Are Large Language Models Capable of Assessing Students' Written Products?
A Pilot Study in Higher Education

Daniele Agostini
University of Trento

Keywords: Large Language Models (LLMs), AI-Assisted Assessment, Technology-Enhanced Assessment, Artificial Intelligence in Education, Assessment Rubrics, Higher Education, Student Assessment, Authentic Tasks, Academic Assessment, Educational Technology

Introduction

Since the public release of ChatGPT on November 30, 2022, and, consequently, that of all its competitors, the use of Large Language Models (LLMs) has spread by the public. One of the areas in which their use has had the most significant impact from the outset is that of Education and Instruction (Baytak, 2023; Elbanna & Armstrong, 2023; Extance, 2023; Roy et al., 2023; Saif et al., 2023; Tiwari et al., 2023). The use that has been made of them from the beginning in the context of higher education, both by teachers and students (Perkins, 2023; Roy et al., 2023; Sullivan et al., 2023) is of particular interest for the subject of this paper.

The speed with which Large Language Models (LLMs) have been integrated into the fabric of higher education, both at the level of teachers and students, raises fundamental questions about their effectiveness and reliability. In this context, LLMs promise to revolutionise how teachers interact with students, manage workload, and personalise the learning experience (Elbanna & Armstrong, 2023).

Although the potential of this technology for advancements in accessibility and personalisation of learning is acknowledged, a crucial question arises concerning their capability to objectively and impartially assess student performance. The application of such advancements in learning assessment is relatively unexplored, which has significant implications for educational practice and pedagogical theory. This study explores the use of leading LLMs in the specific context of assessing student written products, focusing on their accuracy and ability to evaluate according to a rubric developed by the teacher.

Context

Since the launch of ChatGPT, which occurred more than a year ago, and following the release of competing models, Large Language Models (LLMs) have begun to play a significant role in the technological landscape. Although LLMs had existed for some time, their impact remained relatively limited until the introduction of simple and intuitive user interfaces, such as the "chat" layer, which made these tools more accessible to the general public. This democratisation has catalysed the commercial and general use of LLMs, with a consequent increase in investments in this sector by institutions, companies, and individuals (Babina et al., 2023; Bloomberg, 2023; Hammond, 2023; Lee et al., 2022).

OpenAI ChatGPT, Anthropic Claude, Microsoft Copilot, and Google Bard 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 significant boost. At the same time, there has been a crisis in search engines due to LLMs offering new modes of querying and analysing knowledge and data, a more natural interaction, and quite precise and exhaustive answers without requiring advanced search skills. LLMs enable users to skip over the various inconvenient steps that are typically involved in standard search engine use, such as the need to select from lists of websites, accept cookies, and navigate through advertising banners.

In response to this trend, educational institutions and agencies have begun incorporating LLMs and generative AI into their curricula at varying levels. Specialised courses have been developed to harness the potential of these innovative technologies. There is now an emphasis on AI Literacy, which enables professionals across diverse sectors, including education, to comprehend the fundamental aspects of generative AI, the range of tools at their disposal, their functionality, and effective ways of implementing them in their respective fields (Biagini et al., 2023; Cetindamar et al., 2024; Kong et al., 2023; B. Wang et al., 2023; Weber et al., 2023).

However, this rapid development has also raised critical issues related to information handling. While LLMs offer enormous data analysis and generation potential, concerns arise regarding accuracy, privacy, and ethics in information management and output ownership. These challenges represent a continually evolving field, requiring constant attention and critical evaluation to ensure the responsible use of LLMs (Gerdes, 2022; Jang, 2023; Majeed & Hwang, 2023; Samuelson, 2023).

The initial, brief reaction of higher education institutions was defensive. Some universities have reverted to handwritten exams and oral tests to counter students' possible use of LLMs for completing exam papers (Perkins, 2023; Yeo, 2023). In parallel, the market began to offer software for identifying tasks written by LLMs. Such software has proven to be ineffective and has caused management and legal issues for institutions when students have been wrongly accused of submitting AI-generated texts (van Oijen, 2023; J. Wang et al., 2023; Weber-Wulff et al., 2023).

National and international bodies and university groups have promptly provided guidelines that, while maintaining some attention, have moved in the direction of accepting the use of LLMs in an ethical and effective manner for tasks that can benefit institutions, teachers, and students. Some significant examples are UNESCO (Miao et al., 2023; Sabzalieva & Valentini, 2023), the JISC National Centre for AI (Webb, 2023), the Russell Group (Russell Group, 2023), the French National Ministry of Education (GTnum, 2023), the U.S. Department of Education (Cardona et al., 2023), and University College London (UCL, 2023).

Assessment is an area of great potential benefit, particularly in terms of sustainability. However, caution is necessary as LLMs without specific task adjustments appear unable to manage student assessments independently (Swiecki et al., 2022; Webb, 2023), while LLMs adapted for such tasks demonstrate the ability to produce satisfactory results (Martin et al., 2023). As with students, AI usage carries ethical considerations and responsibilities that teachers must uphold when assessing tasks that could impact students' careers (e.g., personal motivation, grades, scholarships, acceptance into master's or doctoral programs).

Theoretical Framework

The idea of using AI to assist educators in their tasks and to reach precise, unbiased, and informed decisions has been present in much literature since the 1980s (Lepage & Roy, 2023).

The opportunity to use LLMs for learning assessment was also analysed in the era immediately preceding ChatGPT, where, however, transformer models, including OpenAI's GPT-3, were already well-established. Tamkin et al. (2021) emphasised their educational uses, which included:

- Summarising: LLMs can summarise long sections of text. This can help provide concise summaries of lengthy student submissions. The summary can consider various parameters in the text, providing precise information on the aspects the educator wants to evaluate.

- Questioning and Answering: LLMs can "understand" a piece of text, answer questions about it, and ask questions if requested. This can be used to create interactive feedback and learning experiences.
- Classifying: LLMs can classify text into predefined categories. This can be used for assisted assessment or to classify student feedback.
- Plagiarism Detection: By comparing the similarity between different pieces of text, LLMs can help detect potential cases of plagiarism among students and between students and original material.
- Measuring Semantic Similarity: LLMs can measure the semantic similarity between two text parts. This can be used to match student queries with relevant answers or resources and assist the teacher in assessing the student's work.
- Generating Feedback: LLMs can generate personalised feedback based on assessing a student's work. It would work even better if the LLM had the teacher's notes on the task to work on.
- Assessing Knowledge: LLMs can assess a student's understanding of a topic based on their written communications, especially if they are adequately trained on correct tasks and have a grading rubric to refer to.

Each of these seven applications is fundamental for using LLMs in learning assessment.

Following the introduction of ChatGPT and other universally accessible LLMs, UNESCO released the guidelines "AI and education: Guidance for policy-makers" (Miao et al., 2023), suggesting the following actions regarding learning assessment:

1. Test and implement artificial intelligence technologies to support the assessment of various dimensions of skills and outcomes.
2. Exercise caution when adopting automated assessment with closed-ended questions based on rules.
3. Use formative assessment aided by artificial intelligence as an integrated function of Learning Management Systems (LMSs) to analyse student learning data more accurately and efficiently and reduce human bias.
4. Progressive assessments based on artificial intelligence to provide regular updates to teachers, students, and parents.
5. Examine and evaluate the use of facial recognition and other artificial intelligence for user authentication and monitoring in remote online assessments.

Drawing upon these various theoretical approaches, indications, and guidelines, the AI-Mediated Assessment for Academics and Students (AI-MAAS) model has been developed and is currently undergoing validation. The model proposes two potential implementations of LLMs for learning assessment: one for formative assessment and one for summative assessment (Agostini & Picasso, 2023). In either case, the selected LLM must possess the ability to assess according to a grading rubric provided by either the teacher or the students.

However, there have not been many experiences in this regard thus far. Martin et al. (2023) worked on this possibility, starting from the need to assign reasoning, conceptualisation and processing tasks, and the fact that correcting large quantities of open-ended responses and tasks often proves unsustainable. The researchers then demonstrated in a chemistry task that LLMs can be used for this purpose. In this case, a near-perfect match between human scores and the LLM scores was achieved. It should be noted, however, that Martin and colleagues did not limit themselves to using an LLM to achieve this result. They followed the this procedure:

1. Used the unsupervised machine learning technique HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise) to cluster student topics. This helped to uncover patterns in discussions.
2. Mapped the clustering results onto structures guided by the theory of reasoning modes and levels of granularity to create a holistic scoring rubric with 20 categories.
3. Assigned a score to all topics manually based on the rubric to create a labelled dataset.
4. Compared the performance of different pre-trained language models (BERT, RoBERTa, SciBERT) on classifying topics into the 20 rubric categories. BERT large uncased achieved the best results.
5. Trained a deep neural network classifier using the labelled dataset to automate the scoring of new topics based on the rubric. The model achieved an accuracy of 87% on an extended test set.
6. Validated the model using techniques like generating artificial topics, conducting black-box analyses, and computing feature importance scores. This helped ensure that the model relied on keywords similar to human raters.

This excellent result is, therefore, the fruit of models trained on a specific task and population, and it cannot be expected that the procedure applied can be used by any teacher not specialised in Machine Learning. The idea of this study was to present an operational model and demonstrate its feasibility.

Other studies have utilised LLMs for assessment purposes without comparing the AI's evaluation to that of a teacher. These studies have had satisfactory results in assessing English L2 tasks (Koraishi, 2023) and supporting self-assessment (Ali et al., 2023). There are other examples of using Machine Learning in assessing STEM-related tasks, but without using LLMs (Ouyang et al., 2023).

This article presents a pilot study on the capabilities of leading LLMs to assess authentic tasks using a rubric. Identifying an LLM reasonably capable in this task would be a good starting point for pursuing AI-supported assessment experiments.

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 (Biggs, 1996) and thereby improve the quality and effectiveness of university teaching.

Methodology and Tools

This study 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 specialisation course for support teachers in the module concerning the use of new technologies for inclusion. Eighty-eight students participated anonymously, divided into 21 groups, along with 2 evaluating teachers, experts in experimental pedagogy and assessment. No data regarding the students' demographics were collected. The groups were tasked with carrying out an authentic task, namely to design an educational intervention targeted at a specific class (which could range from 1st grade of primary school to 5th grade of secondary school, depending on the group composition) that considered the integration of educational technologies and the inclusion of students with special educational needs. This educational intervention could take place over two or more sessions. 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, Objectives and Practical Implications (outcomes), Context and Environment (formal, informal, type of setting, etc.), Planned Technologies with pros and cons and types of use, Evaluation and Schedule with details. 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 a 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 five levels for each criterion.

Table 1: Rubric for the Assessment of the Educational Intervention

Assessment Criteria	**Absent (0 points)**	**Low Level (1 point)**	**Intermediate Level (2 points)**	**High Level (3 points)**	**Advanced Level (4 points)**
Objectives and Outcomes	No Objectives	Objectives unclear or inconsistent	Objectives are clear but not fully integrated	Objectives are clear and well-integrated	Excellent objectives and perfectly integrated with innovative outcomes
Use and Interaction with Technologies and Generative AI	No Use and interaction with technologies	Limited or inadequate use and interaction, absence of generative AI	Adequate use and interaction, generative AI is present but not fully exploited	Good use and interaction, effective inclusion of generative AI	Excellent and innovative use and interaction, optimal exploitation of generative AI
Design: Consistency with Objectives, Internal Coherence, and Originality	No Design	Incoherent design, not aligned with stated objectives or lacking originality.	Coherent design but with limited originality and innovation	Design both coherent and original/innovative	Remarkably coherent design and highly original/innovative
Inclusivity	No Inclusive thinking	Educational inclusivity is inadequate or ineffective	Inclusive thinking is present but without distinctive elements or underdeveloped	Proactive and well-integrated inclusivity with the design	Pronounced inclusivity, integrated throughout the design and activities of the intervention
Assessment	No Assessment	Unclear or inadequate assessment plans and methods	Fairly adequate assessment plans and methods, but not very clear	Good assessment plans and methods, coherent and with some innovative elements	Excellent and highly innovative assessment plans, consistent with educational objectives

Two expert human evaluators and five LLMs evaluated all the student groups' products. The LLMs selected for this study were the most popular competing models at the time, plus an outsider:

1. **ChatGPT 3.5-turbo by OpenAI**: The improved version of the initial release model of ChatGPT. It is currently the model available for the free version of ChatGPT and remains one of the best models to be used. Link: https://chat.openai.com/
2. **ChatGPT 4 by OpenAI**: This is the evolution of ChatGPT 3.5-turbo, with significantly improved text comprehension and generation capabilities. This model offers more accurate and detailed responses, making it suitable for various applications. This model is only available to customers who pay for the premium plan. Link: https://chat.openai.com/
3. **Claude 2 Chat by Anthropic**: An advanced version of Anthropic's artificial intelligence model. Claude 2 stands out for its ability to understand and respond to complex questions with high accuracy and contextual sensitivity. Indeed, its prominent feature is the ability to process

files and maintain an extensive context, i.e., the ability to "read" and take into account very long prompts and textual files (up to 100,000 tokens, approximately 75,000 words). The use of this model is free but limited to 50 messages per day. The paid version is not available in Italy. Link: https://claude.ai

4. **Bing Chat/Copilot by Microsoft**: An artificial intelligence model integrated with the Bing search engine. Copilot, when the correct option is activated, is based on a GPT-4 model, the same type of model used by ChatGPT 4. It is beneficial for its intrinsic ability to perform web searches. Moreover, it can be used for free. Link: https://www.bing.com/chat

5. **Bard by Google**: Google's AI model is known for its integration with Google's vast database of information. Bard is designed to provide quick and accurate answers to various questions, leveraging the vast knowledge available online. It is a free model but seems to lack the accuracy and reasoning capabilities of the other models presented. Link: https://bard.google.com/chat

6. **OpenChat 3.5 by OpenChat/AlignmentLabs**: An open-source artificial intelligence model with only 7 billion parameters (for reference, ChatGPT 3.5 is estimated to have 175 billion parameters) that can also operate locally on desktop computers. In many benchmarks for LLMs, it reaches the levels of ChatGPT 3.5. Being an open-source model, it is free. Link: https://openchat.team

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). Moreover, most of these models provide privacy solutions, even if, in this case, the tasks were entirely anonymous and free of sensitive data. Both OpenAI models and Copilot offer the option not to save conversations and not use them for model training. Claude 2 by Anthropic does not perform any training on chat data, while Google Bard does not record activity and conversations and allows to manage saved data. OpenChat 3.5 does not save anything in its online version, and, being open-source, it is intended to be used locally on one's own computer; this way, no data will be transmitted over the Internet.

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.

Here are the two prompts that were given to the LLMs to assess the products:

Prompt 1 (originally written in Italian):

> *The following document is a task completed by students. You will assess it after receiving the evaluation rubric in my second prompt. If you understand, respond only with "understood" to this first prompt, without adding anything else.*

> *Note that the following texts are part of the instructions given to students for the various sections of the document and are not produced by them:*
>
> *<start of texts already present as instructions in the document>*
>
> *[here all the texts already present in the template were provided]*
>
> *<end of texts already present in the document as instructions>*
>
> *<start of document produced by students>*
>
> *[here the document produced by the groups was pasted]*
>
> *<end of document produced by students>*

Prompt 2 (originally written in Italian):

> *Assess the educational intervention that I provided in the first prompt, created by students of the specialisation course for support teachers. The key competence of this task was to be able to plan an educational intervention that was inclusive for students and that also used the available technologies in this sense (better if generative AI technologies were used). At the same time, the educational design had to prove effective in achieving the objectives they set for themselves. Students did not have much time, so it was not required to plan in great detail. Use the following assessment rubric to assign scores and then present the assessment in the form of a list of the rubric criteria with the corresponding scores.*
>
> *<start of Learning Unit assessment rubric>*
>
> *Assessment Rubric LU:*
>
> *Assessment Criterion: Objectives and Outcomes*
>
> *- Absent (assign 0 points): No Objectives*
> *- Low Level (assign 1 point): Objectives unclear or inconsistent*
> *- Intermediate Level (assign 2 points): Objectives clear but not fully integrated*
> *- High Level (assign 3 points): Objectives clear and well integrated*
> *- Advanced Level (assign 4 points): Excellent objectives and perfectly integrated with innovative outcomes*
>
> *[... continues with other criteria ...]*
>
> *<end of Learning Unit assessment rubric>*

The first prompt was adapted for Chat GPT-4 and Claude 2, which can accept attached documents, so, for these two models, the first prompt did not contain the text of the group product, but directly the file attached. Copilot does not support text documents attached to prompts, but it can read documents opened with the Microsoft Edge browser, and it was used in this way; in this case, as well, it was not necessary to copy the text of the product in the first prompt.

Finally, 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 educator 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, in order to conduct more comprehensive research. The optimisation of the results will only be considered at a later stage.

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. This is why, in the first prompt, the model was asked to respond with "understood" if everything was clear, aiming to preserve as much context as possible for the final assessment response. Without this precaution, LLMs tended to write long responses that prematurely analysed the product without having the rubric available.

To preserve the context window, some LLMs 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. This is why the instructions were divided into two separate prompts. Below is a table illustrating 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 Chat versions, not the APIs. Note that this feature may change with updates.

Large Language Model (versions available in Italy, November 2023)	**Context Window (in tokens)**
ChatGPT 3.5-turbo	8.192
ChatGPT 4	32.000
Claude 2	100.000
Bing Chat/Copilot	Not declared, approx. 13,500 (maximum prompt 4000 characters)
Google Bard	Not declared, approx. 1024
OpenChat 3.5	8.192

Method of Analysis

The analysis method for evaluating the data involved examining the levels assigned by each evaluator (both LLMs and humans) to the various criteria of the rubric for each of the 21 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 105 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. Software tools like Microsoft Excel and JASP (based on R) were used for the statistical analyses.

Results

After an initial review of all the results, the decision was made to exclude Google Bard from the study before proceeding with further analysis. This LLM consistently assigned high scores (3) 93.3% of the time and intermediate scores (2) for the remaining percentage, demonstrating an inability to effectively evaluate any product, likely due to its currently limited context window.

Another check implemented during the trials with the different LLMs concerned the consistency of assessment. This was done by requesting the assessment of the same product three times from each LLM and starting a new session each time. From these tests, the following behaviours were observed:

- ChatGPT-4 was always consistent. There was no variation in assessment across the three attempts.
- ChatGPT-3.5-turbo was inconsistent. The assessment varied by one or two points for two or more criteria.
- Bing Chat / Copilot was always consistent. There was no variation in assessment across the three attempts.
- Claude 2 was almost always consistent. There was a variation of one criterion by one point.
- OpenChat 3.5 was somewhat inconsistent. There was a variation of two criteria by one point.

Claude 2 proved to be the only model capable of evaluating multiple products simultaneously while maintaining consistency, likely thanks to its wide context window.

Copilot was the only model that, upon being informed that students created the products, advised the user that no data from the current conversation would be saved due to potentially sensitive material.

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 Components Loadings

	RC1	RC2	RC3	Uniqueness
e3	0.687		-0.455	0.347
e4	0.654			0.481
e6	0.652			0.498
e2	0.462			0.684
e0		0.838		0.319
e1		0.698		0.393

Table 3. PCA Components Loadings

	RC1	RC2	RC3	Uniqueness
e7			0.921	0.154

Note. The rotation method applied is promax.

The first component (RC1) is formed by evaluators e2, e3, e4, and e6 loadings, which correspond respectively to the LLMs ChatGPT-4, ChatGPT-3.5, Claude 2, and Copilot. The second component (RC2) comprises those of e0 and e1, corresponding to human evaluators 1 and 2. Finally, the third component (RC3) is mainly constituted by e7, OpenChat 3.5, and the negative loading of ChatGPT-3.5.

As can be appreciated in Fig. 1, ChatGPT-4 contributes less to the RC1 component than the others. Trying to name the identified components, RC1 could be called "Proprietary/High-Parameter LLM Evaluation", RC2 "Human Evaluation", and RC3 "OpenSource Low-Parameter LLM Evaluation". The graph also shows how ChatGPT-4 contributes positively to RC2 with a loading of about 0.3.

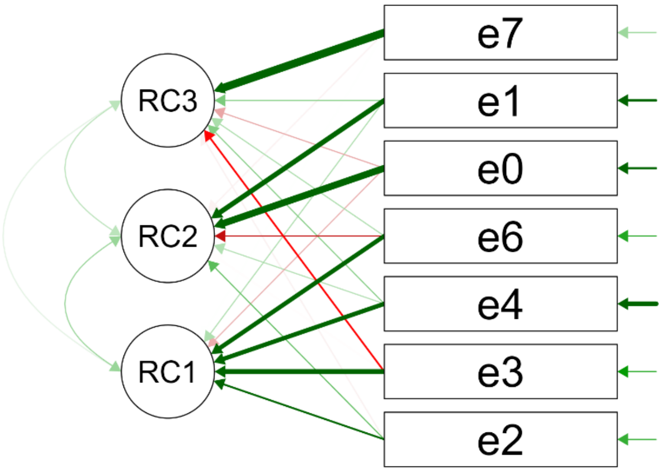


Figure 1. PCA Loading Diagram

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 1 and correspond as follows: Criterion 0: Objectives and Outcomes, 1: Use and Interaction with Technologies and Generative AI, 2: Design: Consistency with Objectives and Outcomes, Internal Consistency and Originality, 3: Inclusivity, 4: Assessment.

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.

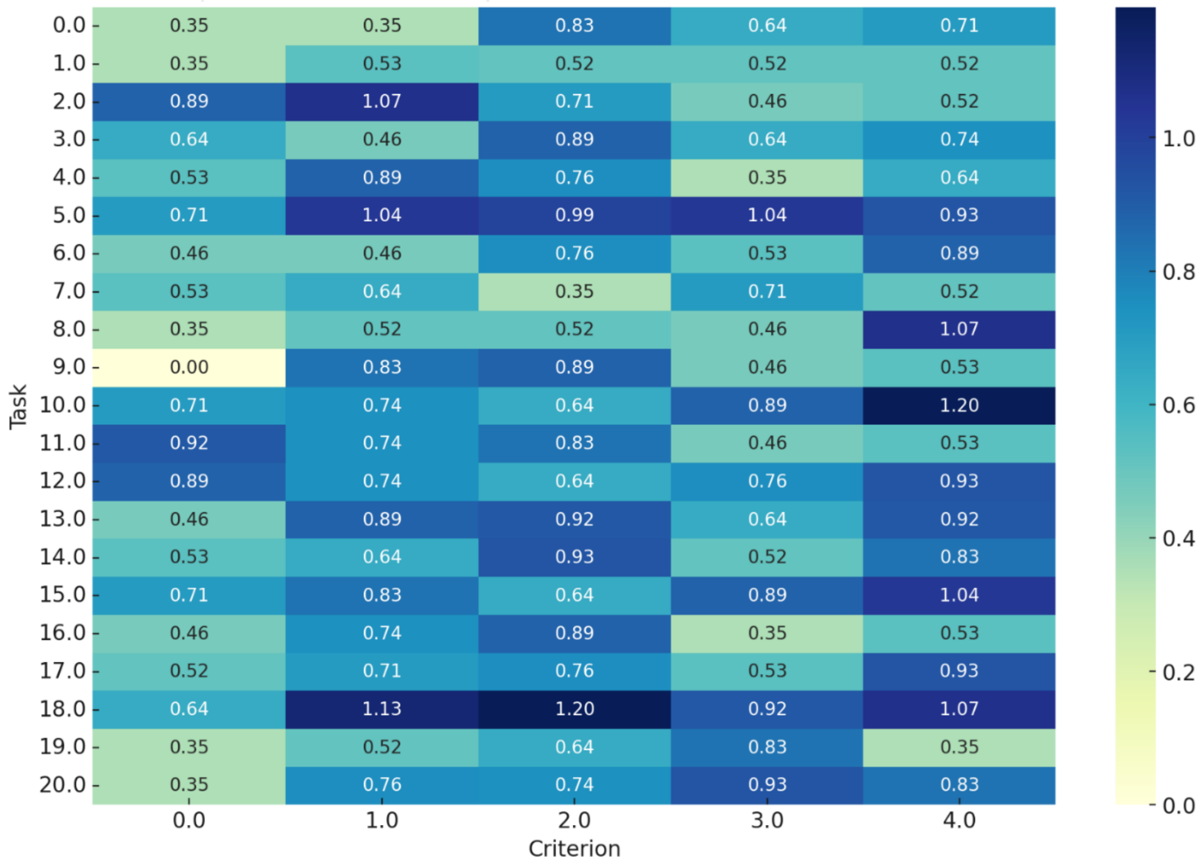
The criterion with the minimum SD across all products is Criterion 0 (Objectives), with an average of about 0.5. This suggests a high level of agreement among evaluators in assessing the quality of the objectives, outcomes and the designed implications. On the other hand, the criterion with the maximum SD among all activities is Criterion 4 (Assessment), with an average of about 0.772. This indicates a higher level of disagreement or inconsistency in how evaluators assessed the correctness and appropriateness of the planned evaluations.

Table 4. Average standard deviation of scores assigned to criteria

Rank	Criterion	Average Standard Deviation (SD)	Average SD % of the score from 0 to 4
1	4 - Assessment	0.772	19.31
2	2 - Design	0.763	19.08
3	1 - Use of Technologies	0.726	18.14
4	3 - Inclusivity	0.644	16.11
5	0 - Objectives	0.541	13.54

In Table 5, it is possible to see how some products, in particular, have caused disagreement among evaluators, especially product number 18. Other products, such as numbers 0 and 1, received more homogeneous evaluations. In other products, like product number 8, it is only a single criterion that is not uniformly assessed.

Table 5. The standard deviation of scores for different Products (Tasks) and Criteria.



Task \ Criterion	0.0	1.0	2.0	3.0	4.0
0.0	0.35	0.35	0.83	0.64	0.71
1.0	0.35	0.53	0.52	0.52	0.52
2.0	0.89	1.07	0.71	0.46	0.52
3.0	0.64	0.46	0.89	0.64	0.74
4.0	0.53	0.89	0.76	0.35	0.64
5.0	0.71	1.04	0.99	1.04	0.93
6.0	0.46	0.46	0.76	0.53	0.89
7.0	0.53	0.64	0.35	0.71	0.52
8.0	0.35	0.52	0.52	0.46	1.07
9.0	0.00	0.83	0.89	0.46	0.53
10.0	0.71	0.74	0.64	0.89	1.20
11.0	0.92	0.74	0.83	0.46	0.53
12.0	0.89	0.74	0.64	0.76	0.93
13.0	0.46	0.89	0.92	0.64	0.92
14.0	0.53	0.64	0.93	0.52	0.83
15.0	0.71	0.83	0.64	0.89	1.04
16.0	0.46	0.74	0.89	0.35	0.53
17.0	0.52	0.71	0.76	0.53	0.93
18.0	0.64	1.13	1.20	0.92	1.07
19.0	0.35	0.52	0.64	0.83	0.35
20.0	0.35	0.76	0.74	0.93	0.83

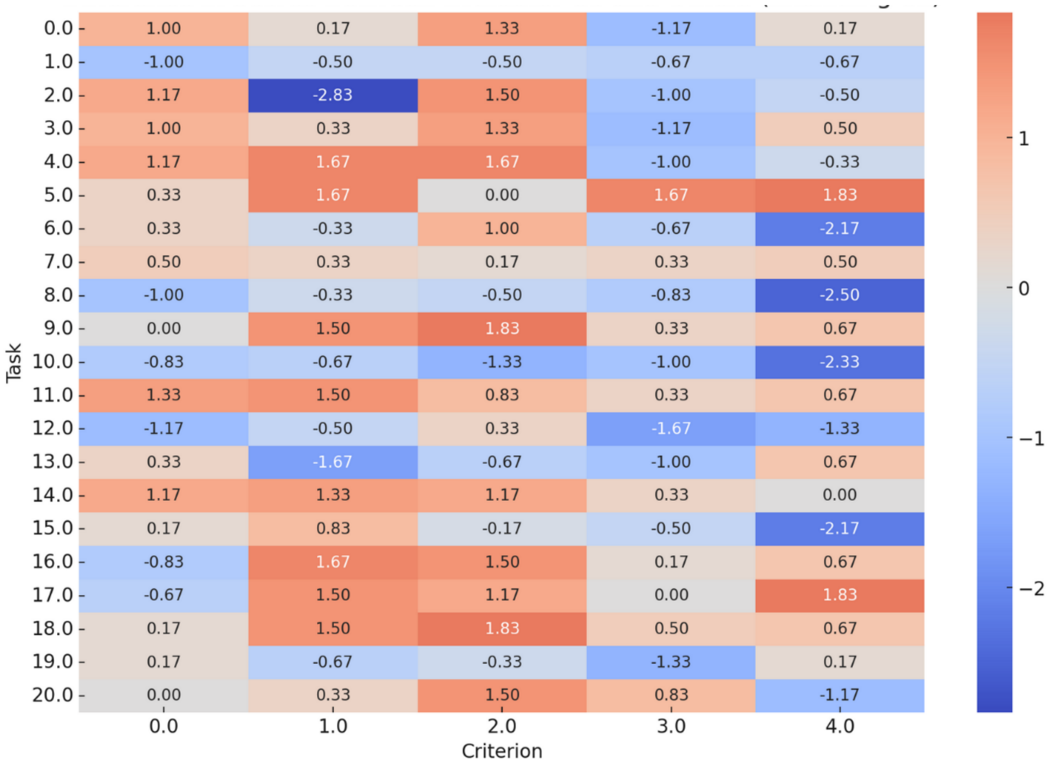
Disagreement Index

For each criterion, the difference between the scores assigned by human evaluators (e0 and e1) was compared, and their difference from the average of the LLM evaluators was calculated—tables 6 and 7 show how the LLMs assigned scores more similar to evaluator e1. Table 8 also shows the comparison results between the scores assigned by the two human evaluators. Between humans there are sometimes even more marked differences than between human and LLM, reaching a three-point difference on more than one occasion. However, it can also be noted that, when analysing criterion by criterion, or even product by product, where there is the most significant disagreement between human evaluators and LLMs, there is a significant agreement among the human evaluators.

For example, let us analyse the product 18 (named "task" in the table). It can be seen how it was problematic in terms of standard deviation (SD)(Table 5) and disagreement (Tables 6 and 7), especially regarding criteria 1 and 2 (Use of technologies and Design). If we look at Table 8, we can see how the assessments of the two human evaluators are in agreement. A "Disagreement Index" (DI) 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 (including the option that the most similar evaluator is the other human evaluator).

It is constructed as follows: Disagreement Index = (Average difference + Variability of the difference) / 2.

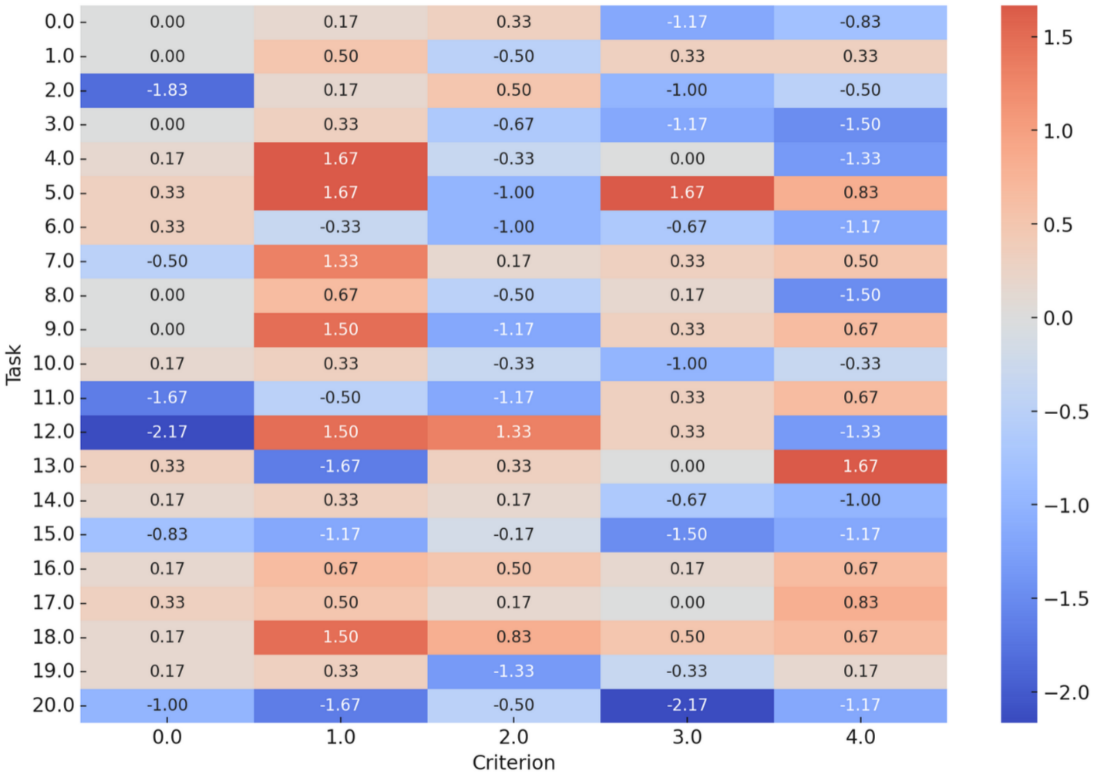
Table 6. The difference in scores assigned between e0 (human evaluator 0) and the average of the LLM evaluators.



Task	0.0	1.0	2.0	3.0	4.0
0.0	1.00	0.17	1.33	-1.17	0.17
1.0	-1.00	-0.50	-0.50	-0.67	-0.67
2.0	1.17	-2.83	1.50	-1.00	-0.50
3.0	1.00	0.33	1.33	-1.17	0.50
4.0	1.17	1.67	1.67	-1.00	-0.33
5.0	0.33	1.67	0.00	1.67	1.83
6.0	0.33	-0.33	1.00	-0.67	-2.17
7.0	0.50	0.33	0.17	0.33	0.50
8.0	-1.00	-0.33	-0.50	-0.83	-2.50
9.0	0.00	1.50	1.83	0.33	0.67
10.0	-0.83	-0.67	-1.33	-1.00	-2.33
11.0	1.33	1.50	0.83	0.33	0.67
12.0	-1.17	-0.50	0.33	-1.67	-1.33
13.0	0.33	-1.67	-0.67	-1.00	0.67
14.0	1.17	1.33	1.17	0.33	0.00
15.0	0.17	0.83	-0.17	-0.50	-2.17
16.0	-0.83	1.67	1.50	0.17	0.67
17.0	-0.67	1.50	1.17	0.00	1.83
18.0	0.17	1.50	1.83	0.50	0.67
19.0	0.17	-0.67	-0.33	-1.33	0.17
20.0	0.00	0.33	1.50	0.83	-1.17

Criterion

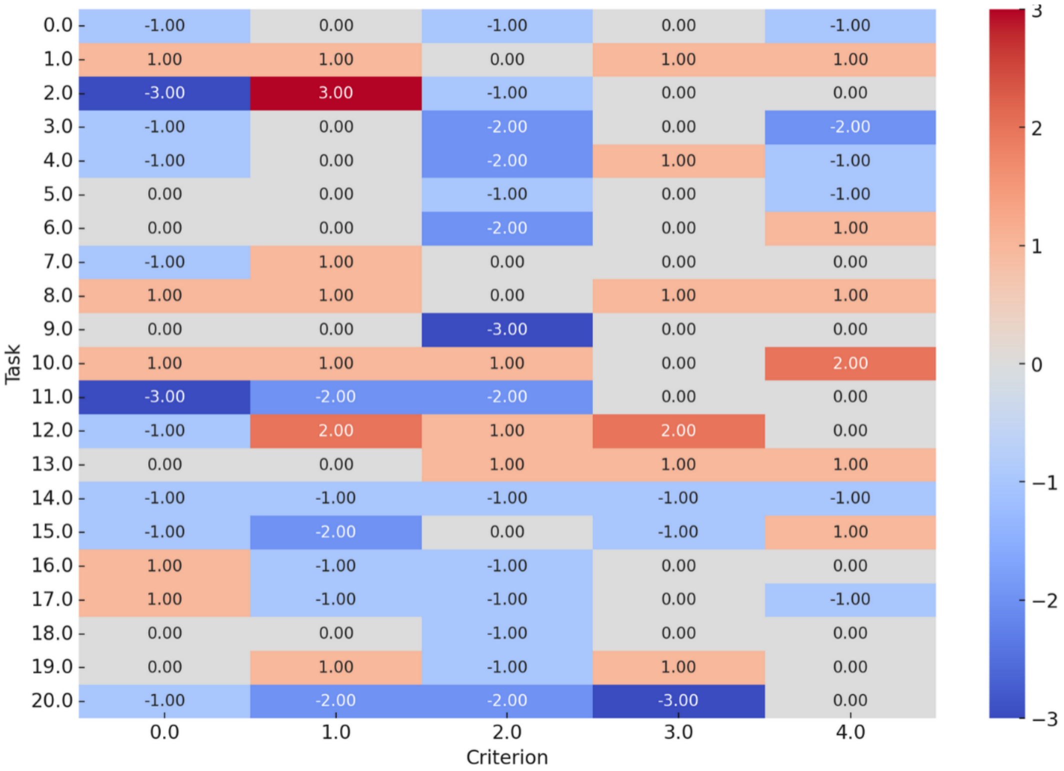
Table 7. The difference in scores assigned between e1 (human evaluator 1) and the average of the LLM evaluators.



Task	0.0	1.0	2.0	3.0	4.0
0.0	0.00	0.17	0.33	-1.17	-0.83
1.0	0.00	0.50	-0.50	0.33	0.33
2.0	-1.83	0.17	0.50	-1.00	-0.50
3.0	0.00	0.33	-0.67	-1.17	-1.50
4.0	0.17	1.67	-0.33	0.00	-1.33
5.0	0.33	1.67	-1.00	1.67	0.83
6.0	0.33	-0.33	-1.00	-0.67	-1.17
7.0	-0.50	1.33	0.17	0.33	0.50
8.0	0.00	0.67	-0.50	0.17	-1.50
9.0	0.00	1.50	-1.17	0.33	0.67
10.0	0.17	0.33	-0.33	-1.00	-0.33
11.0	-1.67	-0.50	-1.17	0.33	0.67
12.0	-2.17	1.50	1.33	0.33	-1.33
13.0	0.33	-1.67	0.33	0.00	1.67
14.0	0.17	0.33	0.17	-0.67	-1.00
15.0	-0.83	-1.17	-0.17	-1.50	-1.17
16.0	0.17	0.67	0.50	0.17	0.67
17.0	0.33	0.50	0.17	0.00	0.83
18.0	0.17	1.50	0.83	0.50	0.67
19.0	0.17	0.33	-1.33	-0.33	0.17
20.0	-1.00	-1.67	-0.50	-2.17	-1.17

Criterion

Table 8. The difference in scores assigned between e0 (human evaluator 0) and e1 (human evaluator 1).



Task	0.0	1.0	2.0	3.0	4.0
0.0	-1.00	0.00	-1.00	0.00	-1.00
1.0	1.00	1.00	0.00	1.00	1.00
2.0	-3.00	3.00	-1.00	0.00	0.00
3.0	-1.00	0.00	-2.00	0.00	-2.00
4.0	-1.00	0.00	-2.00	1.00	-1.00
5.0	0.00	0.00	-1.00	0.00	-1.00
6.0	0.00	0.00	-2.00	0.00	1.00
7.0	-1.00	1.00	0.00	0.00	0.00
8.0	1.00	1.00	0.00	1.00	1.00
9.0	0.00	0.00	-3.00	0.00	0.00
10.0	1.00	1.00	1.00	0.00	2.00
11.0	-3.00	-2.00	-2.00	0.00	0.00
12.0	-1.00	2.00	1.00	2.00	0.00
13.0	0.00	0.00	1.00	1.00	1.00
14.0	-1.00	-1.00	-1.00	-1.00	-1.00
15.0	-1.00	-2.00	0.00	-1.00	1.00
16.0	1.00	-1.00	-1.00	0.00	0.00
17.0	1.00	-1.00	-1.00	0.00	-1.00
18.0	0.00	0.00	-1.00	0.00	0.00
19.0	0.00	1.00	-1.00	1.00	0.00
20.0	-1.00	-2.00	-2.00	-3.00	0.00

Therefore:

- The "Average Difference" is the absolute average difference in scores assigned between the evaluator in question and the reference evaluator (e0 or e1) across all tasks and criteria.
- The "Variability of the Difference" is the standard deviation of the difference scores between the evaluator and the reference evaluator, reflecting how consistent these differences are across different tasks and criteria.

The Disagreement Index 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 higher value indicates a more significant overall disagreement in evaluation relative to the reference evaluator. The highest possible value for the index for an evaluator would be achieved if they constantly evaluated at the maximum difference, for example, from e0 (4 points) with no variation. In this case, the average dissimilarity and the standard deviation would be 4. Thus, the composite index, being the average of these two values, would also be 4.

Evaluators Similar to Human Evaluator e0

It was discovered that the evaluator who provided scores most similar to evaluator e0 was another human evaluator, e1 (see the DI metric in Table 9). The artificial intelligence LLM that agreed the most with e0 was e2, which is ChatGPT-4, followed at a distance by Claude 2 (e4) and OpenChat 3.5. The evaluator that differed the most from the scores assigned by e0 was e6, which is Copilot.

Table 9. Disagreement Indices (DIs) between human evaluator e0 and other evaluators.

Evaluators	Average Difference	Variability of Difference	Standardised Average Difference (Z-score)	Disagreement Index (DI)
Human e1 - Human e0	0.8381	0.8101	-0.8343	-0.0121
ChatGPT-4 (e2) - Human e0	0.8857	0.8004	-0.4230	0.1887
ChatGPT-3.5 (e3) - Human e0	0.9714	0.7527	0.3173	0.5350
Claude 2 (e4) - Human e0	0.9333	0.7754	-0.0118	0.3818
Copilot (e6) - Human e0	1.1619	0.8562	1.9623	1.4093
OpenChat 3.5 (e7) - Human e0	0.9429	0.8184	0.0705	0.4445

Focusing on the single criterion (Table 10), it can be noted how DI with other evaluators varies from criterion to criterion; overall, this confirms what was learned from the general calculation. Interestingly, for the criteria recognised as most critical for evaluation, such as Assessment, Design, and Use of Technologies (Table 4), the evaluators most similar to e0 are the other human evaluator and OpenChat 3.5.

Table 10. Disagreement Indices (DIs) of evaluators compared to human evaluator e0 divided by criterion:

Criteria	DI e1-e0	DI e2-e0	DI e3-e0	DI e4-e0	DI e6-e0	DI e7-e0	Most Similar Evaluator	Second Most Similar
0 - Objectives	0.8679	0.7311	0.5583	0.5749	0.6606	0.8107	ChatGPT-3.5	Claude 2
1 - Use of Technologies	0.8969	0.9841	0.9922	1.0105	1.1209	0.8412	OpenChat 3.5	Human e1
2 - Design	0.9678	0.9183	0.9700	1.0345	1.1464	0.8107	OpenChat 3.5	ChatGPT-4
3 - Inclusivity	0.6910	0.5749	0.8026	0.6374	0.9700	0.7796	ChatGPT-4	Claude 2
4 - Assessment	0.6625	0.9629	0.9262	0.9183	1.0909	1.1342	Human e1	Claude 2

Evaluators Similar to Human Evaluator e1

Upon a quick analysis of Table 7, it was evident that the human evaluator e1 tended to give scores more similar to those of the LLMs than the ones of evaluator e0. This observation was further supported by the data presented in Table 11, which indicated that the evaluator who agreed the most with e1 was an LLM, specifically Claude 2. ChatGPT-4 closely followed, whereas ChatGPT-3.5 lagged behind significantly.

It is interesting to note that the other human evaluator, e0, is the second evaluator most in disagreement with e1. Finally, the most in disagreement is Copilot, as in the previous case.

Table 11. Disagreement Indices (DIs) between human evaluator e1 and other evaluators.

Evaluators	**Average Difference**	**Variability of Difference**	**Standardised Average Difference (Z-score)**	**Disagreement Index (DI)**
Human e0 - Human e1	0.8381	0.8101	0.7328	0.7715
ChatGPT-4 (e2) - Human e1	0.7333	0.6831	-0.5211	0.0810
ChatGPT-3.5 (e3) - Human e1	0.7810	0.7336	0.0489	0.3912
Claude 2 (e4) - Human e1	0.7048	0.7196	-0.8631	-0.0718
Copilot (e6) - Human e1	0.8952	0.8077	1.4168	1.1123
OpenChat 3.5 (e7) - Human e1	0.8286	0.7398	0.6188	0.6793

In the case of evaluator e1 as well, observing individually the criteria (Table 12), it can be noted from the DI for the assessment of the most critical criteria (Evaluation, Design, and Use of Technologies) the evaluators most similar to e1 are the other human evaluator e0 and ChatGPT-4. Also, in this case, OpenChat 3.5 turns out to be the most aligned on the criterion of the use of technologies and is generally quite in agreement across the board.

Table 12. Disagreement Indices (DIs) of evaluators compared to human evaluator e1 divided by criteria:

Criteria	**DI e0-e1**	**DI e2-e1**	**DI e3-e1**	**DI e4-e1**	**DI e6-e1**	**DI e7-e1**	**Most Similar Evaluator**	**Second Most Similar**
0 - Objectives	0.8679	0.6440	0.6589	0.6367	0.5779	0.7797	Copilot	Claude 2
1 - Use of Technologies	0.8969	0.8873	0.8107	0.8366	1.0532	0.7652	OpenChat 3.5	ChatGPT-3.5
2 - Design	0.9678	0.5583	0.6790	0.6238	0.8366	0.7559	ChatGPT-4	Claude 2
3 - Inclusivity	0.6910	0.6985	0.7490	0.7157	0.8969	0.7796	Human e0	ChatGPT-4
4 - Assessment	0.6625	0.7311	0.8786	0.7446	0.8186	0.8536	Human e0	ChatGPT-4

Discussion

Regarding the goal of understanding whether educators can use current Large Language Models (LLMs) without expertise in Machine Learning to assess student-written products, even in the presence of open tasks and questions, using assessment rubrics, the analyses have revealed several interesting elements:

- The breadth of the context window is extremely important for these types of tasks.
- Two of the most potent models tested, ChatGPT-4 and Claude 2, performed very well, assessing students' products quite similarly to human evaluators. In contrast, another equally powerful model, BingChat/Copilot, was the most distant from those of human assessment.
- OpenChat 3.5, an OpenSource model with only 7 billion parameters, provided assessments more similar to those of humans compared to models that are theoretically much more powerful.
- From the PCA, it appears that human evaluators generally have a different pattern of evaluation compared to LLMs. ChatGPT -4 is probably the LLM that, in general, is closest to the human modality.

- The Disagreement Index (DI) points in the same direction, but detailing the differences relative to individual human evaluators brings out additional details: for human evaluator e0, the most concordant evaluator was human evaluator e1, followed by ChatGPT-4. The converse is not valid. The evaluator most in agreement with human evaluator e1 was Claude 2, followed by ChatGPT-4.
- The concordance of human evaluators is seen in specific cases (such as task 18) and, on average, in the assessment of more complex criteria, such as the quality of the planned assessment and the correct use and implementation of technologies.

Based on the available data, it appears that OpenAI's ChatGPT-4 and Anthropic's Claude 2 are the most suitable models for assisting university educators in evaluating students' written products in the presence of an assessment rubric. This is because they are both closely aligned with the assessments of the two expert human evaluators.

OpenChat 3.5 deserves a mention because, despite being a 7 billion parameter model, thus capable of running locally on most people's desktop computers, it adhered more closely to human assessment compared to much larger models like Copilot, ChatGPT-3.5, and Google Bard (which was excluded from the analysis).

Another criterion to remember is the impact of assessment differences among various evaluators. In an assessment rubric like the one proposed, with evaluations translating into scores from 0 to 4, moving by one point at a time, standard deviation (SD) has a different impact depending on its value: for Criterion 0 (Objectives), for example, with an SD of about 0.54, the effect is not particularly significant. This is because the SD is less than 1, implying that most evaluations are clustered around the average score and rarely move the level by a whole point. For Criterion 4 (Evaluation), the effect is more significant, with an SD of about 0.77. This level of SD indicates that the evaluations are more distributed and can potentially move the detected learning level by a whole point or more.

Considering this, by looking at Tables 10 and 12, it can be understood that, at the moment, none of the LLMs can be used for autonomous evaluation for all criteria, especially regarding the more complex ones (Webb, 2023). However, ChatGPT-4, Claude 2, and, with some caution, even OpenChat 3.5 have the potential to be used as support for evaluation, both for summative and formative evaluation levels (in the latter case, interacting mainly with students) as described in the AI-MAAS model (Agostini & Picasso, 2023).

Conclusions

The fundamental question of this study was whether and which current Large Language Models (LLMs) can be used by university educators, even those without technical experience, to assess student-written 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.

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. This finding confirms the latest guidelines as stated by Miao et al. (2023) and Webb (2023).

However, not all models examined are advisable to assist teachers in assessment. Some lack the "capabilities" (or training) to perform such a complex task (such as Copilot); others, like Google Bard, do not have a context window broad enough to assess elaborate texts. On the practical side, the ability of Claude 2 to assess several texts at once is very beneficial.

This study has allowed us to sift through the examined LLMs and select two for more extensive and in-depth analyses: ChatGPT-4 and Claude 2. To select a third and possibly a fourth model, in addition to the two mentioned, it could be interesting to conduct a similar experiment only with open-source models, given the good performance of OpenChat 3.5 compared to much larger models. The open-source LLM community is very active and, during 2023, has been able to raise the quality level of the models exponentially, also thanks to companies and institutions like Meta (which released LLAMA and LLAMA2), TII (Technology Innovation Institute of Abu Dhabi, which released the Falcon models), Mistral (with the Mistral model), and HuggingFace (a platform for the development and sharing of LLMs) which have created an ecosystem supporting these efforts and that has proved to be very effective.

Once a few models are selected, a subsequent study can proceed using a multi-shot prompting type, thus presenting the LLMs with examples of optimally performed assessments (exemplars) before moving on to the assessment task. Textual feedback to tasks is one of LLMs' most disruptive capabilities, and those that could be provided during the assessment and following the criteria provided in a rubric deserve particular exploration (Agostini & Picasso, 2023; Sabzalieva & Valentini, 2023; Sullivan et al., 2023; Tamkin et al., 2021).

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 to be more precise. However, it is not considered necessary to overly simplify the criteria. That is to avoid the error of evaluating LLMs on bespoke rubrics, which would not be those that a teacher would create in the day-to-day work. This study also did not consider all ethical, data protection, and legislative implications and issues. In those matters, too, the landscape is rapidly evolving.

The rapid pace of updates to LLMs within their platforms must be highlighted: model speed, context window breadth, and capabilities can, therefore, change in a short time, and some conclusions of this study may soon need to be updated as well.

References

Agostini, D., & Picasso, F. (2023, November 6). Large Language Models for Sustainable Assessment and Feedback in Higher Education: Towards a Pedagogical and Technological Framework. *Proceedings of the First International Workshop on High-Performance Artificial Intelligence Systems in Education Co-Located with 22nd International Conference of the Italian Association for Artificial Intelligence (AIxIA 2023)*. AIxEDU 2023 High-performance Artificial Intelligence Systems in Education, Aachen. https://ceur-ws.org/Vol-3605/

Ali, F., Choy, D., Divaharan, S., Tay, H. Y., & Chen, W. (2023). Supporting Self-Directed Learning and Self-Assessment Using TeacherGAIA, a Generative AI Chatbot Application: Learning Approaches and Prompt Engineering. *Learning: Research and Practice*, *9*(2), 135–147. https://doi.org/10.1080/23735082.2023.2258886

Babina, T., Fedyk, A., He, A. X., & Hodson, J. (2023). *Firm Investments in Artificial Intelligence Technologies and Changes in Workforce Composition* (Working Paper 31325). National Bureau of Economic Research. https://doi.org/10.3386/w31325

Baytak, A. (2023). The Acceptance and Diffusion of Generative Artificial Intelligence in Education: A Literature Review. *Current Perspectives in Educational Research*, *6*(1), Article 1. https://doi.org/10.46303/cuper.2023.2

Biagini, G., Cuomo, S., & Ranieri, M. (2023, November 6). Developing and Validating a Multidimensional AI Literacy Questionnaire: Operationalizing AI Literacy for Higher Education. *Proceedings of the First International Workshop on High-Performance Artificial Intelligence Systems in Education Co-Located with 22nd International Conference of the Italian Association for Artificial Intelligence (AIxIA 2023)*. AIxEDU 2023 High-performance Artificial Intelligence Systems in Education, Aachen. https://ceur-ws.org/Vol-3605/

Biggs, J. (1996). Enhancing teaching through constructive alignment. *Higher Education*, *32*(3), 347–364. https://doi.org/10.1007/BF00138871

Cetindamar, D., Kitto, K., Wu, M., Zhang, Y., Abedin, B., & Knight, S. (2024). Explicating AI Literacy of Employees at Digital Workplaces. *IEEE Transactions on Engineering Management*, *71*, 810–823. https://doi.org/10.1109/TEM.2021.3138503

Elbanna, S., & Armstrong, L. (2023). Exploring the integration of ChatGPT in education: Adapting for the future. *Management & Sustainability: An Arab Review*, *3*(1), 16–29. https://doi.org/10.1108/MSAR-03-2023-0016

Extance, A. (2023). ChatGPT has entered the classroom: How LLMs could transform education. *Nature*, *623*(7987), 474–477. https://doi.org/10.1038/d41586-023-03507-3

Generative AI to Become a $1.3 Trillion Market by 2032, Research Finds | Press | Bloomberg LP. (2023, June 1). *Bloomberg L.P.* https://www.bloomberg.com/company/press/generative-ai-to-become-a-1-3-trillion-market-by-2032-research-finds/

Gerdes, A. (2022). A participatory data-centric approach to AI Ethics by Design. *Applied Artificial Intelligence*, *36*(1). https://doi.org/10.1080/08839514.2021.2009222

GTnum. (2023, April). Intelligence artificielle et éducation: Apports de la recherche et enjeux pour les politiques publiques. *Carnet Hypothèses 'Éducation, numérique et recherche'*. https://edunumrech.hypotheses.org/8726

Hammond, G. (2023, December 27). *Big Tech outspends venture capital firms in AI investment frenzy*. https://www.ft.com/content/c6b47d24-b435-4f41-b197-2d826cce9532

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

Kong, S.-C., Cheung, W. M.-Y., & Zhang, G. (2023). Evaluating an Artificial Intelligence Literacy Programme for Developing University Students' Conceptual Understanding, Literacy, Empowerment and Ethical Awareness. *Educational Technology & Society*, *26*(1), 16–30.

Koraishi, O. (2023). Teaching English in the Age of AI: Embracing ChatGPT to Optimize EFL Materials and Assessment. *Language Education and Technology*, *3*(1), Article 1. https://langedutech.com/letjournal/index.php/let/article/view/48

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

Lepage, A., & Roy, N. (2023). 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*, 30–50. https://doi.org/10.52358/mm.vi16.304

Majeed, A., & Hwang, S. O. (2023). When AI Meets Information Privacy: The Adversarial Role of AI in Data Sharing Scenario | IEEE Journals & Magazine | IEEE Xplore. *IEEE Access*, *11*, 76177–76195. https://doi.org/10.1109/ACCESS.2023.3297646

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

Miao, F., Holmes, W., Ronghuai, H., & Hui, Z. (2023). *AI and education: Guidance for policy-makers*. UNESCO. https://unesdoc.unesco.org/ark:/48223/pf0000376709?posInSet=1&queryId=dc9add31-5176-42f3-9537-4819566551e9

Miguel A. Cardona, E. D., Rodríguez, R. J., & Ishmael, K. (2023). *Artificial Intelligence and the Future of Teaching and Learning: Insights and Recommendations*. https://policycommons.net/artifacts/3854312/ai-report/4660267/

Ouyang, F., Dinh, T. A., & Xu, W. (2023). A Systematic Review of AI-Driven Educational Assessment in STEM Education. *Journal for STEM Education Research*, *6*(3), 408–426. https://doi.org/10.1007/s41979-023-00112-x

Perkins, M. (2023). Academic Integrity Considerations of AI Large Language Models in the Post-Pandemic Era: ChatGPT and Beyond. *Journal of University Teaching and Learning Practice*, *20*(2). https://eric.ed.gov/?id=EJ1382355

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

Russell Group. (2023). *New principles on use of AI in education*. Russell Group. https://russellgroup.ac.uk/news/new-principles-on-use-of-ai-in-education/

Sabzalieva, E., & Valentini, A. (2023). *ChatGPT and artificial intelligence in higher education: Quick start guide*. UNESCO. https://unesdoc.unesco.org/ark:/48223/pf0000385146

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

Samuelson, P. (2023). Generative AI meets copyright. *Science (New York, N.Y.)*, *381*(6654), 158–161. https://doi.org/10.1126/science.adi0656

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

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

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

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

UCL. (2023, September 12). *Using generative AI (GenAI) in learning and teaching*. Teaching & Learning. https://www.ucl.ac.uk/teaching-learning/publications/2023/sep/using-generative-ai-genai-learning-and-teaching

van Oijen, V. (2023, March 31). AI-generated text detectors: Do they work? | SURF Communities. *Surf Communities*. https://communities.surf.nl/en/ai-in-education/article/ai-generated-text-detectors-do-they-work

Wang, B., Rau, P.-L. P., & Yuan, T. (2023). Measuring user competence in using artificial intelligence: Validity and reliability of artificial intelligence literacy scale. *Behaviour & Information Technology*, *42*(9), 1324–1337. https://doi.org/10.1080/0144929X.2022.2072768

Wang, J., Liu, S., Xie, X., & Li, Y. (2023). *Evaluating AIGC Detectors on Code Content* (arXiv:2304.05193). arXiv. https://doi.org/10.48550/arXiv.2304.05193

Webb, M. (2023). A Generative AI Primer. *National Centre for AI*. https://nationalcentreforai.jiscinvolve.org/wp/2024/01/02/generative-ai-primer/

Weber, P., Pinski, M., & Baum, L. (2023). Toward an Objective Measurement of AI Literacy. *PACIS 2023 Proceedings*. https://aisel.aisnet.org/pacis2023/60

Weber-Wulff, D., Anohina-Naumeca, A., Bjelobaba, S., Foltýnek, T., Guerrero-Dib, J., Popoola, O., Šigut, P., & Waddington, L. (2023). Testing of detection tools for AI-generated text. *International Journal for Educational Integrity*, *19*(1), 26. https://doi.org/10.1007/s40979-023-00146-z

Yeo, M. A. (2023). Academic integrity in the age of Artificial Intelligence (AI) authoring apps. *TESOL Journal*, *14*(3), e716. https://doi.org/10.1002/tesj.716