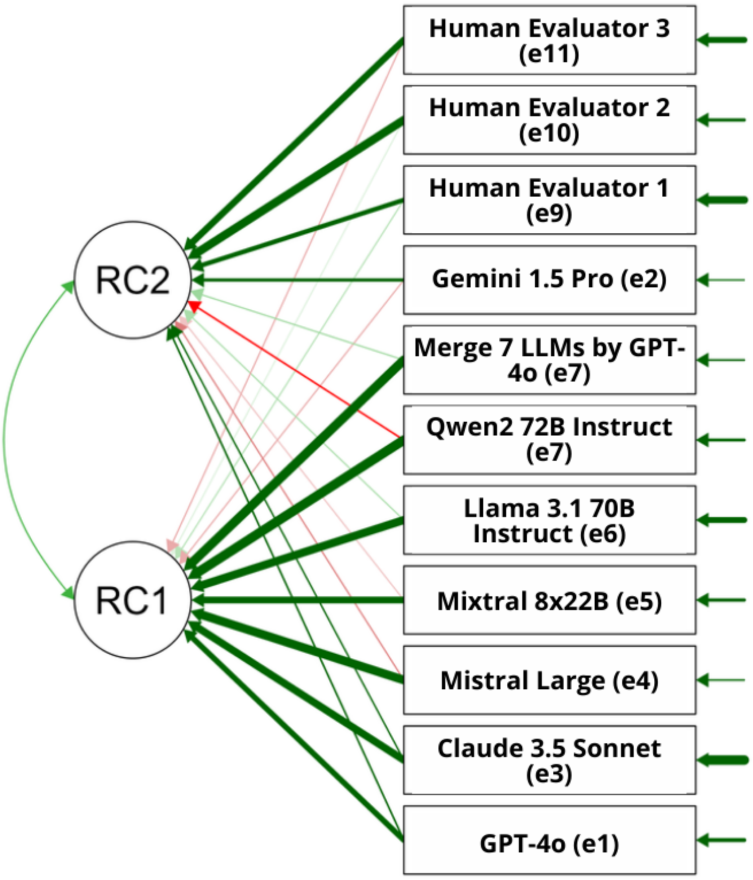

Figure 1: PCA Path Diagram. The diagram shows two main components (RC1 and RC2) and their relationships with different evaluators. RC1 represents the "LLM Evaluation Pattern" while RC2 represents the "Human Evaluation Pattern".

Table 4
Average standard deviation of scores assigned to criteria

Criterion	Description	Average Standard Deviation	Percentage of Total Range (1-4)
5	Critical reflection on redesign	0.76	25%
2	Selection and implementation of teaching strategies	0.64	21%
3	Definition of learning objectives	0.61	20%
1	Understanding and application of teaching architectures	0.54	18%
4	Detailed scanning of the intervention	0.53	17%

$$\text{AIdx} = \frac{\text{Average difference} + \text{Variability of the difference}}{2} \tag{1}$$

Therefore:

- The "Average Difference" is the absolute average difference in scores assigned between the evaluator in question and the average of human evaluators across all tasks and criteria.

- The "Variability of the Difference" is the standard deviation of the difference scores between the tested evaluator and the reference evaluator, reflecting how consistent these differences are across different tasks and criteria.

AIdx is calculated individually for each evaluator. It provides a single measure that encapsulates the average magnitude of evaluation differences relative to the reference evaluator and the consistency of such differences. A lower value indicates a more significant overall agreement in evaluation relative to the human evaluator. The highest possible value for the index for an evaluator would be achieved if they constantly evaluated at the maximum difference from the human evaluators (3 points).

The LLM evaluator who provided assessments most similar to the average of human evaluators (calculated through the AIdx) is GPT-4o, followed at a negligible distance by Claude 3.5 Sonnet. On the other hand, the LLM evaluator with the worst AIdx is Qwen2 72B (Table 5).

Table 5
Agreement Indices (AIdx) with reference to the average of human evaluators.

Evaluator	Agreement Index with "average human" (lower is better)
Human 3	0.39
Human 2	0.41
Human 1	0.45
GPT 4o	0.46
Claude Sonnet 3.5	0.47
Merge (GPT4o)	0.49
Llama 3.1 70B Instruct Turbo	0.52
Mixtral 8x22B (2404)	0.53
Mistral Large (2402)	0.53
Gemini1.5 Pro Latest	0.56
Qwen2 72B	0.72

Focusing on the single criterion (Table 6), it can be noted how AIdx with other evaluators vary from criterion to criterion. Unexpectedly, Qwen2 72B, the worst on the general AIdx with the “average” human evaluator, is the single model that is most human-like in three criteria out of five. Its main problem is that it assessed in a very different way from humans the most difficult criterion: criterion number 5 “Critical reflection on redesign” (Table 6). It also did not fare optimally in criterion number 3 “Definition of learning objectives”. Other LLMs like GPT-4o and Claude 3.5 Sonnet, as well as the Merge of the different LLMs feedback, keep a good Agreement Index across the board.

7.4. Assessment correlations among LLM and Human evaluators

As reported in Table 7 the model with higher correlation with human evaluation is by far Gemini 1.5 Pro (r = 0.84), followed at a distance by Claude Sonnet 3.5 (r = 0.66), then by GPT-4o (r = 0.59), the merge of the LLMs feedback by GPT-4o (r = 0.58) and Llama 3.1 70B (r = 0.45). This suggests that Gemini 1.5 Pro’s pattern of scores across the criteria is the most similar to that of the human evaluators.

But what happens excluding single criteria from the correlation analysis? That could help in understanding what criteria makes the assessment “human” and what LLMs struggle with:

- **Excluding Criterion 3 (Definition of learning objectives)**: When Criterion 3 is excluded all LLMs’ correlation indexes significantly improve. Noticeably, for Mistral Large and Qwen2 72B the jump is from being hardly correlated, or not at all (r = 0.27 and -0.06 respectively), to being significantly correlated (r = 0.86 and 0.88). Excluding Criterion 3 also significantly reduces the correlation of Gemini 1.5 Pro suggesting that this was the Criterion that it got right and

Table 6
Agreement Indices (AIdx) of evaluators compared to the average human evaluator divided by criterion

Crit.	GPT 4o	Gemini 1.5 Pro Latest	Claude Sonnet 3.5	Mistral Large (2402)	Mixtral 8x22B (2404)	Meta Llama 3.1 70B	Qwen2 72B	Merge (GPT-4o)	Best LLM	Second Best LLM
1	0.50	0.41	0.50	0.43	0.41	0.43	0.39	0.47	Qwen2 72B	Gemini 1.5 Pro / Mixtral 8x22B
2	0.56	0.64	0.57	0.47	0.55	0.57	0.40	0.55	Qwen2 72B	Mistral Large
3	0.43	0.60	0.43	0.49	0.68	0.62	0.70	0.41	Merge (GPT-4o)	GPT-4o / Claude Sonnet 3.5
4	0.37	0.46	0.35	0.37	0.48	0.38	0.33	0.32	Merge (GPT-4o)	Qwen2 72B
5	0.41	0.63	0.45	0.80	0.44	0.55	1.24	0.62	GPT-4o	Mixtral 8x22B

Table 7
Pearson correlation coefficients calculated between the average scores per criterion for each Large Language Model (LLM) and the average scores per criterion of human evaluators. Single criteria have been excluded to understand what makes the pattern "human".

LLM	Total	Excl. Crit. 1	Excl. Crit. 2	Excl. Crit. 3	Excl. Crit. 4	Excl. Crit. 5
GPT-4o	0.59	0.45	0.60	0.65	0.74	0.66
Gemini1.5 Pro Latest	0.84	0.79	0.99	0.47	0.82	0.86
Claude Sonnet 3.5	0.66	0.54	0.66	0.78	0.71	0.81
Mistral Large (2402)	0.27	0.16	0.21	0.86	0.21	0.88
Mixtral 8x22B (2404)	0.15	0.24	0.01	0.53	0.06	0.19
Llama 3.1 70B Instrict Turbo	0.45	0.34	0.42	0.74	0.47	0.69
Qwen2 72B	-0.06	-0.20	-0.13	0.88	-0.11	-0.77
Merge (GPT4o)	0.58	0.45	0.58	0.75	0.63	0.75

mostly contributed to its excellent general correlation to humans' assessment. This suggests that Criterion 3 may be peculiarly human-like in its application, which these models struggle to mimic accurately. The high increase implies that Criterion 3 might involve a complex judgment that those models are incapable to handle or contextual information that is not being passed to the model.

- **Excluding Criterion 2 (Selection and implementation of teaching strategies)**: Excluding Criterion 2 doesn't change LLMs correlation with human assessment, except for Gemini 1.5 Pro Latest. Gemini shows an almost perfect correlation of r = 0.99 when Criterion 2 is excluded, which is remarkable, but, even in this case, this criterion doesn't seem to be crucial.
- **Excluding Criterion 5 (Critical reflection on redesign)**: The exclusion leads to a substantial

increase in correlation for Mistral Large (from r = 0.27 to 0.88) and a notable improvement for several other models. This criterion, similarly to Criterion 3, may also represent aspects of human judgement that are challenging for models to replicate accurately.

8. Discussion

Regarding the goal of understanding whether educators without expertise in machine learning can employ current Large Language Models (LLMs) to assess students' written authentic tasks using assessment rubrics, the analyses have revealed several interesting elements:

- Differently from a previous iteration of the study [10], all the models have enough context window to perform this task.
- From the PCA, it appears that human evaluators generally have a different pattern of evaluation compared to LLMs.
- In contrast with the evaluation pattern, the Agreement Index (AIdx) measures both the magnitude of the score differences and their consistency. A high Agreement Index value suggests significant discrepancies between the model's scores and the average human scores, despite possibly similar trends in the pattern. Transforming in percentage the AIdx of each model referred to the average human an accuracy metric has been achieved. This helps to better visualise each model's performance (Fig. 2 and Fig. 3)
- Only Llama 3.1 70B was inconsistent in the repeated assessment of the same task.
- Gemini 1.5 Pro is the LLM model with the evaluation pattern more similar (with by far the higher correlation) to the human's (see Table 7). It is the only model that in the PCA results only in the component of human assessment (Fig. 1). On the other hand, its AIdx was the second worst, just before Qwen2 72B (Table 5, Fig. 2).
- GPT-4o and Claude 3.5 Sonnet have evaluation patterns not too dissimilar from the human's (Fig 1, Table 7) and on average attribute marks more similar to humans than any other model (Table 5).
- Llama 3.1 70B Instruct was the best of the open models, and the fourth in total (Table 5), after the already mentioned three proprietary models. It behaved quite well in the correlation index with the humans' assessment pattern with a moderate correlation (Table 7) and has a good AIdx. The problem with this model is the inconsistency of the assessment of the same task, where it "changed its mind" 19 times out of 45. It would be interesting to understand if that inconsistency has to do with the quantisation applied by Together AI, the API provider used.
- Mixtral 8x22B and Mistral Large fared similarly with patterns quite dissimilar to the human's and AIdx which are pretty decent (similar to Llama's). The correlation of Mistral Large with the human pattern of evaluation, when Criterion 3 is removed, is the second highest, thus giving reasons to follow it closely and keep it in the test pool.
- Qwen2 72B, an open LLM, would have been by far the best LLM overall (and Mistral Large would have been the second) if it weren't for Criterion 3. Criterion 3 posed a grave problem for Qwen2 both from the assessment pattern and from the AIdx point of view (Fig. 2 and Fig. 3).
- Criterion 3 (Definition of learning objectives), in a larger part, and Criterion 5 (Critical reflection on redesign), in a smaller part, appear to be the most discriminative criteria in terms of capturing what makes the human evaluation pattern unique for this assessment task (Fig. 2 and Fig. 3). These criteria likely involve nuances and complexities in judgment that are particularly human-like and challenging for LLMs to capture accurately, or the authors might have failed to provide all the relevant contextual information regarding these criteria to LLMs. This last hypothesis seems relevant because, in the previous iteration of the study, this same criteria was the easiest one for LLM to assess in a human-like manner [10].

Based on the available data, it appears that the more suitable LLMs for the assessing students' authentic tasks using an assessment rubric are Claude 3.5 Sonnet and GPT-4o. That is because they fared well both on the assessment pattern (PCA and correlations) and in the agreement index (magnitude and consistencies of scores). On the other hand, Gemini 1.5 Pro is the one that had by far the most human-like assessment pattern, but fell short on the AIdx, attributing marks that were very different from the humans'.

Qwen2 7B and Llama 3.1 70B deserve a mention as they are open models, and if not for some flaw would have been at the level (or better than) the aforementioned proprietary models. Llama has a problem of inconsistency of the marks assigned for each criterion, while Qwen2 really just got one criterion very wrong. It might be useful to know that for both of them, Together AI (https://www.together.ai/) was used as an API provider. It applies quantisation of Floating Point 8-bit (FP8) for Llama 3.1 70B Instruct Turbo, while Qwen2 72B Instruct is run at full-precision Floating Point 16-bit (FP16).

Human evaluators have a pattern of evaluation (see the PCA) that can be usually distinguished from the LLMs' one, but Gemini 1.5 Pro, if not for its very different score attribution, has very similar patterns. It is interesting to note that human evaluators among themselves have different score attributions (see Table 5 and Figure 3), but as for critical criteria they assess similarly.

All considered, presently, none of the LLMs can be used for autonomous evaluation for all criteria, especially regarding the more complex and the less contextualised ones. This confirms what Webb [33] highlighted. However, Claude 3.5 Sonnet, GPT-4o and, with some caution, Qwen2 72B Instruct have the potential to be used as solid support for evaluation for the summative evaluation level as described in the AI-MAAS (AI-Mediated Assessment for Academics and Students) model [11].

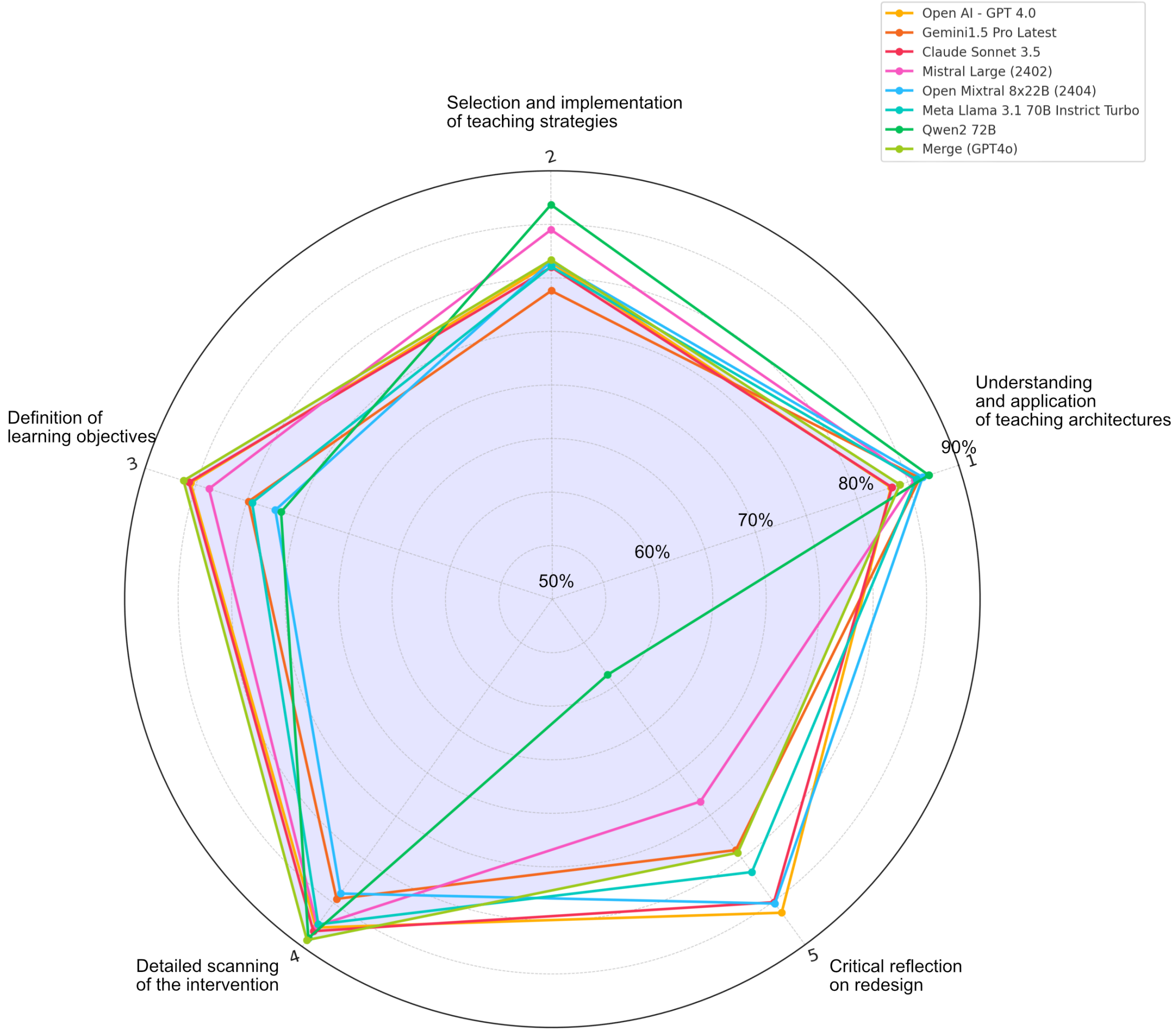

Figure 2: Radar graph of the LLMs' AIdx (transformed in percentage) with the average human for each criterion.

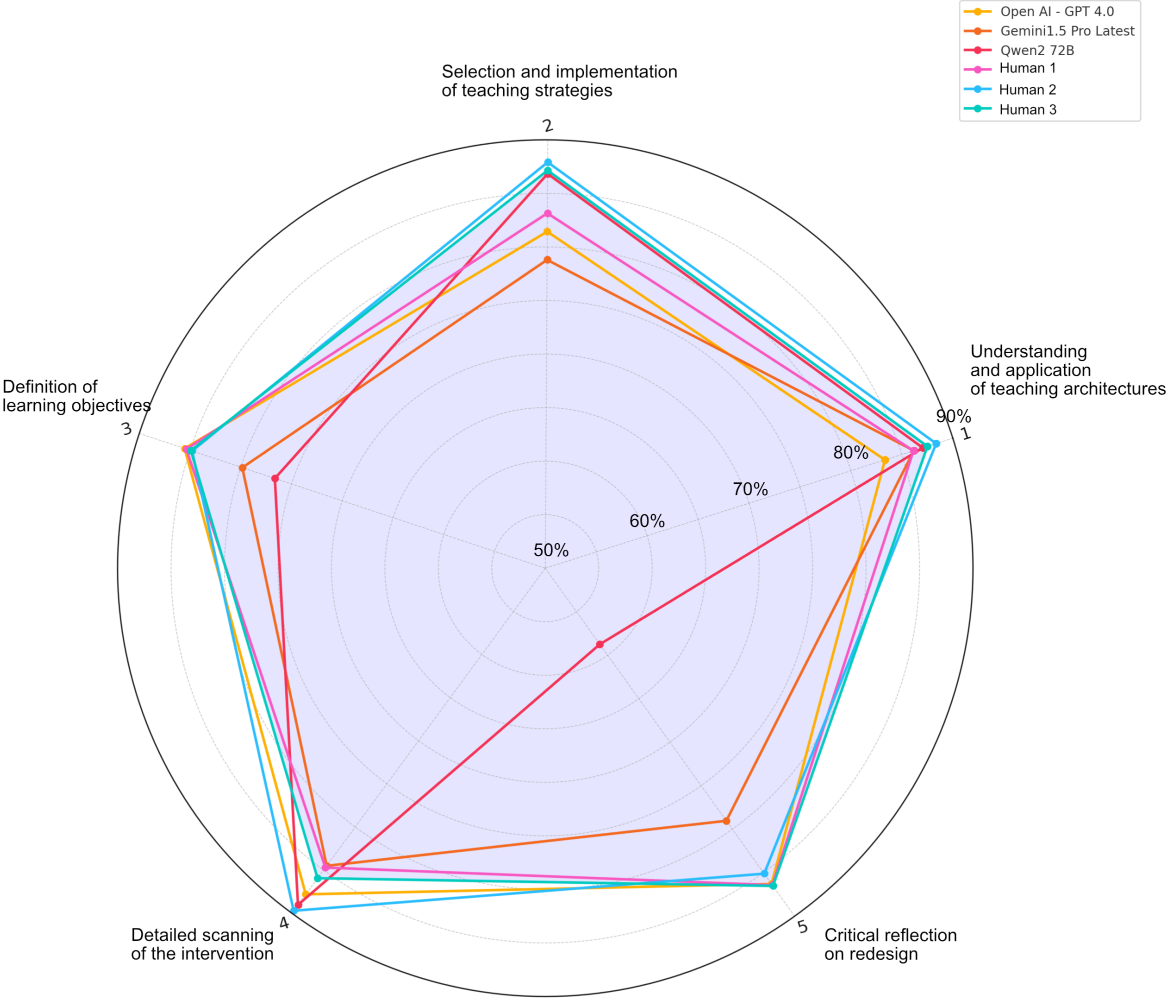

Figure 3: Radar graph of the LLMs' and Humans' AIdx (transformed in percentage) with the average human for each criterion.

9. Conclusions

The fundamental question of this study was whether and which current Large Language Models (LLMs) can be used by university educators (but it applies to other educators and instructors, too), even those without technical experience, to assess student-written authentic products in the presence of open tasks and questions using assessment rubrics. Indeed, using these technologies could make assessment more sustainable and scalable, allowing for more consistent alignment with declared learning objectives. This study has allowed us to determine that Claude 3.5 Sonnet, GPT-4o and, with some caution, Qwen2 72B Instruct have the potential to be used as solid support for summative evaluation. According to this study, the use of LLMs can be beneficial, but only if they are used under proper supervision. They should be seen as assistance for university educators and not as a substitute for assessments. The available data does not indicate that they are reliable enough to perform assessments independently, even if they are getting close to it. In fact, some criteria that is too complex or needs additional information about the context or specific subject can be evaluated in a way that is not in line with human assessment. This finding confirms the guidelines as stated by Miao et al. [31] and Webb [33]. The limitations of the present study lie in the sample size of student products that need to be significantly increased, as well as the number of human expert evaluators and the disciplines involved in the tests. The assessment rubric can also be optimised and, especially for the most critical criteria (such as Criterion 3), it would be important to experiment on its formulation to understand if it could have been a human error in defining the criteria that made it difficult to interpret by the LLMs. The idea behind this study is that it should be expanded and updated on a rolling basis to adjust the discussion and bring useful novelties into the assessment practice. Future evolutions of the study might include multi-shot prompting and the evaluation of textual feedback and assessment to tasks. Feedback that could be provided during the assessment for each of the criteria provided in a rubric deserve particular exploration [11, 32, 9, 42].

Acknowledgments

The authors would like to thank Elena Benini, PhD Student, for her contribution on the assessment of students' authentic tasks. Thanks also to Prof. Massimo Stella for the fruitful discussion about statistical methods. Both of them work at the Department of Psychology and Cognitive Sciences of the University of Trento.

References

[1] A. Baytak, The acceptance and diffusion of generative artificial intelligence in education: A literature review, Current Perspectives in Educational Research 6 (2023) Article 1. doi:10.46303/cuper.2023.2.

[2] S. Elbanna, L. Armstrong, Exploring the integration of ChatGPT in education: Adapting for the future, Management & Sustainability: An Arab Review 3 (2023) 16–29. doi:10.1108/MSAR-03-2023-0016.

[3] A. Extance, ChatGPT has entered the classroom: How LLMs could transform education, Nature 623 (2023) 474–477. doi:10.1038/d41586-023-03507-3.

[4] S. Roy, V. Gupta, S. Ray, Adoption of AI ChatBot like Chat GPT in Higher Education in India: A SEM Analysis Approach, Economic Environment 4 (2023) 130–149. doi:10.36683/2306-1758/2023-4-46/130-149.

[5] N. Saif, S. U. Khan, I. Shaheen, A. Alotaibi, M. M. Alnfiai, M. Arif, Chat-GPT; validating Technology Acceptance Model (TAM) in education sector via ubiquitous learning mechanism, Computers in Human Behavior (2023) 108097. doi:10.1016/j.chb.2023.108097.

[6] C. K. Tiwari, M. A. Bhat, S. T. Khan, R. Subramaniam, M. A. I. Khan, What drives students toward ChatGPT? An investigation of the factors influencing adoption and usage of ChatGPT, Interactive Technology and Smart Education (2023). doi:10.1108/ITSE-04-2023-0061.

[7] F. Kamalov, D. Santandreu Calonge, I. Gurrib, New era of artificial intelligence in education: Towards a sustainable multifaceted revolution, Sustainability 15 (2023) 12451.

[8] M. Perkins, Academic Integrity Considerations of AI Large Language Models in the Post-Pandemic Era: ChatGPT and Beyond, Journal of University Teaching and Learning Practice 20 (2023).

[9] M. Sullivan, A. Kelly, P. Mclaughlan, ChatGPT in higher education: Considerations for academic integrity and student learning, Journal of Applied Learning & Teaching (2023). doi:10.37074/jalt.2023.6.1.17.

[10] D. Agostini, Are large language models capable of assessing students' written products? A pilot study in higher education, Research Trends in Humanities Education & Philosophy 11 (2024) 38–60.

[11] D. Agostini, F. Picasso, Large language models for sustainable assessment and feedback in higher education, Intelligenza Artificiale 18 (2024) 121–138.

[12] T. Babina, A. Fedyk, A. X. He, J. Hodson, Firm Investments in Artificial Intelligence Technologies and Changes in Workforce Composition, Working Paper 31325, National Bureau of Economic Research, 2023. doi:10.3386/w31325.

[13] Generative AI to become a $1.3 trillion market by 2032, re-

search finds, 2023. URL: https://www.bloomberg.com/company/press/generative-ai-to-become-a-1-3-trillion-market-by-2032-research-finds/.

[14] G. Hammond, Big tech outspends venture capital firms in AI investment frenzy, 2023. URL: https://www.ft.com/content/c6b47d24-b435-4f41-b197-2d826cce9532.

[15] Y. S. Lee, T. Kim, S. Choi, W. Kim, When does AI pay off? AI-adoption intensity, complementary investments, and R&D strategy, Technovation 118 (2022) 102590. doi:10.1016/j.technovation.2022.102590.

[16] D. Long, B. Magerko, What is AI literacy? Competencies and design considerations, in: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, 2020, pp. 1–16.

[17] D. T. K. Ng, J. K. L. Leung, S. K. W. Chu, M. S. Qiao, Conceptualizing AI literacy: An exploratory review, Computers and Education: Artificial Intelligence 2 (2021) 100041.

[18] G. Biagini, S. Cuomo, M. Ranieri, Developing and validating a multidimensional AI literacy questionnaire: Operationalizing AI literacy for higher education, in: Proceedings of the First International Workshop on High-Performance Artificial Intelligence Systems in Education, AIxEDU 2023, Aachen, 2023. URL: https://ceur-ws.org/Vol-3605/.

[19] D. Cetindamar, K. Kitto, M. Wu, Y. Zhang, B. Abedin, S. Knight, Explicating AI literacy of employees at digital workplaces, IEEE Transactions on Engineering Management 71 (2024) 810–823. doi:10.1109/TEM.2021.3138503.

[20] S.-C. Kong, W. M.-Y. Cheung, G. Zhang, Evaluating an Artificial Intelligence Literacy Programme for Developing University Students' Conceptual Understanding, Literacy, Empowerment and Ethical Awareness, Educational Technology & Society 26 (2023) 16–30.

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

[22] P. Weber, M. Pinski, L. Baum, Toward an Objective Measurement of AI Literacy, in: PACIS 2023 Proceedings, 2023. URL: https://aisel.aisnet.org/pacis2023/60.

[23] UNESCO, Report "Guidance for generative AI in education and research", 2023. URL: https://www.unesco.org/en/articles/guidance-generative-ai-education-and-research.

[24] A. Gerdes, A participatory data-centric approach to AI ethics by design, Applied Artificial Intelligence 36 (2022). doi:10.1080/08839514.2021.2009222.

[25] C. Jang, Coping with vulnerability: The effect of trust in ai and privacy-protective behaviour on the use of ai-based services, Behaviour & Information Technology (2023). doi:10.1080/0144929X.2023.2246590.

[26] A. Majeed, S. O. Hwang, When AI Meets Information Privacy: The Adversarial Role of AI in Data Sharing Scenario, IEEE Access 11 (2023) 76177–76195. doi:10.1109/ACCESS.2023.3297646.

[27] P. Samuelson, Generative AI meets copyright, Science 381 (2023) 158–161. doi:10.1126/science.adi0656.

[28] M. A. Yeo, Academic integrity in the age of Artificial Intelligence (AI) authoring apps, TESOL Journal 14 (2023) e716. doi:10.1002/tesj.716.

[29] V. van Oijen, AI-generated text detectors: Do they work?, 2023. URL: https://communities.surf.nl/en/ai-in-education/article/ai-generated-text-detectors-do-they-work.

[30] D. Weber-Wulff, A. Anohina-Naumeca, S. Bjelobaba, T. Foltýnek, J. Guerrero-Dib, O. Popoola, P. Šigut, L. Waddington, Testing of detection tools for AI-generated text, International Journal for Educational Integrity 19 (2023) 26. doi:10.1007/s40979-023-00146-z.

[31] F. Miao, W. Holmes, H. Ronghuai, Z. Hui, AI and education: Guidance for policy-makers, Technical Report, UNESCO, 2023. URL: https://unesdoc.unesco.org/ark:/48223/pf0000376709.

[32] E. Sabzalieva, A. Valentini, ChatGPT and artificial intelligence in higher education: Quick start guide, Technical Report, UNESCO, 2023. URL: https://unesdoc.unesco.org/ark:/48223/pf0000385146.

[33] M. Webb, A Generative AI Primer, Technical Report, National Centre for AI, 2023. URL: https://nationalcentreforai.jiscinvolve.org/wp/2024/01/02/generative-ai-primer/.

[34] Russell Group, New principles on use of AI in education, 2023. URL: https://russellgroup.ac.uk/news/new-principles-on-use-of-ai-in-education/.

[35] GTnum, Intelligence artificielle et éducation: Apports de la recherche et enjeux pour les politiques publiques, 2023. URL: https://edunumrech.hypotheses.org/8726.

[36] M. A. Cardona, R. J. Rodríguez, K. Ishmael, Artificial Intelligence and the Future of Teaching and Learning: Insights and Recommendations, Technical Report, 2023. URL: https://policycommons.net/artifacts/3854312/ai-report/4660267/.

[37] UCL, Using generative AI (GenAI) in learning and teaching, 2023. URL: https://www.ucl.ac.uk/teaching-learning/publications/2023/sep/using-generative-ai-genai-learning-and-teaching.

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

[39] P. P. Martin, D. Kranz, P. Wulff, N. Graulich, Exploring new depths: Applying machine learning for the analysis of student argumentation in chemistry, Journal of Research in Science Teaching (2023). doi:10.1002/tea.21903.

[40] E. Kasneci, K. Seßler, S. Küchemann, M. Bannert, D. Dementieva, F. Fischer, et al., ChatGPT for good? On opportunities and challenges of large language models for education, Learning and Individual Differences 103 (2023) 102274.

[41] A. Lepage, N. Roy, A review of the literature from 1970 to 2022 on the roles of teachers and artificial intelligence in the field of AI in education, Médiations et Médiatisations 16 (2023) 30–50. doi:10.52358/mm.vi16.304.

[42] A. Tamkin, M. Brundage, J. Clark, D. Ganguli, Understanding the Capabilities, Limitations, and Societal Impact of Large Language Models, arXiv preprint arXiv:2102.02503 (2021). URL: http://arxiv.org/abs/2102.02503.

[43] O. Koraishi, Teaching English in the Age of AI: Embracing ChatGPT to Optimize EFL Materials and Assessment, Language Education and Technology 3 (2023) Article 1.

[44] F. Ali, D. Choy, S. Divaharan, H. Y. Tay, W. Chen, Supporting self-directed learning and self-assessment using TeacherGAIA, a generative AI chatbot application: Learning approaches and prompt engineering, Learning: Research and Practice 9 (2023) 135–147. doi:10.1080/23735082.2023.2258886.

[45] F. Ouyang, T. A. Dinh, W. Xu, A Systematic Review of AI-Driven Educational Assessment in STEM Education, Journal for STEM Education Research 6 (2023) 408–426. doi:10.1007/s41979-023-00112-x.

[46] J. Biggs, Enhancing teaching through constructive alignment, Higher Education 32 (1996) 347–364. doi:10.1007/BF00138871.