Interdisciplinary Description of Complex Systems 23(6), 668-684, 2025

LARGE LANGUAGE MODELS IN PHYSICS: ANALYSIS OF ACCURACY AND TEACHER PERCEPTION

Boško Lišnić and Marija Gaurina*

University of Split, Faculty of Science
Split, Croatia

DOI: 10.7906/indecs.23.6.5
Regular article

Received: 29 July 2025.
Accepted: 12 November 2025.

ABSTRACT

This article explores the educational potential of large language models (LLMs) in physics teaching for secondary schools. The aim is to examine the accuracy, consistency and execution time of LLMs in solving tasks from national physics exams for secondary schools in the Republic of Croatia. In addition, we want to collect teachers' perceptions of their use in teaching and as a tool in preparation for state maturas. The quantitative part analysed the responses obtained from 15 models from four different platforms (OpenAI, Perplexity, Microsoft Copilot, DeepSeek) in solving tasks from national physics exams. In the qualitative part, a focus group was conducted with physics teachers. The results show a high level of accuracy of individual models, and problems in solving tasks with a graphical display were revealed. Teachers recognised the potential of LLMs as auxiliary tools but emphasised the necessity of students' prior knowledge in physics and teachers' need for critical literacy and control. The research provides insight into the possibilities and limitations that LLMs bring with them in STEM education and opens space for further research into the application of artificial intelligence tools in education.

KEY WORDS

LLM, education, physics, STEM, AI literacy, evaluation

CLASSIFICATION

APA: 3580, 3600

JEL: I21, I25

*Corresponding author, η: mgaurina@pmfst.hr; -; -

INTRODUCTION

The emergence of large language models (LLMs) has brought new opportunities to the field of education [1]. LLMs have the potential to provide a wide range of benefits and opportunities for students at all stages of education [2]. Today, students actively use various LLM tools, such as ChatGPT, as a tool in class or at home. According to research from 2023, already then 89% of surveyed American students used ChatGPT to do homework [3], and today those numbers are certainly even higher. Artificial intelligence tools provide numerous advantages. Thus, some findings highlight the positive impact of AI on the advancement of conceptual understanding, providing personalised learning, facilitating social interaction and assessment methods [4]. Special emphasis is placed on various learning supports that are based on the personal needs of students [5]. An improvement in more efficient learning and increased autonomous learning was recorded [6]. LLMs have excellent interdisciplinary opportunities, allowing students to connect in integrated learning and develop interdisciplinary thinking skills [5]. With the aim of improving learning methods, the use of LLMs is on the rise as a tool that achieves student achievement levels in subjects such as mathematics, computer science, and physics [1], therefore, it is not surprising that such tools are increasingly used in solving factual, as well as mathematical-logical tasks. Although students appreciate the explanations provided, they may also lose confidence in the tools due to inaccuracies [7]. These tools bring with them challenges in their use in teaching. Challenges have been identified in the form of technical infrastructure, training data, and data privacy [4]. These tools are limited to the data they are trained on and can give fictional answers in a convincing tone [8].

Therefore, we aim to examine how useful LLMs can be in preparing students for the high school physics exam. First, we want to determine the accuracy of the solutions obtained from such models. There is also a need to check how consistent they are in their answers, whether they differ in time of execution, and what teachers' attitudes towards their use are. The research question is: How accurate, consistent, and time-efficient are LLMs in solving tasks for the high school physics exam, and how are they used, perceived and interpreted by physics teachers?

RELATED WORKS

LLMs show safety deficiencies, especially in situations involving unclear, ambiguous and ethical tasks, according to the SafetyBench study [9]. Although the performance is formally good, there is a tendency to hallucinate facts, which in an educational context can seriously undermine students' trust in the systems [9]. Maitland et al. [10] highlighted a high proportion of factual and conceptual errors that are unacceptable in the context of the reliability of clinical decision-making systems. As with the clinical context, we can draw a parallel with the educational context: models make false claims in a convincing tone that causes misconceptions in students. In physics, such mistakes reinforce misconceptions students already have, which [11] also highlights as a critical risk: "LLMs often reinforce misconceptions because they respond with great certainty and authority". Rong et al. [12] introduce the concept of exclusionary reasoning, which refers to the model's ability to know when not to intervene, depending on the situation. This plays an especially important role in education when unwanted corrections of student answers can reduce self-confidence. Wu et al. [13] point out that multi-step explanations (e.g. chain-of-thought prompting) can reduce accuracy because hallucinating the model can lead to incorrect answers. Sonkar et al. [14] define Student Data Paradox, the concept related to model training on student-tutor dialogue, which results in better imitation of misconceptions but weaker reasoning. Thus, he offers a solution in the form of hallucinatory tokens, which can also be used in physics because they are models that distinguish between situations, i.e. when to "act" and when to give the correct answer. Although LLMs show good performance in factual responses, they can also show deficiencies in responses that involve visual elements, understanding symbols, or that require deep reasoning [15, 16].

Careful pedagogical integration is needed, given that uncritical and unmoderated use of LLMs in education can lead to over-reliance on models and reduce student engagement. On the other hand, some research highlights the potential of LLMs in supporting reflective learning if used in the right way through a structured framework and teacher guidance [11, 17]. Regarding evaluation, Chang et al. [15] propose a multidimensional framework for evaluating LLMs, addressing the questions of what to evaluate (type of task), where (benchmark tasks), and how (methods and metrics). Such and similar models allow us to assess the appropriateness of LLMs in fields such as education [15].

Given that AI tools surround students, the importance of AI literacy is emphasised. This concept primarily refers to a critical review of the answers provided by AI, the evaluation of sources, and the interaction of AI tools as an addition to learning, not a replacement [18, 19]. Such an approach supports the desire to use LLMs as tools and objects for critical reflection. Similarly, Chiu et al. [18] develop a concept of *AI competencies that* includes assessing the plausibility of responses, awareness of model limitations, and caution against hallucinations. Liang et al. [20] emphasise that the clarity and usefulness of the responses provided by LLMs should be increased, with supervision and structured feedback. Therefore, a prerequisite for this is appropriate prompts and constant validation of results. As pointed out by Wang et al. [21], multiple verification, prompt engineering and pedagogical supervision play a major role so that LLMs can be used to generate assignments and automatically evaluate student responses.

This article will contribute to the use of LLMs in the educational field by evaluating their performance, consistency, and execution time on the tasks of the national physics exams in the Republic of Croatia.

METHODOLOGY

Mixed research was conducted, which had a quantitative and qualitative component. The quantitative part of the research focused on comparative research with the aim of checking the accuracy of different LLMs, similarities between the models, consistency, execution time, and relationship between accuracy and execution time in solving state matura tasks in physics in Croatia. The qualitative part of the research was based on the focus group technique to gain insight into the opinions and perceptions of physics teachers. The process of collecting, testing and analysing data was from April to June 2025.

QUANTITATIVE PART OF THE RESEARCH

The research began by testing 15 different large language models (LLMs), to assess their accuracy in solving tasks from the state matura in physics in the Republic of Croatia. The state maturas are administered in the 4th grade of secondary school and are one of the criteria for university admission. The questions relate to the fields of physics: mechanics, electromagnetism, thermodynamics, optics and modern physics. Tasks from the last five years (2019-2024) were considered to ensure the representativeness of current tasks and to include possible changes in the content and structure of the state matura in physics. This range allows for analysing trends and comparison of models in recent exams. The tasks are divided into two test booklets:

- Exam booklet 1 – Multiple choice questions (A, B, C, D), each task carries one point.
- Exam booklet 2 – Extended answer questions, worth two to four points depending on the complexity of the task.

Each year has an average of 24 multiple-choice questions and 11 extended-response questions. In total, there were 122 multiple-choice questions and 57 extended-response questions. It is important to note that both test booklets also contained questions with graphical displays (e.g.,

pictures, graphs, diagrams). These questions were further separated in the statistical analysis so that their accuracy could be observed separately compared to questions without visual elements. The aim was to examine the extent to which graphical elements affect the model's ability to provide an accurate answer. There were 45 such questions (36 of which were multiple-choice and nine extended-response).

Platforms and Models

We tested questions on four different platforms across a total of the following 15 models.

1. ChatGPT (OpenAI): ChatGPT 4o, ChatGPT o3, ChatGPT o4-mini, ChatGPT o4-mini-high, ChatGPT 4.5, ChatGPT 4o mini,
2. Perplexity: Sonar, Claude 3.7 Sonet, GPT-4.1, Grok 3 Beta, R1 1776, o4-mini, Claude 3.7 Sonet Thinking,
3. Microsoft Copilot: GPT-4 Turbo,
4. DeepSeek: DeepThink (R1).

The platforms were selected because, according to the authors' observations, they represent the most widely used LLM tools in Croatia's educational environment, providing results relevant to real-world use. All available model variants from the mentioned platforms were tested at that time. Two models were excluded from the analysis. OpenAI ChatGPT 4 (because it stopped working during the testing phase) and Perplexity's Gemini 2.5 Pro (because it would not parse part of the task files, despite multiple requests and attempts). The goal was to compare not only different LLMs across platforms, but also the performance of different versions within the same platform.

It should be noted that the paid versions of ChatGPT Pro and Perplexity Pro were used. Microsoft Copilot was used via a Microsoft 365 Education license. For DeepSeek, only an email address was used for login.

Method of Conducting Quantitative Analysis

Each model was tested in turn via the mentioned platforms. A query was set and a file, i.e. an exam booklet with tasks for a particular year, was attached. A separate query was run for multiple-choice questions and a query for extended-response questions. The same queries were repeated for each year of the state matura. All models received identical, neutrally formulated queries, without additional instructions that could direct the model towards the correct answer. The first answer generated by the model was accepted, without subsequent changes, additional clarifications or retries. The tasks were set in new sessions for each year of the state matura to avoid potential bias due to the context of previous questions. In order to check the consistency of the model, the same queries were tested in later stages of the research. Example of a query for a multiple-choice task: "*I am sending you the tasks from the state matura in physics for the year 2023/2024. Please solve them for me. Give me the solutions in CSV format. In the first column, the solutions should be (A, B, C, D), and in the second column, the explanations.*" Similarly, queries are formulated for extended-response tasks, for example: "*I am sending you the tasks from the state matura in physics for the year 2023/2024. Please solve them for me. Give me the solutions in CSV format. In the first column, the final solution should be, and in the second column, the step-by-step solution with explanation"*.

The answers obtained were structured in a common table, which enabled a systematic comparison between the model and the official solutions and served as a basis for quantitative analysis. The correct answers were taken from the official state matura website. The model answer was compared with the correct answer from the solution library. For extended answer questions, the final solutions were compared. If the procedure is correct, and the final result deviates due to rounding within the allowed tolerance, no points are lost during the assessment.

In some cases, if the result is rounded to more or fewer decimal places than recommended, and the value is within the acceptable range, the correct answer is also recognized. The models were compared based on the percentage of correct answers, with special attention to differences in performance on questions with graphical displays.

In addition to checking the accuracy of solving the tasks, similarities between the models will be analyzed to understand the extent to which their solutions match or differ. This analysis will provide insight into the similarities between the models.

Consistency was also tested by giving the same tasks to the models at different time intervals over several weeks. This was to determine how stable and repeatable the responses were, as inconsistency in responses can significantly reduce their usefulness in an educational context.

Execution time was measured for all tested models to gain insight into the practicality of their real-time application. This metric is especially important for scenarios where quick feedback is needed, such as interactive learning or exam preparation.

The relationship between accuracy and execution time was analysed to determine the relationship between these two factors. Such analysis is crucial for understanding situations where users must choose between a faster but potentially less accurate model and a slower but more precise approach.

QUALITATIVE PART OF THE RESEARCH

A focus group was conducted to gather additional information and gain insight into the teachers' opinions. A focus group is a discussion group that discusses a topic based on appropriate questions that stimulate debate. Its main goal is to identify the participants' perceptions and ideas regarding the given topic [22]. We limited the group to 10 participants because a smaller group allows for greater depth of expression from each member, which is crucial for the quality of the discussion. The focus group members were carefully selected to ensure the group was homogeneous. After the focus group, participants were given the opportunity to further express their opinions and reflections through a follow-up survey.

Participants

This focus group consisted of 10 participants, namely high school physics teachers, two retired. All teachers had more than ten years of experience in school, and most of them had over twenty years of experience. All were employees of public schools in the Republic of Croatia. Participants were carefully selected to be in line with the research objectives. In Table 1, we present some information from the participants of this focus group. All participants were informed about the research objectives, the method of data collection and processing, and the possibility of withdrawing from the research at any time. Participants gave informed consent, and the research was conducted in accordance with ethical guidelines.

Table 1. Profile of teachers who participated in the Focus Group.

Participant	School	Years of experience
P1	Grammar school	20+ years
P2	Grammar school	20+ years
P3	Grammar school	20+ years
P4	Grammar school	11-20 years
P5	Grammar school	11-20 years
P6	Vocational school	11-20 years
P7	Vocational school	11-20 years
P8	Vocational school	11-20 years
P9	Retired (grammar school)	20+ years
P10	Retired (vocational school)	20+ years

Focus Group Organization

The focus group was held online during May 2025 via the Google Meet platform, due to the convenience of the participants being geographically dispersed throughout the Republic of Croatia. Before the meeting itself, the participants received a document with a carefully selected sample of questions from the state matura, as well as the answers and explanations provided by the models. This was done so that the participants could gain insight into the quality of the models' answers and to make the focus group meeting as concrete as possible. After that, a 1-hour meeting was held. The moderator of this focus group, one of the authors of this article, took a discreet approach and a non-directive role. Pre-defined and open-ended questions were asked, minimally interfering in the dialogue during the semi-structured discussion.

The content of the meeting was as follows:

1. Teachers' use of LLMs in teaching.
2. Use of LLMs by students.
3. Accuracy of answers provided by LLMs.
4. Can LLMs be used in preparation for the high school exam and to what extent?
5. The rest and the conclusion.

Method of Conducting Qualitative Analysis

After the focus group meeting, the recorded conversation was analyzed. Transcription and analysis were performed manually by the researcher, without the use of specialized automatic transcription software. The recording was analyzed by authors of the article to ensure multiple perspectives. This ensured greater precision and the ability to record nonverbal elements of communication. The three main dimensions of the analysis are:

- using LLMs in teaching
- accuracy and clarity of answers by LLMs
- recommendations and potential dangers when using the LLM when preparing for the state exam (*matura*) in physics

Along with the focus group, a follow-up survey was also conducted to obtain additional information after the focus group meeting. The data obtained were analyzed and linked to a specific participant and their statements during the meeting. This follow-up survey aimed to allow participants to further reflect on the topics raised during the group discussion in a more relaxed environment, without time pressure, and to express their views, experiences and recommendations regarding the use of AI in physics teaching. Some participants may not have felt comfortable expressing certain views in a group setting or were influenced by the opinions of others. The individual survey allowed for a more honest and independent expression of views. The survey also enabled a systematic collection of responses to specific questions that may have only been partially addressed during the focus group, ensuring that all key research areas were adequately covered. This methodological approach, which combines group discussion with individual surveying, provided a more comprehensive insight into the attitudes and perceptions of physics teachers regarding the use of AI in education.

RESULTS

The key research findings are presented and divided into quantitative and qualitative results in this section.

QUANTITATIVE RESULTS

Model Accuracy

In total, 15 large language models were tested on 122 multiple-choice questions and 57 extended-answer questions. The accuracy results are shown in Table 2. Analysis of the percentage of resolution shows significant differences between the models. The best results on all types of tasks were achieved by ChatGPT o3 and o4-mini, with accuracy above 90% on multiple choice and above 75% on extended response. The results of the ChatGPT o3 model show that this model was the most accurate for both types of questions. The weakest results were recorded with ChatGPT 4o mini (36% and 42%) and Copilot (59% and 35%). The relationship between the percentage of accuracy of the two types of tasks for each model can be seen in Figure 1.

Additionally, tasks of both types that contained graphical representations were analysed. This group of tasks proved to be the most demanding for LLMs, with an average accuracy of 51%. Here again, ChatGPT o3 proved to be the most accurate (73%), and besides it, o4-mini and R1 1776 exceeded 70% accuracy (71%). Here, too, ChatGPT 4o mini (22%) and Copilot (33%) proved to be the worst.

Table 2. Accuracy of LLMs by task type on the physics state exam.

Model	Platform	Multiple choice questions, %	Extended answer questions, %
ChatGPT 4o	ChatGPT (OpenAI)	66	39
ChatGPT o3	ChatGPT (OpenAI)	93	89
ChatGPT o4-mini	ChatGPT (OpenAI)	84	75
ChatGPT o4-mini-high	ChatGPT (OpenAI)	75	84
ChatGPT 4.5	ChatGPT (OpenAI)	75	63
ChatGPT 4o mini	ChatGPT (OpenAI)	36	42
Sonar	Perplexity	75	53
Claude 3.7 Sonet	Perplexity	68	63
GPT-4.1	Perplexity	64	63
Grok 3 Beta	Perplexity	69	49
R1 1776	Perplexity	82	70
o4-mini	Perplexity	93	77
Claude 3.7 Sonet Thinking	Perplexity	76	68
Copilot	Microsoft Copilot	59	35
DeepThink (R1)	Deepseek	77	54
Average		**73**	**62**

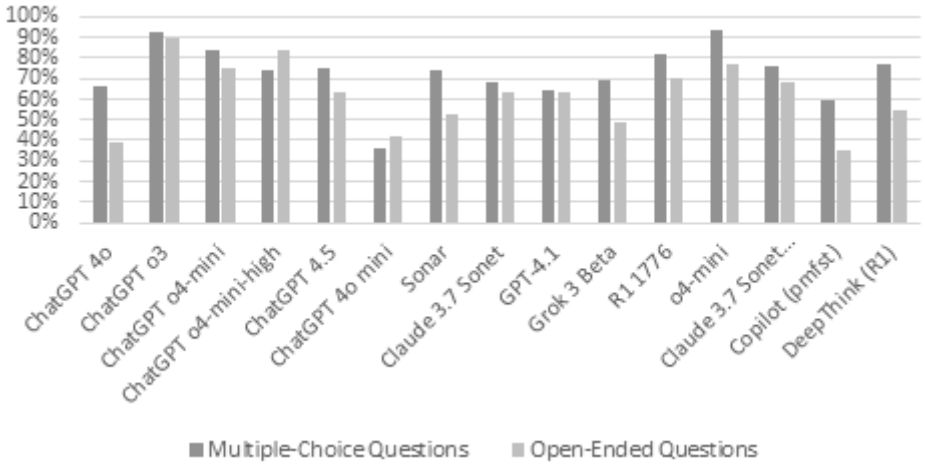

Figure 1. Accuracy of LLMs on multiple-choice and open-ended physics exam questions.

Similarities Between Models

We analyzed the mutual similarity of the answers of 15 models for 122 multiple choice questions. Instead of the classical Pearson correlation, which is not optimal for categorized ordinal answers (A, B, C, D), we used measures that better reflect the actual agreement: percentage of agreement (how often two models give identical answers to the same questions) and Cohen's kappa (a statistical measure of agreement that corrects the effect of chance).

The highest agreement was shown by the ChatGPT o3 and o4-mini models, which gave identical answers in as many as 91,8% of cases, while their Cohen's kappa was 0,89, indicating excellent agreement above chance. This is the only pair that exceeds the 0,8 threshold for strong agreement.

ChatGPT 4o mini model has the lowest average Cohen's kappa value compared to the other models (0.16), indicating a very low level of agreement. This means that the responses of this model differ significantly from the responses of the other models, almost at the level of chance agreement. This value falls into the category of weak agreement.

The value of Cohen's kappa for other models in relation to the others is between 0,4 and 0,8.

Consistency of Model Responses

The consistency of the model was tested on multiple choice tasks. Only non-graphic tasks were considered, since models often do not read visual elements well, which can result in random and variable answers. Also, extended response tasks are not included because, although the models often provide similar solutions, differences in rounding and wording can make objective comparison difficult.

Consistency was tested across multiple time intervals over several weeks, attempting to capture possible variations in model behaviour due to updates or changes in platform infrastructure.

The average consistency of all models and platforms is 86%. The ChatGPT platform (OpenAI) shows the highest average consistency of 93%, which indicates the stability and reliability of their models. The Perplexity platform shows significant variability among models, with an average consistency of 79%, but also with individual models reaching a high level of consistency (e.g. o4-mini – 94%). The lowest consistency was shown by Perplexity's Cluade 3.7 Sonet model of 64%. The percentage of consistency of all models and platforms is visible in Table 3.

Table 3. Response consistency on multiple-choice tasks, excluding questions with graphical content.

Model	Platform	Consistency by model, %	Consistency across platforms, %
ChatGPT 4o	ChatGPT (OpenAI)	86	93
ChatGPT o3	ChatGPT (OpenAI)	99	
ChatGPT o4-mini	ChatGPT (OpenAI)	92	
ChatGPT o4-mini-high	ChatGPT (OpenAI)	90	
ChatGPT 4.5	ChatGPT (OpenAI)	97	
ChatGPT 4o mini	ChatGPT (OpenAI)	93	
Sonar	Perplexity	86	79
Claude 3.7 Sonet	Perplexity	64	
GPT-4.1	Perplexity	72	
Grok 3 Beta	Perplexity	84	
R1 1776	Perplexity	70	
o4-mini	Perplexity	94	
Claude 3.7 Sonet Thinking	Perplexity	85	
Copilot	Microsoft Copilot	84	84
DeepThink (R1)	Deepseek	88	88
Average		86%	86%

Execution Time

The research measured the execution time of each model while solving a particular exam booklet. The time is expressed in seconds and represents the average duration of processing all tasks in a particular category for a particular year. Thus, the average execution time for multiple-choice questions from a single exam booklet is 91 seconds, and for extended-response questions it is 101 seconds. Recall that the average number of multiple-choice questions per exam booklet is 24, and for questions with extended answers it is 11.

The results show a significant difference in execution time between individual models, with some models, e.g. ChatGPT o3 (327 s and 220 s) requiring significantly more time than others, such as Sonar (20 s and 40 s) or Copilot (22 s and 25 s). All results are visible in Table 4.

Table 4. Average execution time per exam.

Model	Platform	Multiple choice questions, s	Extended answer questions, s
ChatGPT 4o	ChatGPT (OpenAI)	25	70
ChatGPT o3	ChatGPT (OpenAI)	327	220
ChatGPT o4-mini	ChatGPT (OpenAI)	106	65
ChatGPT o4-mini-high	ChatGPT (OpenAI)	100	165
ChatGPT 4.5	ChatGPT (OpenAI)	110	155
ChatGPT 4o mini	ChatGPT (OpenAI)	54	100
Sonar	Perplexity	20	40
Claude 3.7 Sonet	Perplexity	90	70
GPT-4.1	Perplexity	20	40
Grok 3 Beta	Perplexity	70	60
R1 1776	Perplexity	35	110
o4-mini	Perplexity	150	200
Claude 3.7 Sonet Thinking	Perplexity	155	125
Copilot	Microsoft Copilot	22	25
DeepThink (R1)	Deepseek	75	75
Average		**91**	**101**

Relationship Between Response Accuracy and Execution Time

Correlation analysis between execution time and model accuracy indicates a moderate to strong positive association. Thus, for multiple-choice tasks, the Pearson correlation coefficient was 0,58 (p = 0,022), which suggests that models with longer processing times generally achieve higher accuracy. For extended response tasks, the correlation was even more pronounced. A value of 0,73 (p = 0,002) indicates an even stronger association in more complex tasks between execution time and accuracy. Both p-values are well below the usual threshold of 0,05.

Additionally, a scatter plot (Figure 2) is presented that illustrates the relationship between model execution time and their accuracy on the tasks of the state matura in physics. The graph clearly shows a positive correlation, suggesting that models with longer processing times generally achieve higher accuracy.

In Table 5 one can see the difference between the platforms in average accuracy and average execution time.

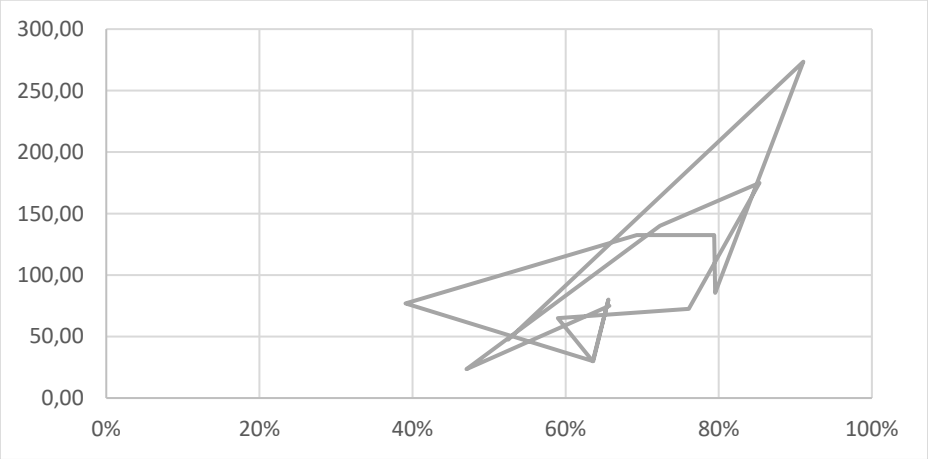

Figure 2. Relationship between execution time and accuracy of LLMs on physics exam tasks.

Table 5. Average accuracy and execution time of language models by platform**.**

Platform	Average accuracy, %	Average execution time, s
ChatGPT (OpenAI)	68	125
Perplexity	69	85
Copilot	47	24
DeepSeek	66	75

QUALITATIVE RESULTS

Based on the transcription of a focus group with ten physics teachers and a subsequent survey, three main thematic units were identified that depict the attitudes and experiences of teachers with large language models in the context of physics teaching and preparation for the state exam.

Using LLMs in Teaching

Most teachers use LLMs primarily as an auxiliary tool in preparing lessons – for generating ideas, creating materials and designing problem tasks. However, the analysis shows polarization among physics teachers in their approach to using LLMs. Some teachers completely avoid using these tools in teaching, preferring traditional methods. Participant P2 clearly emphasized: "*I do not use LLMs in teaching. I use written materials and collections*".

Another group of teachers uses LLMs selectively, mainly for generating ideas and preparing materials, but not directly with students. Participant P1 stated: "*I don't use LLMs with students. I use them if I need an idea... I use Copilot and ChatGPT most often, but mostly Copilot because it's handy in Teams*". Some also use them to evaluate students' work, and participant P6 points out that he uses Perplexity to create evaluation rubrics and check sources, emphasizing its transparency: "*Perplexity lists sources. It approaches the solution through procedures and then provides the solution*". Participant P9, with over 20 years of experience, took a pragmatic stance: "*In regular classes, I would give an assignment and if no one solves it, I would recommend that they solve it together with the help of artificial intelligence*".

Some participants already use tools such as ChatGPT and Microsoft Copilot with their students, but they believe that this should be done with strict supervision and critical evaluation of the solutions obtained. "*Students often know how to use models for copying, without their own understanding, so this requires a different approach to assessing their work*" – warns participant P6. The importance of educating students in a critical approach to the results obtained from LLMs was emphasized, including checking and understanding each step of the solution. "*They should be taught not to automatically accept everything that the model says, but to understand and check each part of the solution"* – agreed the participants in the discussion.

Participants unanimously pointed out that students use LLMs on their own, massively, often without critical reflection. Participant P1 emphasized: "*Students use them massively, much more than teachers. We as teachers need to keep up with the times as soon as possible and somehow teach children to look at them critically*". Teachers easily recognize when students use LLMs because of the specific markings and approaches. Thus, participant P6 warned of a deeper problem: *"They literally copy, i.e. they copy what they see. They do not understand the symbols they copy at all"*. Participant P7 stated: "*I notice that students use the markings given to them by ChatGPT regardless of the fact that we did not use them in class*". Participant P8 supported this with specific examples: "*So they write multiplication as 'x'. We write oscillation for the length of a pendulum as a lowercase 'l', and they write it as a capital 'L'. So you can see from the markings that they did not do it*". Participant P1 adds: "*Students use it massively, much more than teachers. We as teachers need to keep up with the times as soon as possible*".

Perception of Accuracy and Quality of LLM Responses

Some participants noted significant problems with the accuracy and consistency of LLMs when solving physics problems. Participant P5 described her insight: "*I gave ChatGPT a problem to solve. I asked him if he was sure that was the case... It took him three or four attempts to solve the problem correctly, but he didn't do it right the first time*".

Models often produce concise answers that can leave out key parts, which can lead to misunderstandings or errors if not monitored and corrected. Tasks with graphical representations are particularly problematic, as confirmed by the quantitative part of the research. Some participants note that models sometimes change important details, which can be confusing for students. Participant P9 noted: "*One of the tasks given gave the wrong solution because it started solving a task with harmonic oscillation with cosine, but in the exam booklet it is sine*". Participant P1 adds: "*I noticed a lot of errors in the tasks given. They are incomplete and I don't know how much students would learn*".

Participant P3 warned about the importance of critical checking: "*I used it. It makes mistakes. Only when it is pointed out to him about the mistakes will he say 'You're right' and correct it*".

Participant P9 points out: "*He solved 34 out of 35 tasks so accurately with explanations. I think that one was not a good formulation of the task. After warning him what was incorrect, he immediately corrected it and stated 'I immediately noticed where I was wrong' and in the end, he solved it correctly*".

Recommendations for Use in Preparation for the State Matura

Participants recognised numerous advantages of LLMs, such as automated assistance in preparing materials, explaining concepts, and encouraging research-based learning. However, they cautioned that additional education of teachers and students is necessary for successful integration. It was suggested that a hybrid approach be used where LLM serves as an auxiliary tool, but under the constant supervision of the teacher, especially for more complex tasks.

Teachers' views on the suitability of LLMs for preparing for the state matura show caution and recognition of potential. Participant P4 cautioned: "*I think it is a bit questionable to use it to prepare for the state matura... Some main concepts should be mastered for them to use ChatGPT*". Most did not use the models as a primary tool for students to create solutions to problems, but they recognise their potential for such use with adequate education and supervision. They cite asking the model for a step-by-step explanation as the correct approach: "*Write a detailed solution procedure – it allows students to follow the logic*" (participant P9). LLMs can also encourage students to think critically. Participant P5 describes a task in which

students analyse a text generated by ChatGPT and compare it with a textbook: "*So, to critically reflect on it and compare it with other sources*".

The most experienced participant, a retired professor with many years of experience, P9, was more optimistic: "*I think it can be used to prepare for the state matura... I think that students can use it as a great tool to teach them to prepare for the state matura*". However, he also urges caution. When asked, "Does he think that students could rely solely on it?" he thinks not. "*You can't just come off the street and ask someone who has never studied physics at the high school level and tell them to solve one problem. There must always be some knowledge of the student*".

The moderator summed up the discussion with the words: "*LLMs can help, and we have to keep up with the times. They can be useful to a certain extent for both us and the students, but with additional control. As for preparing for the state matura, they can help the students, but they still need to have a certain fund of knowledge to be able to communicate with the LLM in a quality way*".

DISCUSSION

This section will interpret and elaborate on the quantitative and qualitative findings of the research. First, the performance of large language models will be discussed through various metrics such as accuracy, speed of execution, and consistency. Then, teachers' perceptions and attitudes about the application of these models in physics teaching will be commented, based on the results of the focus group and subsequent reflections. The broader implications of our findings for educational practice will be brought out. Finally, limitations of the research that should be kept in mind when interpreting the results will be highlighted and directions for future research will be suggested.

MODEL ANALISYS

The results show that the best LLMs are already quite reliable on the state matura in physics, especially on multiple-choice questions. Xu et al. compared the performance of LLMs on undergraduate-level physics questions. While their best model achieved 49.8% [23], our research shows that several models achieve over 90% accuracy on Croatian high school physics exams. Such findings may mean that today's models have sufficient knowledge for teaching at the high school level, while there is room for improvement at the undergraduate level. We can also compare the results of our research with the findings of another study, which took place two years ago. According to Ding et al. [24], GPT-3.5 achieves 49,3% (zero-shot) and 73,2% (few-shot) accuracy on elementary school physics tasks, showing that the development from GPT-3.5 to the current models (GPT-4, o3) demonstrates rapid progress in LLMs' abilities.

The models are significantly more accurate in multiple-choice questions than in extended-response or especially in graphical tasks. The greatest challenge is in graphical tasks, where most models still have significant difficulties [10, 25]. Current LLMs appear to have limitations in interpreting data from graphical tasks. Also, the models show difficulties in determining values from visual tasks, which is reflected in the resulting solution. Models often ask the user to read data from an image or describe an image. Thus, our findings show significantly lower accuracy on tasks with graphical representations compared to textual tasks. Variability and inconsistency were also observed in such tasks. This indicates that current models are not fully prepared for the independent and reliable solving of problems that require visual interpretation. Therefore, a human factor is needed to check and control the interaction when it comes to graphical tasks. In an educational context, this calls for additional caution and supervision by teachers.

The obtained results of agreement between the models provide indications of similarities and differences in providing solutions to multiple-choice questions. The largest Cohen's kappa obtained between the ChatGPT o3 and o4-mini models shows their considerable agreement in answers, which may suggest that they have similar architectures, approaches or training data. On the other hand, the ChatGPT 4o mini model shows a low level of agreement with the other models, which indicates that this model probably makes decisions in a different way. Such a model can be useful when diversifying opinions or obtaining alternative solutions for creative tasks is needed. It can also indicate frequent errors. From the obtained findings, it is evident that models with high agreement achieve better accuracy (e.g. ChatGPT o3 and o4-mini) while models that agree less with the others show a low level of accuracy. In general, most models show moderate to good agreement, which suggests that despite the differences, there is a certain consensus among the models. Of course, the context and purpose of using an individual model should be considered, as each individual model may be more suitable for a different context or type of task.

Consistency analysis adds strength to the answers obtained and reduces the randomness factor, especially in multiple-choice tasks. It has been observed that identical queries can result in different answers, even in the same model session [26]. Novikova et al. [27] warn that models often fail to maintain consistent answers. The high consistency of most models in our study (86% average) suggests that LLMs have become sufficiently stable and reliable to consistently solve standardised physics problems. However, there is room for improvement in terms of even greater consistency. However, significant differences between the consistency of models (range 64% to 99%) highlight the importance of careful selection of models for educational purposes. Consistency of responses, with accuracy, is a key factor for the reliability of LLMs in an educational context. The results show that models within the same platform can be quite heterogeneous in terms of consistency, and therefore, certain unstable models would need additional validation before being used for educational or evaluation purposes.

The results indicate that models that provide more accurate solutions for more complex and demanding tasks also have longer execution times, which may be a consequence of more complex computational processes within the model. Therefore, when choosing a model for a particular task, it is necessary to consider the trade-off between the desired accuracy and acceptable computation time, especially in tasks where response speed plays an important role. However, it should be emphasised that execution time is not the only factor and that there are deviations in this regard. It is not a rule that a slower model is always more accurate, as our analysis shows variability and models with similar times can have different accuracies.

TEACHERS' PERCEPTION

The focus group analysis reveals several important insights. Participants agree on the need to educate students on the critical use of LLMs. They should be introduced to teacher supervision [24], especially in preparation for the state matura. They agree that state matura preparation should teach students to think, not just solve numerical problems. They believe preparing for the state matura should teach students to think, not just solve numerical problems. The consensus among them is that LLMs are not a substitute for teachers, but a tool that requires supervision. Although our teachers' views showed some caution, an example of a structured teaching system, Physics-STAR, that uses a personalized physics learning framework shows a 100% improvement in student performance on complex physics tasks [28]. This may mean that the key difference lies in the implementation approach, where pedagogically structured and informed systems are contrasted with the unsupervised use by students that our teachers have observed. Students have been reported to use LLMs, often without thinking, skipping steps in solutions or not understanding the symbols they copy. Some students rely entirely on the given answers without further checking. Therefore, students need

to be educated on how to recognise and correct errors in the generated answers. Students and teachers need to be aware of the limitations [6]. Although some models can provide clarity in their solutions, clarity still does not guarantee accuracy [13], hence the need for verification and control in their educational application. From their own experience, teachers noted variability among different models, which corresponds to quantitative research findings. In addition to the incorrect answers provided by the models, problems such as superficial explanations were also mentioned.

Wu et al. [13] findings provide implications for the debate about trust in LLMs and the reliability of reasoning in education. They showed that chain reasoning can negatively affect accuracy, which supports caution about interpretable but unreliable responses. Their work, as well as the experiences of our teachers, further confirms that clarity does not guarantee accuracy, justifying the need for verification and control in the educational application of LLMs, especially in disciplines that require logical reasoning, such as physics.

IMPLICATIONS

The results show that LLMs can be a valuable tool for teachers in lesson preparation and problem generation. The best models are already helpful in preparation for the state matura in physics. Although LLMs show relatively high accuracy (e.g. ChatGPT o3 over 90% on multiple choice), graph tasks pose a challenge (average accuracy 51%). The focus group recognises the usefulness of the models for students with some knowledge and supervision: "*The model can be useful, but students need to have basic knowledge to use it correctly*" (P9).

Recommendation 1: Preparation for the state exam in physics can include an LLM as an assistant. It is essential to teach students AI literacy, including recognizing hallucinations and critically evaluating the reliability of responses.

Using these tools requires technological and pedagogical literacy. Teachers who have used them report that they are useful "*for ideas, concepts, but also for creating rubrics*", but under the condition of "*critical supervision and verification of sources*" (P6).

Recommendation 2: Educating teachers about AI tools and assessing the quality of AI responses should become an integral part of the professional development of physics teachers.

The findings point to the need to introduce pedagogical scenarios for the use of artificial intelligence in schools, including curriculum guidelines, evaluation protocols and ethical frameworks.

Recommendation 3: Develop guidelines that include the responsible use of LLMs in schools.

LIMITATIONS

The study analysed responses from 15 models across four platforms. Given the rapid advancement of the models, the results are primarily valid for the state of technology at the time of the study. They should be interpreted with caution after some time. Changes in performance and model updates over time can lead to instability in the findings and should be monitored longitudinally [29, 30]. On the positive side, these results should trend upward as technology advances.

The research was based exclusively on physics tasks for the 2019-2024 state maturas, which include standardised test items. Although the sample consists of mechanics, thermodynamics, electromagnetism, optics, and modern physics, the focus was on tasks with a unique final solution, which did not assess the role of LLMs in interpretive and ambiguous tasks.

While the participating teachers provided valuable insights through focus groups, students did not directly participate in the research. This led to a lack of qualitative insights from the perspectives of the most sensitive users of the models in education, particularly on the reasons

for relying on LLMs, barriers to use, and perceptions of accuracy. It is planned to be part of a future study.

CONCLUSION

Traditional education faces challenges such as individual differences among students and assessment of teaching effectiveness. Research on educational models is constantly evolving, providing new methods and approaches to achieve the goals of personalised learning, intelligent teaching, and academic assessment [31]. The focus group results confirm that LLMs are, somewhat informally, already integrated into the educational process. The focus group shows that while LLMs bring significant opportunities to improve physics teaching, there are important challenges related to the accuracy, consistency and understanding of the generated solutions. Accuracy and consistency over time are crucial when integrating LLMs into sensitive fields such as educational or clinical [30]. Their successful implementation requires a thoughtful approach, ongoing education of all stakeholders, and expert oversight. LLMs can be additional learning tools, but teachers must emphasise critical thinking and source checking [30]. It is crucial to develop such responsible education systems where critical evaluation methods and effective use of these tools in teaching will be of utmost importance. The advancement of AI in education requires initiatives to address the ethics and privacy issues of AI and requires interdisciplinary and transdisciplinary collaboration [32]. A report on the security risks and biases of LLMs [9] warrants caution in their use, particularly in contexts that require accuracy and ethical neutrality. These findings demonstrate the potential but also support the need for careful and thoughtful integration of LLMs into physics teaching.

REFERENCES

[1] Wang, S., et al.: *Large Language Models for Education: A Survey and Outlook*.
preprint arXiv:2403.18105,
http://dx.doi.org/10.48550/arXiv.2403.18105,

[2] Kasneci, E., et al.: *ChatGPT for good? On opportunities and challenges of large language models for education*.
Learning and Individual Differences **103**, No. 102274, 2023,
http://dx.doi.org/10.1016/j.lindif.2023.102274,

[3] Huang, J. and Li, S.: *Opportunities and Challenges in the Application of ChatGPT in Foreign Language Teaching*.
International Journal of Education and Social Science Research **6**(4), 75-89, 2023,
http://dx.doi.org/10.37500/ijessr.2023.6406,

[4] Jalilova, S. and Musayeva, G.: *Artificial Intelligence In Physics Teaching.*
http://dx.doi.org/10.5281/ZENODO.14744940,

[5] Xu, H.; Gan, W.; Qi, Z.; Wu, J. and Yu, P.S.: *Large Language Models for Education: A Survey*.
preprint arXiv:2405.13001,
http://dx.doi.org/10.48550/arXiv.2405.13001,

[6] Peláez-Sánchez, I.C.; Velarde-Camaqui, D. and Glasserman-Morales, L.D.: *The impact of large language models on higher education: exploring the connection between AI and Education 4.0*.
Frontiers in Education **9**, No. 1392091, 2024,
http://dx.doi.org/10.3389/feduc.2024.1392091,

[7] Gabbay, H. and Cohen, A.: *Combining LLM-Generated and Test-Based Feedback in a MOOC for Programming*.
In: Joyner, D.; Kim, M.K.; Wang, X. and Xia, M., eds.: *Proceedings of the Eleventh ACM Conference on Learning @ Scale*. ACM, New York, pp.177-187, 2024,
http://dx.doi.org/10.1145/3657604.3662040,

[8] Keat, K., et al.: *PGxQA: A Resource for Evaluating LLM Performance for Pharmacogenomic QA Tasks*.
In: Biocomputing 2025. World Scientific, Kohala Coast, pp.229-246, 2024,
http://dx.doi.org/10.1142/9789819807024_0017,

[9] Huang, R.S., et al.: *Assessment of Resident and AI Chatbot Performance on the University of Toronto Family Medicine Residency Progress Test: Comparative Study*.
JMIR Medical Education **9**, No. e50514, 2023,
http://dx.doi.org/10.2196/50514,

[10] Maitland, A.; Fowkes, R. and Maitland, S.: *Can ChatGPT pass the MRCP (UK) written examinations? Analysis of performance and errors using a clinical decision-reasoning framework*.
BMJ Open **14**(3), No. e080558, 2024,
http://dx.doi.org/10.1136/bmjopen-2023-080558,

[11] Bitzenbauer, P.: *ChatGPT in physics education: A pilot study on easy-to-implement activities*.
Contemporary Education Technology **15**(3), No. ep430, 2023,
http://dx.doi.org/10.30935/cedtech/13176,

[12] Rong, Y.; Du, T.; Li, R. and Bao, W.: *Integrating LLM-based code optimization with human-like exclusionary reasoning for computational education*.
Journal of King Saud University Computer and Informatino Sciences **37**, No. 87, 2025,
http://dx.doi.org/10.1007/s44443-025-00074-7,

[13] Wu, J., et al.: *Large language models leverage external knowledge to extend clinical insight beyond language boundaries*.
Journal of the American Medical Informatics Association **31**(9), 2054-2064, 2024,
http://dx.doi.org/10.1093/jamia/ocae079,

[14] Sonkar, S.; Liu, N. and Baraniuk, R.: *Student Data Paradox and Curious Case of Single Student-Tutor Model: Regressive Side Effects of Training LLMs for Personalized Learning*.
In: *Findings of the Association for Computational Linguistics* EMNLP 2024. Association for Computational Linguistics, Miami, pp.15543-15553, 2024,
http://dx.doi.org/10.18653/v1/2024.findings-emnlp.912,

[15] Chang, Y., et al.: *A Survey on Evaluation of Large Language Models*.
preprint arXiv:2307.03109,
http://dx.doi.org/10.48550/arXiv.2307.03109,

[16] Jiang, Q.; Gao, Z. and Karniadakis, G.E.: *DeepSeek vs. ChatGPT vs. Claude: A comparative study for scientific computing and scientific machine learning tasks*.
Theoretical and Applied Mechanics Letters **15**(3), No. 100583, 2025,
http://dx.doi.org/10.1016/j.taml.2025.100583,

[17] Krupp, L., et al.: *Challenges and Opportunities of Moderating Usage of Large Language Models in Education*.
preprint arXiv:2312.14969,
http://dx.doi.org/10.48550/arXiv.2312.14969,

[18] Chiu, T.K.F.; Ahmad, Z.; Ismailov, M. and Sanusi, I.T.: *What are artificial intelligence literacy and competency? A comprehensive framework to support them*.
Computers and Education Open **6**, No. 100171, 2024,
http://dx.doi.org/10.1016/j.caeo.2024.100171,

[19] Getenet, S.: *Pre-service teachers and ChatGPT in multistrategy problem-solving: Implications for mathematics teaching in primary schools*.
International Electronic Journal of Mathematics Education **19**(1), No. em0766, 2024,
http://dx.doi.org/10.29333/iejme/14141,

[20] Liang, Y.; Zou, D.; Xie, H. and Wang, F.L.: *Exploring the potential of using ChatGPT in physics education*.
Smart Learning Environonments **10**, No. 52, 2023,
http://dx.doi.org/10.1186/s40561-023-00273-7,

[21] Wang, J.; Xiao, R. and Tseng, Y.-J.: *Generating AI Literacy MCQs: A Multi-Agent LLM Approach*.
In: Proceedings of the 56th ACM Technical Symposium on Computer Science Education V.2. ACM, pp.1651-1652, 2025,
http://dx.doi.org/10.1145/3641555.3705189,

[22] Rezende Junior, M.F. and López-Simó, V.: *What are the perceptions of physics teachers in Brazil about ChatGPT in school activities*?
Journal of Physics: Conference Series **2693**, No. 012011, 2024,
http://dx.doi.org/10.1088/1742-6596/2693/1/012011,

[23] Xu, X., et al.: *UGPhysics: A Comprehensive Benchmark for Undergraduate Physics Reasoning with Large Language Models*.
preprint arXiv:2502.00334,
http://dx.doi.org/10.48550/arXiv.2502.00334,

[24] Ding, J.; Cen, Y. and Wei, X.: *Using Large Language Model to Solve and Explain Physics Word Problems Approaching Human Level*.
preprint arXiv:2309.08182,
http://dx.doi.org/10.48550/arXiv.2309.08182,

[25] Dinh, T.A., et al.: *SciEx: Benchmarking Large Language Models on Scientific Exams with Human Expert Grading and Automatic Grading*.
In: *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Miami, pp.11592-11610, 2024,
http://dx.doi.org/10.18653/v1/2024.emnlp-main.647,

[26] Seo, H., et al.: *Large Language Models as Evaluators in Education: Verification of Feedback Consistency and Accuracy*.
Applied Sciences **15**(2), No. 671, 2025,
http://dx.doi.org/10.3390/app15020671,

[27] Novikova, J.; Anderson, C.; Blili-Hamelin, B.; Rosati, D. and Majumdar, S.: *Consistency in Language Models: Current Landscape, Challenges, and Future Directions*.
preprint arXiv:2505.00268,
http://dx.doi.org/10.48550/arXiv.2505.00268,

[28] Jiang, Z. and Jiang, M.: *Beyond Answers: Large Language Model-Powered Tutoring System in Physics Education for Deep Learning and Precise Understanding*.
preprint arXiv:2406.10934,
http://dx.doi.org/10.48550/arXiv.2406.10934,

[29] Cambaz, D. and Zhang, X.: *Use of AI-driven Code Generation Models in Teaching and Learning Programming: a Systematic Literature Review*.
In: *Proceedings of the 55th ACM Technical Symposium on Computer Science Education V.1*. ACM, Portland pp.172-178, 2024,
http://dx.doi.org/10.1145/3626252.3630958,

[30] Gupta, M.; Virostko, J. and Kaufmann, C.: *Large language models in radiology: Fluctuating performance and decreasing discordance over time*.
European Journal of Radiology **182**, No. 111842, 2025,
http://dx.doi.org/10.1016/j.ejrad.2024.111842,

[31] Gan, W.; Qi, Z.; Wu, J. and Lin, J.C.-W.: *Large Language Models in Education: Vision and Opportunities*.
preprint arXiv:2311.13160,
http://dx.doi.org/10.48550/arXiv.2311.13160,

[32] Zhang, K. and Aslan, A.B.: *AI technologies for education: Recent research & future directions*.
Computers and Education: Artificial Intelligence **2**, No. 100025, 2021,
http://dx.doi.org/10.1016/j.caeai.2021.100025.