An AI Score to Objectively Assess the Performance of Educational Chatbots

Miguël Dhyne[1], Jean-Roch Meurisse[2], Laurence Dumortier[2], Michaël Lobet[1,*]

1 *Department of Physics and IRDENa, University of Namur, Rue de Bruxelles 51, 5000 Namur, Belgium*
2 *FaSEF and IRDENa, University of Namur, Rue de Bruxelles 51, 5000 Namur, Belgium*
** corresponding author: michael.lobet@unamur.be*

❑ **Abstract**

The rapid integration of AI chatbots into education has created a need for objective methods to evaluate their pedagogical performance. This study introduces an AI Score, a composite metric designed to benchmark educational chatbots across four criteria: their initial performance, their robustness, their self-correction ability and their lack of reliability. The AI Score is calculated using a weighted formula and validated through a standardized test comprising highly discriminant multiple-choice questions. To validate the test, six platforms—ChatGPT, Copilot Studio, NotebookLM, Grok, Mistral, and ClaudeAI—were evaluated under identical conditions using Retrieval-Augmented Generation and class-specific resources. Results demonstrate the AI Score's ability to differentiate chatbot performance. The methodology aligns with ISO/IEC standards for AI reliability and governance, offering educators a reproducible framework for pre-deployment assessment. Limitations and future directions, including longitudinal studies, qualitative evaluation of answer quality, and adaptation to other domains, are further discussed.

Keywords

Generative Artificial Intelligence, AI tutor, educational chatbots, AI benchmark, science of teaching and learning

❑ Introduction

After the burst of ChatGPT in November 2022, the development and use of chatbots in educational contexts has expanded rapidly in recent years. Several recent studies highlight the potential of chatbots and large language models (LLM) to personalize learning, foster engagement, provide tailored explanations, and diversify access to content—while also emphasizing the need to develop critical thinking skills and verification strategies [Debets et al., 2025 ; Kooli, 2023 ; Dempere et al., 2023]. Such findings are supported by international analyses [Kasneci et al., 2023] and feedback from higher education contexts [Marchal et al., 2024].

By virtue of their interactivity and constant availability, these AI tools meet the expectations of flexible, personalized, and on-demand support in the learning process [Kasneci et al. 2023 ; Lee, 2024]. Therefore, bringing AI tools into the classroom has become a specific target market for big tech compagnies such as OpenAI, Microsoft or Google [Extance, 2023] to name a few. Moreover, the accessibility of these AI tools enables any teacher to create its own educational chatbot according to their specific topic and education level [Adiguzel et al., 2023 ; Elkot et al., 2025 ; Elnaffar et al., 2025 Labadze, 2023 ; Dhyne et al., 2024]. Nevertheless, the multiplicity of platforms makes it challenging for educators to select the most pedagogically suitable option.

A global survey of 23,218 students across 109 countries found that 85% had used ChatGPT, with 60% reporting regular use for academic tasks such as essay writing, concept explanation, and exam preparation [Kasneci et al., 2023]. In the United States, 60% of students used ChatGPT on more than half of their assignments, indicating a deep integration into academic workflows [Essel et al., 2022]. Numbers of July 2025 estimated ChatGPT had more than 700 million total weekly active users with education being a major use case [Chatterji et al., 2025]. 10.2% of all user messages and 36% of practical guidance messages are requests for tutoring or teaching. While tools like Microsoft Copilot Studio are gaining traction, particularly in computer science education, where 45% of students use it for code generation, other platforms such as Grok, Mistral, or NotebookLM lack published data on educational adoption [Jin et al., 2025; Cabezas-Clavijo, 2025]. This disparity underscores the need for comparative evaluations: current usage is dominated by a few platforms, yet pedagogical suitability remains largely unmeasured. As AI tools evolve rapidly, longitudinal studies are needed to track not only adoption but also the impact on learning outcomes, equity, and critical thinking; dimensions that remain underexplored in existing literature [Lee, 2024; Extance, 2023].

Educational chatbots offer tangible benefits for both teachers and students. For teachers, these tools save significant time through the automation of repetitive administrative tasks (grading assignments, managing frequently asked questions, preparing teaching materials) and provide more personalized feedback for students [Elkot et al., 2025]. A study by Dhyne and Plumat (2024) illustrates an innovative pedagogical use: GPT-based chatbots were deployed to evaluate and improve lesson preparations for preservice science teachers, leveraging the CDR model (Contextualization, Decontextualization, Recontextualization) and Bloom's taxonomy [Bloom, 1956 ; Anderson et al., 2001]. These tools helped identify imbalances in the cognitive levels targeted and improved the pedagogical coherence of lesson sequences, while freeing up time for qualitative exchanges between trainers and trainees.

For students, chatbots provide personalized support (tailored explanations, immediate feedback) and help reduce anxiety related to learning obstacles or the fear of asking questions in public [Elkot et al., 2025]. However, these benefits come with major risks. A study by Cabezas-Clavijo (2025) revealed that only 26.5% of the bibliographic references generated by eight chatbots (including ChatGPT, Grok, and DeepSeek) were fully accurate, while 39.8% were erroneous or fabricated, highlighting a critical issue of reliability and hallucinations. Additionally, ethical concerns, such as privacy protection (data collection and use of student information) and

algorithmic biases (reinforcement of stereotypes in responses), remain central, particularly in contexts where resources to oversee these tools are limited [Jin et al., 2025].

Furthermore, the field of AI is evolving fast; therefore, an AI conversational agent used as an educational chatbot can evolve with time and versioning, sometimes with regressions [Labadze, 2023 ; Cabezas-Clavijo, 2025]. Additionally, the metaprompt implemented inside the educational chatbot can have an impact on the output received by the students [Davar et al., 2025]. The widespread adoption of these technologies also raises questions about technological dependence and superficial learning. As noted by Elnaffar et al. (2025), uncritical use of chatbots may lead students to favor "ready-made" answers over critical thinking and independent problem-solving, especially in demanding fields. To mitigate these risks, architectures such as Retrieval-Augmented Generation (RAG) show promise. Swacha & Gracel (2025) demonstrate that RAG, by grounding responses in verified external sources, significantly reduces hallucinations and improves the traceability of information, a crucial advantage for educational uses where precision and transparency are essential.

Given these considerations, a fair comparison between different available educational chatbot is essential to highlight their strengths and weaknesses. Therefore, we propose to define an AI score based on four criteria: initial accuracy, robustness, self-correction ability, and lack of reliability. We derive a strict methodology to be able to derive such AI score to differentiate the performance of educational chatbots using RAG on a strictly defined class material [Swacha & Gracel, 2025; Jin et al., 2025] and then apply it to 6 cases.

The selection of chatbots was guided by practical and strategic considerations relevant to the academic context. An OpenAI assistant [OpenAI, 2025] was included as ChatGPT is the most widely used AI among students, making it essential for understanding current usage patterns. Copilot Studio [Microsoft, 2025] was selected due to the university's existing Microsoft licensing agreement, which provides students with free access to this tool. Le Chat [Mistral AI, 2025] represents a European alternative to US-dominated platforms, offering a relevant perspective on data sovereignty and regional AI development. Grok [xAI, 2025] was included to examine emerging alternatives in the conversational AI landscape. Finally, we incorporated two specialized educational tools: NotebookLM [Google, 2025], using Gemini, designed for document analysis and knowledge synthesis, and Amanote [Fery, 2017], an innovative tool for synchronized note-taking with digital materials like Moodle, powered by Claude 3.7 Sonnet [Anthropic, 2025]. This selection enables a comparison of diverse approaches, from generalist models to targeted solutions, and assesses their suitability for specific pedagogical needs.

Results of AI score to those 6 cases are discussed. Limitations of AI score as well as educational perspectives are then discussed.

❑ Definition of an AI score

As explained above, there is a need for comparison method to assess the pedagogical performance of an AI conversational agent acting as educational chatbot, during the testing phase before releasing it to students.

Therefore, the AI score evaluates several criteria which are essential to discriminate between different educational chatbots. These criteria are

- The **Initial Performance (IP),** defined as the ability to provide a correct answer on the first attempt, following an initial prompt. This aspect is of prime importance because the initial answer provided by an "AI tutor" is what students will initially see. Most students may not ask for a follow-up question to check the validity of the answer, as one of the key advantages of AI tutor is to provide a fast answer to their concern. Therefore, the higher the IP, the better the AI score and it is the key criterion evaluated in the AI score.
- The **Robustness (R)**, defined as the ability to maintain a correct answer despite external questioning or follow-up prompt. Students may doubt the provided answers either because they did not fully understand it as they are still in a learning process or because the AI-generated response may be incorrect since AI tools are prone to hallucinations. A robust AI tutor will achieve a higher AI score. However, since this criterion requires a second step from the student, its weight is set lower than IP.
- **Self-correction ability (SCA)** is defined as the ability to revise an initially incorrect answer when challenged and subsequently converge toward the correct answer in a second step. Indeed, since AI tools may hallucinate, a careful student might challenge the AI tutor. If the chatbot can self-correct, it's beneficial for the AI score but again, as it requires a second step from the student, its weight is lower than IP.
- **Lack of reliability (LR)**, defined as the tendency of an AI tutor to alter its response without logical justification. As AI tools are inherently probabilistic, one should penalize the lack of reliability, i.e. changing an answer without rational justification, such as switching from an incorrect answer to another incorrect one or losing all the context, the latter being equivalent to a memory loss. Such behavior can be detrimental to learning, especially when a student holds misconceptions. This criterion is like a situation where a student is trying to guess the right answer at an examination without knowing the right answer. Therefore, LR should be a penalty, hence the negative sign in the final formula.

Overall, the AI score is defined as a weighted average of the above indicators as follows:

$$AI\ score\ [\%] = 70\%\ IP\ +\ 20\%\ R\ +\ 10\%\ SCA\ -\ 25\%\ LR \quad (1)$$

The results obtained by applying this formula are then translated into a letter grade scale, as shown in Table 1.

Table 1: Conversion of AI score [%] in the final letter-scale AI score

$\boldsymbol{AI\ score}$ [%]	$AI\ score\ badge$

$\mathbf{91\% \leq AI\ score\ [\%] \leq 100\%}$	
$\mathbf{81\% \leq AI\ score\ [\%] < 91\%}$	
$\mathbf{71\% \leq AI\ score\ [\%] < 81\%}$	
$\mathbf{61\% \leq AI\ score\ [\%] < 71\%}$	
$\mathbf{AI\ score\ [\%] < 61\%}$	

The formula for the AI score and the methodology of its calculation are the main results of the present work.

❑ **Methodology for calculating the AI score**

The goal of this section is to describe the reproducible test to perform in order to obtain the AI Score for any tested platform.
Before the test, one should clearly define the available database that the educational chatbots will be able to consult to perform the test. This documentation corpus must be the same for all compared chatbots to have a fair benchmark. In the present case, all chatbots have been provided with the class syllabus, annotated lecture slides and audio transcripts of the lectures.

The test will contain multiple-choice questions (MCQs), each having an unique answer. As the goal of the test is to discriminate between different chatbots, the selection of the MCQs is an important step (Step 1). Indeed, too easy questions would not enable us to differentiate between the different candidates or versions of the AI tutors. Hence, we suggest an optional pre-test to help with the choice of MCQs. Here, we use a student validated examination for which we have solid students' statistics.

Based on those statistics, we calculate a discriminant value Δ which constitutes a statistical measure of the discriminatory power of an examination question; that is, its ability to

differentiate high-performing candidates from those who are less proficient. This metric is based on comparing success rates between two groups of students previously identified according to their overall performance. Students are thus classified into two categories: strong students, whose overall average ranges between 12 and 20, and weak students, whose overall average is below 7 out of 20 (the maximum score at the exam being 20 points).

For each MCQ, two success proportions are calculated. The first, denoted PropSup, corresponds to the ratio between the number of correct answers provided by strong students and the total number of strong students. The second, denoted PropInf, is calculated analogously for weak students, representing the ratio between the number of correct answers from this group and the total number of weak students. The delta is then obtained by the difference between these two proportions: Δ = PropSup - PropInf.

The interpretation of Δ allows for the evaluation of each question's discriminatory quality. A high value, close to 1, indicates that the question is highly discriminant, with high-performing students answering correctly in a significantly higher proportion than weak students. Conversely, a Δ close to zero suggests that the question does not effectively differentiate between the two groups, either because it is too easy and successfully answered by all, or because it is too difficult and failed by the majority. A negative Δ, an undesirable situation, reveals a problematic question where weak students paradoxically achieve better results than strong students, which may indicate ambiguous or misleading wording. For the construction of the final test, the ten questions with the highest Δ values are selected to maximize the test's ability to discriminate between different performance levels of the evaluated chatbots. The selected questions for the present case are placed in Appendix A.

At the end of step 1, as a side benchmark, we submitted the full examination to the different chatbots, following the same evaluation grid as for students: +1 point for a correct answer, -0.25 points for a wrong answer. A second session (i.e. a second run) is given to any chatbot if they have at least one incorrect answer. This optional step helps to answer the question "*how does the chatbot compare to students?*". It is not a mandatory step, but it provides additional data to make the final decision.

The second step consists of defining the initial prompt and the follow-up prompt necessary for running the test (Step 2). They should be general enough to be applied to any of the ten MCQs. It is important to maintain those initial and follow-up prompts through the test.

The third step consists of performing the test by providing the ten selected MCQs to each pedagogical chatbot, with the initial prompt then with the follow-up prompt (Step 3). This is done five times (five runs) to ensure sampling has enough stochastic possibilities while maintaining the test in a reasonable timeframe. Each run corresponds to an independent session for a specific

chatbot. To reduce memory effect or potential contamination, new conversation or new API thread is considered for each run. The chatbot must answer according to the following format "Answer X" (X being one single letter). Initial and follow-up answers are recorded as well as logging time. The test goes on for every question, without any recall of previous items. All in all, in step 3, each chatbot is evaluated on 5 independent runs, producing 100 observations (10 MCQs × 5 runs ×initial or follow-up questions).

The fourth step is the calculation of the different criteria announced above (Step 4).

- The **Initial Performance (IP)** is the sum of the initial answers for each chatbot.
- The **Robustness (R)** compares the initial and follow-up answers. If both answers are the same and correct, the chatbot has a point. Maximum score is 50 if all follow-up answers are always correct and correspond to the correct initial answers.
- The **Self-correction ability (SCA)** examines the follow-up answer after an incorrect initial answer. If the follow-up answer is correct, the chatbot receives a point. If the initial answer was already correct, the chatbot also receives a point, a kind of bonus of being a good student.
- The **lack of reliability (LR)** is split into two sub criteria. First, there is an instability (I) sub-criterion if the initial answer is incorrect and the follow-up answer is also incorrect but different from the initial answer. In that case, the chatbot gets an extra point. The second sub-criterion is memory loss (ML). If the chatbot totally forgets the context between the initial prompt and the follow-up prompt, asking to repeat the original MCQ, the chatbot gets an extra point. The LR criterion is equal to one if either I or ML are equal to one and we perform a sum over the 50 tentatives. As a recall, LR has a negative impact on the AI score.

Maximum score of each criterion is 50.

Step 5 consists of calculating the AI score [%] using the formula (1)

$$AI\ score\ [\%] = 70\%\ IP\ +\ 20\%\ R\ +\ 10\%\ SCA\ -\ 25\%\ LR$$

and determine the letter by looking at Table 1.

As steps 4 and 5 are rather intricate, we developed a script for automatically calculating the AI score and providing the corresponding letter, available at https://aiscore.academy.

❑ Results

The test is done on six different platforms proposing the functionality to developed personalized chatbots, namely (1) ChatGPT from OpenAI [OpenAI, 2025], (2) Copilot Studio from Microsoft [Microsoft, 2025], (3) NotebookLM from Google using Gemini [Google, 2025], (4) Grok[xAI, 2025],

(5) Mistral AI [Mistral, 2025] and (6) Amanote using Claude 3.7-sonnet [Fery, 2017]. Technical specifications of platforms can be found in Appendix B.
The course is chosen to be an introductory physics class, more particularly an optics class for biological and veterinary sciences freshmen. We chose this class because we already have plenty of data regarding this class and GenAI [Henry, 2025]. All chatbots received the class syllabus, annotated lecture slides and audio transcripts of the lectures.

At step 1, we performed the pre-test to discriminate among 20 questions that were coming from a previous year examination. The first benchmark, comparing how the different chatbots perform compared to students, performed on all 20 questions to have a fair comparison with students, is reported in Table 2.

Table 2 Results of the pre-test

Educational chatbot	First attempt [/20]	Second attempt [/20]
ChatGPT	17.5	18.75
Copilot Studio	13.75	18.75
NotebookLM	20	/
Grok	17.5	17.5
Mistral	16.25	16.25
Amanote	17.5	17.5
Students	8.17	/

The presented result for the students is an average of 266 students, with an average score of 8.17 for a maximum score of 20, a median score of 8.25, a variance of 20.05 and a standard deviation of 4.48. No second attempt was made to the students with the same questions. NotebookLM which scored a perfect score on the first attempt, did not get a second chance either.

As observed in Table 2, all educational chatbot perform significantly better than students. It should be noted that all chatbots have access to the class material, equivalent to performing an open-book examination, which is not the case of students. Furthermore, one can see that the result of this first benchmark is not sufficient to effectively discriminate between the different chatbots. Indeed, all of them – except Copilot Studio at first attempt – performed relatively well (results ≥80% times correct) while performing the same 20 MCQs examination as students.

Hence, there is a need to perform steps 2 to 5, which further justifies the need of an AI Score. The initial and follow-up prompts are presented in Appendix C, while the results are displayed in Table 3.

Table 3: AI score of the six different educational chatbots tested.

Educational chatbot	IP	R	SCA	LR	AI Score [%]	AI Score
ChatGPT	43	40	44	0	85	B
Copilot Studio	38	29	39	24	60	E
NotebookLM	50	50	50	0	100	A
Grok	48	46	48	0	95	A
Mistral	42	41	46	0	83	B
Amanote	50	45	50	0	98	A

❑ Discussion

First, by looking at the results presented in Table 3, one can say that the AI score as defined in this work meets its initial goal: being able to discriminate between different educational chatbots, providing different objective criteria and similar testing conditions to rank different AI tutors. Although this process yielded comparative results, these should be interpreted with caution and should not be considered strict performance rankings as an absolute universal answer, but rather as an instantaneous picture comparing educational chatbots in a specific setting, at a peculiar time. Thus, our goal is not to classify chatbots as excellent or mediocre, but rather to establish a reliable rating system for educational chatbots. Nevertheless, several methodological sensitivities must be acknowledged, particularly regarding the heterogeneity of LLM architectures and configurations, which can influence score variations regardless of the actual quality of the responses. Comparing chatbots built on fundamentally different architectures presents significant methodological challenges. This heterogeneity makes it difficult to determine whether performance differences stem from underlying model capabilities, architectural choices, or parameter configurations. Ideally, rigorous comparisons should systematically vary configurations within the same model family to isolate parameter effects from architectural differences. However, practical constraints like API costs, access restrictions, and combinatorial complexity make such comprehensive testing challenging. Hence, once again, our results thus represent a comparative snapshot rather than definitive performance rankings.

The rapid evolution of LLM technologies threatens the reproducibility and long-term validity of comparative studies. Models are frequently updated or deprecated and configurations change without notice. This temporal instability means our findings capture a specific moment in a rapidly shifting landscape. Future research should account for this evolution by conducting

longitudinal comparisons tracking how successive model generations perform on identical benchmarks, while acknowledging that today's conclusions may quickly become obsolete. However, the letter grading system (Table 1) is broad enough to bring robustness to small technical sensitivities.

Nevertheless, one can discuss the objectivity of the selected indicators (IP, R, SCA and LR) used to determine the AI score. Indeed, evaluating educational chatbots requires a rigorous approach that ensures both their technical reliability and their pedagogical effectiveness. The ISO/IEC TR 24028:2020 standard [ISO/IEC, 2020] and ISO/IEC 42001:2023 [ISO/IEC, 2023] provide a conceptual and operational framework to justify the selected indicators and to ensure that the evaluation is based on recognized foundations in terms of reliability, ethics, and risk management. The ISO/IEC TR 24028:2020 standard defines the critical dimensions of the reliability of AI systems: robustness, transparency, absence of bias, and controllability. These concepts are directly integrated into the design of the AI Score. The Robustness (R) indicator, which assesses the chatbot's ability to maintain stable correct responses despite perturbations or follow-up prompts, aligns with the notion of reliability defined by the standard. Initial Performance (IP), which measures the ability to provide a correct response from the very first interaction, reflects the confidence that teachers and learners can have in the tool [Marchal et al., 2024]. The Lack of Reliability (LR) indicator, which penalizes inconsistencies and contradictions as well as loss of memory, is particularly critical in an educational context where incorrect or unstable responses can compromise the quality of learning [Astolfi et al., 1993; Legendre, 2005].
The ISO/IEC 42001:2023 standard, for its part, proposes best practices in governance, risk assessment, and continuous improvement. These principles strengthen the relevance of the AI Score by situating it beyond a purely technical evaluation toward the integration of ethical and operational dimensions. Specifically, the LR indicator quantifies the potential risks associated with using the chatbot and encourages a dynamic use of the AI Score, enabling educational institutions to adjust and optimize their chatbots objectively with time.
Thus, the AI Score combines technical criteria (ISO standards) and pedagogical criteria.

Nevertheless, one could discuss the weight of the different criteria proposed in formula (1). Let us briefly justify the chosen weight of the different criteria.

- **Initial Performance (IP):**

IP with its maximum weighting constitutes the fundamental indicator of the AI Score. Empirically, this priority is based on several behavioral and pedagogical observations. First, students do not sufficiently verify the accuracy of AI results [Martin-Moncunill, 2025]. Moreover, the first response establishes the initial level of trust in the AI tool [Marchal et al., 2024; Arnoldi, 2022],

thus creating a lasting cognitive anchoring effect. Finally, a correct response from the first interaction maximizes pedagogical effectiveness by reducing the risk of anchoring conceptual errors. This priority given to IP also finds its justification in established theoretical frameworks. On one hand, it aligns with Sweller's (1988) cognitive load theory, according to which an immediate correct response reduces extraneous cognitive load and allows the learner to focus their mental resources on understanding the content. On the other hand, it is consistent with the concept of epistemological obstacle developed by Astolfi and Peterfalvi (1993), which demonstrates that an initial incorrect response can create or reinforce misconceptions that are particularly difficult to deconstruct subsequently.

- **Robustness (R):**

R justifies an intermediate weighting due to its importance for trust and learning, although it requires that the student formulate a follow-up question or express doubt concerning the initial response. Active verification, such as cross-referencing information with academic sources, remains minority practice, even in an educational context where accuracy is crucial [Martin-Moncunill, 2025]. Robustness is also of particular importance for the learner's metacognitive development. Indeed, if the student questions the chatbot to obtain more information and the latter can maintain a consistent correct response, the tool can be considered as support and encouragement for critical thinking. Moreover, robustness can be related to the term "reliability" as defined by the ISO/IEC TR 24028:2020 standard [ISO/EIC, 2020]. The proposed weighting reflects its substantial importance while recognizing the primacy of initial performance.

- **Self-Correction Ability (SCA):**

SCA deserves a lower intermediate weighting due to its more limited impact. Indeed, self-correction requires the meeting of a double condition: first, the chatbot must have provided an initially incorrect response, then the student must actively contest this response based on their personal knowledge. This situation presents a low probability of occurrence, estimated at less than 10% of interactions, as it requires a learner sufficiently critical and confident to have detected the error and questioned it [Martín-Moncunill & Alonso Martínez, 2025]. Theoretically, SCA nevertheless presents important relevance for managing hallucinations inherent to language models (LLM), a problem well documented in the literature [Cabezas-Clavijo et al., 2025; Elnaffar et al., 2025; Lee et al., 2024; Pakeh et al., 2025; Swacha et al., 2025; Extance, 2023]. It also fits within the continuous improvement approach advocated by the ISO/IEC 42001:2023 standard concerning AI systems. The proposed weighting reflects its secondary but non-negligible character in the overall evaluation of the educational chatbot's quality.

- **Lack of Reliability (LR):**

LR constitutes the only indicator to receive a negative weighting, justified by its potentially harmful impact on learning. Indeed, changing responses without logical justification creates confusion and distrust in the learner, seriously compromising the system's pedagogical effectiveness. LR constitutes a clear violation of the controllability principle established by the ISO/IEC TR 24028:2020 standard [ISO/EIC, 2020] and goes against the objectives of evaluation and risk management in educational AI as defined by the ISO/IEC 42001:2023 standard [ISO/EIC, 2023]. Furthermore, a loss of memory or of the context would not help for creating trust with the developed AI tutor, hence it has to be penalized. The proposed weighting for LR (25%) represents a penalty proportional to the severity of the observed behavior and its potential impact on learning.

One could object that the chosen weights are somewhat arbitrary and that different versions of formula (1) could give different outcomes, hence hindering the universality of the AI score. This is indeed true and the proposed AI Score is only valid with the corresponding formula clearly indicated. Different weights or additional criteria could modify the present formula (1) depending on the goals of the person willing to perform the test. If someone wants manually adjust the weight of the AI Score for their personal test, this possibility is provided by the AI Score calculator available on https://aiscore.academy/aiscorec.php. As a recall, the main goal of the development of the AI score is to provide a standard test to benchmark different available platforms available either online or commercially and the provided weights are found to be the best to fill that goal.

Although the AI Score makes it possible to objectively measure the technical performance of chatbots through four quantifiable criteria, this approach has significant limitations in real educational contexts. It does not account for essential dimensions such as user acceptability, the motivation it generates, or the perceived quality of the interaction (Marchal et al., 2024; Arnoldi, 2022). Indeed, three studies converge on the need for a multidimensional evaluation. Arnoldi (2022) recommends an approach centered on user experience, integrating usability, emotions, engagement, and perceived presence. Marchal et al. (2024) empirically confirm that students evaluate chatbots according to criteria far broader than simple accuracy: ease of use, fit to needs, response speed, and the quality of the overall experience. Their study also reveals that 95% of interactions relate to pedagogical-intellectual support, with socio-affective dimensions being entirely absent. Furthermore, Qiu et al. (2025) operationalize this holistic vision through a systematic evaluation framework integrating three dimensions: learning gains, engagement, and perception. Their methodology, grounded in learning analytics, combines quantitative data (test scores, interaction logs, time-based metrics) and qualitative data (thematic analyses of feedback). Their quasi-experimental study illustrates the complexity of such evaluation: although no

significant difference was observed in terms of learning gains, the group using a Socratic chatbot showed more sustained engagement and more dynamic interaction patterns on difficult topics.

For a comprehensive evaluation of an educational chatbot, it is therefore appropriate to combine the AI Score as a prior technical validation with the systematic framework of Qiu et al. (2025) to measure longitudinal pedagogical impact, behavioral and cognitive engagement, as well as learners' perceptions. This complementary approach recognizes that technical quality is a necessary but not sufficient condition to guarantee the educational effectiveness of a chatbot, which also depends on contextual, pedagogical, and socio-affective factors.

Finally, the proposed AI score and the related methodology can be objected to be a purely quantitative tool while the output of educational chatbots is inherently qualitative. For example, it occurred that a chatbot provides a wrong answer letter while providing a correct (but unasked) justification, referring to the correct answer. Or oppositely, the AI tutor may provide the correct letter with a false justification and thus sur-estimate the real AI score. In such a case, we would suggest slightly modifying the initial prompt by asking for an additional short justification of the answer and treating that litigious case as one would judge an oral examination rather than as a written MCQ examination. One should however recall that the 5 runs moderate the probability of occurrence of such problems. Nevertheless, an additional analysis of the quality of the answers provided by the chosen best-performing AI tutor is strongly advised before giving it to students. Furthermore, other criteria such as the rapidity of answering, sycophancy, the verbosity of the AI tutor or the status of protection of the data could also be considered before choosing an AI tutor but those are beyond the scope of the present AI score.

❑ Conclusions

The present work proposes a methodology to objectively compare the performances of AI conversational agents used as educational chatbots by defining an AI Score. The score is established by questioning the educational chatbots and evaluating their initial performance, robustness, self-correction ability, and lack of reliability. The rigorous methodology is derived using a set of multiple-choice questions, preferably tested on students beforehand. A letter scale, from A (best) to E (poorest), is then calculated providing an indicator compliant with the ISO/IEC TR 24028:2020 [ISO/IEC, 2020] and ISO/IEC 42001:2023 [ISO/IEC, 2023] standards. Such analysis should guide educators in selecting and developing AI tools for teaching purposes. Limitations of the present study are also discussed in terms of the choice of the different criteria and their respective weights in the final formula.

The present AI Score would benefit from further study by coupling this quantitative indicator to a more qualitative one judging the quality of the answers provided. Longitudinal studies to judge the robustness of the AI Score across time, with various versions of new LLM or assessing the importance of the metaprompt on the AI Score of the AI tutor would also be interesting.

Furthermore, additional studies such as removing the RAG only constraint, a transposition to other academic domains or a generalization of the AI Score to other domains than education where conversational agents are developed and where users are placed in similar situations as students, are interesting perspectives. We believe that the present methodology and AI Score, in addition to the handy associated website https://aiscore.academy will help educators to better assess the development of AI tutor before releasing it to students and spark didactic developments.

Contributions

M.D., JR. M. and M.L. devised the project and the main conceptual ideas. JR. M. carried out the experiments. L.D. helped set up the automatic calculations of the AI Score. JR. M. set up the website with the help of SerTIC service and communication administration of UNamur. All authors discussed the results, adapted the formula and criteria and contributed to the final manuscript. M.L. wrote the first draft of the manuscript, with the help of M.D. and input from all authors. M.L. supervised the project.

Funding

M.L. is a Research Associate of the Fonds de la Recherche Scientifique – FNRS. The authors acknowledge the support of the University of Namur via the PUNCh GenAI4Student project.

Acknowledgements

The authors would like to thank Julie Henry, Guillaume Mele, Clara Depommier, Julien Colaux, Benoit Frenay and Carine Michiels for fruitful discussions leading to this work. The authors also would like to thank the SerTIC and communication administration of UNamur for their help setting up the website.

Bibliography

- Adiguzel, T., Kaya, M. H., & Cansu, F. K. (2023). Revolutionizing education with AI: Exploring the transformative potential of ChatGPT. *Contemporary Educational Technology*, 15(3), ep429. https://doi.org/10.30935/cedtech/13152

- Ait Baha, T., El Hajji, M., Es-Saady, Y., & Fadili, H. (2023). The impact of educational chatbot on student learning experience. Education and Information Technologies. https://doi.org/10.1007/s10639-023-12166-w
- **Anderson, L. W., & Krathwohl, D. R. (Eds.). (2001).** *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives* (Complete ed.). Addison Wesley Longman.
- Anthropic (2025). Claude 3.7 Sonnet [Modèle de langage IA].Claude 3.7 Sonnet: Advanced Hybrid AI for Every Task
- Arnoldi, A. (2022). *Comment évaluer un chatbot assistant de cours utilisé en formation à distance ? Élaboration d'un questionnaire et évaluation d'ADIDBot* [Master thesis, Université de Genève]. Faculté de Psychologie et des Sciences de l'éducation. https://tecfa.unige.ch/tecfa/maltt/memoire/Arnoldi2022.pdf
- Arvin, C. (2025, June 12). *"Check my work?": Measuring sycophancy in a simulated educational context* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2506.10297
- Astolfi, J.-P., & Peterfalvi, B. (1993). Obstacles et construction de situations didactiques en sciences expérimentales. *Aster : Recherches en Didactique des Sciences Expérimentales*, *16*, 103-141.
- Bloom, B.S. (1956) Taxonomy of Educational Objectives, Handbook The Cognitive Domain. David McKay, New York.
- Cabezas-Clavijo, Á., & Sidorenko-Bautista, P. (2025). *Assessing the performance of 8 AI chatbots in bibliographic reference retrieval: Grok and DeepSeek outperform ChatGPT, but none are fully accurate* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2505.18059Chatterji, A., Cunningham, T., Deming, D. J., Hitzig, Z., Ong, C., Shan, C. Y., & Wadman, K. (2025). *How people use chatgpt* (No. w34255). National Bureau of Economic Research.
- Davar, N. F., Dewan, M. A. A., & Zhang, X. (2025). AI Chatbots in Education: Challenges and Opportunities. Information, 16(3), 235. https://doi.org/10.3390/info16030235
- Debets, T., Banihashem, S. K., Brinke, D. J.-T., Vos, T. E. J., De Buy Wenniger, G. M., & Camp, G. (2025). Chatbots in Education: A systematic review of objectives, underlying technology and theory, evaluation criteria, and impacts. Computers and Education, 234, Article 105323. https://doi.org/10.1016/j.compedu.2025.105323
- Dempere, J., Modugu, K., Hesham, A., & Ramasamy, L. K. (2023). The impact of ChatGPT on higher education. Frontiers in Education, 8, 1206936. https://doi.org/10.3389/feduc.2023.1206936
- Dhyne, M., & Plumat, J. (2025). *L'IA au service des stagiaires et formateurs pour évaluer et améliorer les préparations de cours*. *SHS Web of Conferences, 214*, 01008. https://doi.org/10.1051/shsconf/202521401008
- Elnaffar, S., Rashidi, F., & Abualkishik, A. Z. (2025, October 4). *Teaching with AI: A systematic review of chatbots, generative tools, and tutoring systems in programming education* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2510.03884

- Elkot, M. A., Alhalangy, A., Abdalgane, M., & Ali, R. (2025). *Pedagogical influence of AI-chatbots on learning outcomes: A systematic review*. *International Journal of Educational Methodology, 11*(4), 527–540. https://doi.org/10.12973/ijem.11.4.527
- Essel, H.B., Vlachopoulos, D., Tachie-Menson, A. *et al.* (2022) The impact of a virtual teaching assistant (chatbot) on students' learning in Ghanaian higher education. *Int J Educ Technol High Educ* **19**, 57. https://doi.org/10.1186/s41239-022-00362-6
- Extance, A. (2023). ChatGPT has entered the classroom: how LLMs could transform education. *Nature*, *623*(7987). https://doi.org/10.1038/d41586-023-03507-3
- Fery, A. (2017). *Amanote: A modern note-taking application.* (Unpublished master's thesis). Université de Liège, Liège, Belgique. Retrieved from https://matheo.uliege.be/handle/2268.2/2612
- Google. (2025). *NotebookLM* [Outil d'assistance à la recherche (IA)]. https://www.google.com/notebook/
- Henry, J., Lobet M. (2025). *Comprendre les représentations étudiantes de l'intelligence artificielle générative : profils et modèles mentaux*, under revision.
- ISO/IEC. (2020). *Information technology — Artificial intelligence — Overview of trustworthiness in artificial intelligence* (ISO/IEC TR 24028:2020). International Organization for Standardization. https://www.iso.org/standard/77608.htm
- ISO/IEC. (2023). Information technology — Artificial intelligence — Management system (ISO/IEC 42001:2023). International Organization for Standardization. https://www.iso.org/standard/42001.html
- Jin, Y., Yan, L., Echeverria, V., Gašević, D., & Martinez-Maldonado, R. (2025). *Generative AI in higher education: A global perspective of institutional adoption policies and guidelines*. *Computers and Education: Artificial Intelligence, 8*, 100348. https://doi.org/10.1016/j.caeai.2024.100348
- Kasneci, E., Seßler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., … Kasneci, G. (2023). *ChatGPT for good? On opportunities and challenges of large language models for education*. *Learning and Individual Differences, 103*, 102274. https://doi.org/10.1016/j.lindif.2023.102274
- Kooli, C. (2023). Chatbots in education and research: A critical examination of ethical implications and solutions. Sustainability, 15(7), 5614. https://doi.org/10.3390/su15075614
- Krathwohl, D. R. (2001). *A revision of Bloom's taxonomy: An overview*. Theory Into Practice, 41(4), 212–218. https://doi.org/10.1207/s15430421tip4104_2
- Labadze, L., Grigolia, M. & Machaidze, L. (2023) Role of AI chatbots in education: systematic literature review. *Int J Educ Technol High Educ* **20**, 56. https://doi.org/10.1186/s41239-023-00426-1

- Lee, D., Arnold, M., Srivastava, A., Plastow, K., Strelan, P., Ploeckl, F., Lekkas, D., & Palmer, E. (2024). *The impact of generative AI on higher education learning and teaching: A study of educators' perspectives*. *Computers and Education: Artificial Intelligence, 6*, 100221. https://doi.org/10.1016/j.caeai.2024.100221
- Legendre, R. (2005). *Dictionnaire actuel de l'éducation* (3e éd.). Guérin.
- Marchal, P., Kumps, A., Floquet, C., Deruwé, O., et De Lièvre, B. (2024). Perceptions et usages d'un chatbot comme tuteur de cours en sciences de l'éducation. *Revue internationale sur le numérique en éducation et communication*, *18*, 125-147. https://doi.org/10.52358/mm.vi18.410
- Martín-Moncunill, D., & Alonso Martínez, D. (2025). Students' trust in AI and their verification strategies: A case study at Camilo José Cela University. *Education Sciences, 15*(10), 1307. https://doi.org/10.3390/educsci15101307
- Microsoft. (2025). *Microsoft Copilot Studio* [Assistant IA]. https://Copilot Studio.microsoft.com/ Copilot Studio.microsoft.com
- Mistral AI. (2025). *Le Chat* [Assistant IA / modèle de langage]. https://mistral.ai/products/le-chat Mistral AI
- OpenAI. (2025). ChatGPT(version GPT-4o mini) [Modèle de langage IA]. https://openai.com/Parekh, K. V., Saxena, N., & Ansari, M. A. (2025). *A comparative study of retrieval-augmented generation (RAG) chatbots*. In *2025 International Conference on Automatics, Robotics and Artificial Intelligence (ICARAI)* (pp. 1–6). IEEE. https://doi.org/10.1109/ICARAI67046.2025.11137956
- Qiu, W., Su, C. L., Jamil, N. B., Thway, M., Ng, S. S. H., Zhang, L., Lim, F. S., & Lai, J. W. (2025). A systematic approach to evaluate the use of chatbots in educational contexts: Learning gains, engagements and perceptions. Computers, 14(7), 270. https://doi.org/10.3390/computers14070270
- Swacha, J., & Gracel, M. (2025). Retrieval-Augmented Generation (RAG) Chatbots for Education: A Survey of Applications. Applied Sciences, 15(8), 4234. https://doi.org/10.3390/app15084234
- Swellller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science, 12*(2), 257–285. https://doi.org/10.1207/s15516709cog1202_4
- xAI. (2025). *Grok* [Modèle de langage (IA)]. https://x.ai/grok

Appendix A : **Selected questions and their corresponding Δ**

The examination consists of 20 questions and was administered to students, who received a score out of 20. For each question, we have the number of students who selected each answer option (A, B, C, D, or E), and these data are broken down by three performance groups: students who scored less than 7 out of 20, those who scored between 8 and 11, and those who scored at least 12 points out of 20.

From these data, we calculate the Delta index for each question using the following formula:
Δ = PropSup - PropInf
where PropSup is the proportion of students who gave the correct answer among those who scored ≥ 12/20, and ProfInf is the proportion of students who gave the correct answer among those who scored < 7/20.

This Δ index allows us to assess the discriminating power of each question, that is, its ability to distinguish high-performing students from struggling students.

Interpretation of Δ values:

- High Δ (close to 1): The question has excellent discriminating power. High-performing students overwhelmingly succeed on this question while struggling students fail, indicating that the question effectively measures mastery of the assessed content.
- Moderate Δ (approximately 0.3 to 0.5): The question has acceptable discriminating power. It reasonably differentiates between the two groups of students, although some weak students may also answer it correctly.
- Low or zero Δ (close to 0): The question does not discriminate between high-performing and struggling students. This may indicate that the question is either too easy (everyone succeeds), or that it assesses non-essential knowledge, or that it is poorly worded.
- Negative Δ: A problematic situation where weak students perform better than high-performing students. This suggests a poorly designed, ambiguous question, or one whose wording misleads the best students. Such questions should be revised or eliminated.

The following table presents the analysis of the exam questions presented to the students and which have been selected (Δ > 0.50) for the chatbots testing.

Question's Number	PropSuf	PropInf	Delta
1	0.96	0.42	**0.54**
2	0.91	0.40	**0.51**
3	0.87	0.66	0.21
4	0.25	0.01	0.24
5	0.75	0.21	**0.53**
6	0.87	0.21	**0.66**

7	0.94	0.30	**0.64**
8	0.68	0.23	0.45
9	0.25	0.29	-0.04
10	0.60	0.04	**0.56**
11	0.89	0.32	**0.57**
12	0.37	0.12	0.25
13	0.97	0.65	0.31
14	0.76	0.23	**0.53**
15	0.68	0.20	0.48
16	0.79	0.28	**0.51**
17	0.48	0.14	0.34
18	0.95	0.42	**0.53**
19	1.00	0.64	0.36
20	0.95	0.46	0.49

Therefore, the selected questions (in French) are the following, the right answer is in bold:

Q1 Que vaut la vitesse de propagation de la lumière dans le verre ? Les réponses proposées sont

A. 0 m/s

B. 1.82 ×10^8m/s

C. 2.00 ×10^8m/s

D. 3.00×108^m/s

E. 5.49 ×10^−9m/s

Q2 Planck a révolutionné la physique moderne en postulant un lien entre énergie et fréquence. Cette relation s'écrit :

A. E=hf

B. E=hc/λ

C. E=$\hbar\omega$

D. Les propositions de A à C sont toutes correctes

E. Les propositions de A à C sont toutes incorrectes

Q5 Au vu de la distance entre les atomes d'un cristal de NaCl, des rayons X incidents vont typiquement subir un phénomène de

A. diffraction

B. polarisation

C. dispersion

D. absorption

E. Les propositions de A à D sont toutes correctes

Q6 Un faisceau lumineux arrive sur une interface air/eau. Quelle fraction de l'intensité lumineuse – exprimée en % - est transmise ? Supposez qu'il n'y a pas d'absorption.

A. 0.4%

B. 2.0%

C. 98.0%

D. 99,6.0%

E. Les propositions de A à D sont toutes incorrectes

Q7 Un prisme fait de fluorure de magnésium $MgCl$2MgCl2 possède un indice de réfraction de 1.377 à 587 nm. Quel est l'angle critique pour la réflexion totale ? On suppose le prisme entouré d'air

A. 24.4°

B. 46.6°

C. 56.4°

D. 90°

E. Les propositions de A à D sont toutes incorrectes

Q10 Le graphite est composé d'un empilement de plans parallèles de graphène qui est un matériau bidimensionnel (i.e. une couche monoatomique d'atome de carbone). La distance entre chaque plan de graphène est de 0.33 nm. Si un faisceau monochromatique de rayons X de 1 Å de longueur d'onde est dirigé vers du graphite, quel est l'angle le plus petit (et non nul) pour lequel on obtiendra une interférence constructive ?

A. 8.7°

B. 9.6°

C. 19.5°

D. 20.4°

E. 21.5°

Q11 Un faisceau de lumière est incident sur une interface entre un verre borosilicaté d'indice de réfraction 1.55 et d'un matériau inconnu. La lumière est incidente côté verre, l'angle entre l'interface et le faisceau lumineux vaut 60°. Le faisceau réfracté sort avec un angle de 35.6° par rapport à la normale. Quel est ce matériau inconnu ?

A. De l'air

B. De l'eau

C. Du verre

D. Du fluorure de magnésium

E. Il est impossible de déterminer le matériau inconnu avec ces données

Q14 Question pour un champion. Je suis la partie de l'œil jouant le plus grand rôle dans la puissance de l'oeil, je suis

A. La cornée

B. Le cristallin

C. La pupille

D. La rétine

E. L'humeur aqueuse

Q16 Afin de calculer le grossissement commercial d'un microscope, il est nécessaire de connaitre

A. La distance focale de l'oculaire et la distance objet de l'oculaire

B. La distance focale de l'objectif et la distance objet de l'objectif

C. La distance focale de l'objectif et la distance objet de l'oculaire

D. La distance focale de l'oculaire et la distance objet de l'objectif

E. La puissance de l'objectif et de l'oculaire et la longueur optique

Q18 Quel est le grossissement angulaire d'une loupe de 25 cm de distance focale ?

A. 1 fois

B. 2 fois

C. 5 fois

D. 40 fois

E. Les propositions de A à D sont toutes incorrectes

Appendix B : LLM Selection and Technical Specifications

This section provides an overview of the technical characteristics of the six chatbots examined in this study (July 2025), including their underlying language models, architectural features, and key functionalities.

To ensure a fair and rigorous comparison, all chatbots were provided with identical inputs: the same prompt and the same educational resources, including PowerPoint presentations, videos, and text documents. NotebookLM does not have any pre-recorded metaprompt, but still get the same initial prompt. Before analyzing their respective outputs, this section presents the technical characteristics of each platform, including their underlying language models, architectural features, and key functionalities that may influence their processing and interpretation of these materials. It should be noted that few data are available from the companies, so the following are given to the best of our knowledge.

- **Copilot Studio Studio (agent)**
    - Model : GPT-4 Turbo
    - Temperature : 0.2-0.3 (auto-declaration)
    - Top-p : +/-0.9 (auto-declaration)
    - Top-k : +/- 40 (auto-declaration)

- **OpenAI assistant**
    - Type of chatbot : assistant freely accessible through a webpage (API)
    - Model : gtp-4o (model used in ChatGPT)
    - Tools : file search, code interpreter
    - Response format : text
    - Temperature : 1.00, representing a balanced approach between deterministic and creative outputs, allowing for natural variation in responses while maintaining coherence.
    - Top P : 1.00, ensuring no restriction on the token selection pool and allowing the temperature parameter to fully govern response variability.
- **Mistral**
    - Standard interface : Le Chat
    - Model : Mistral Large 2
    - Temperature and TopP : No public documentation available on these parameters
- **Grok**
    - Model : Grok 4
    - Temperature : According to some users, the temperature would be low by default (officially unverified information)
- **NotebookLM**
    - Model : Gemini 1.5 Pro
    - Temperature and TopP : No public documentation available on these parameters
    - Interaction only with user-provided documents, grounded and secured approach
- **Amanote**

- Plugin installed on the university's Moodle platform (version 4.5)
- Model : eu.anthropic.claude-3-7-sonnet-20250219-v1
- Temperature and TopP : No public documentation available on these parameters

Appendix C : Initial and follow up prompts

Initial prompt (French)

Tu es un tuteur d'optique pour des étudiants de première année d'université. Tu dois répondre en français. Pour chaque question QCM qui t'est posée, tu dois indiquer la lettre de la réponse que tu juges correcte sous la forme "Réponse: X" (où X=A, B, C, D ou E). Tu es obligé de donner une réponse parmi celles proposées.

Translation

You are an optics tutor for first-year university students. You must answer in French. For each multiple-choice question you are asked, you must indicate the letter of the answer you consider correct in the form "Answer: X" (where X=A, B, C, D, or E). You must choose one of the answers provided.

Follow-up prompt (French)

Tu viens de répondre « [copier/coller de sa réponse] ». Es-tu sûr de cette réponse ? Vérifie ton raisonnement étape par étape. Si tu détectes une erreur de logique, de calcul ou d'interprétation du cours, corrige-la et indique la lettre correcte. Réponds en conservant le format : « Réponse confirmée : X » ou « Réponse corrigée : Y » + justification courte.

Translation

You just answered "[copy/paste your answer]." Are you sure about this answer? Check your reasoning step by step. If you detect an error in logic, calculation, or interpretation of the course material, correct it and indicate the correct letter. Respond using the following format: "Answer confirmed: X" or "Answer corrected: Y" + brief justification.