XXX Conference on Industrial Systems Engineering
"Francesco Turco" Summer School - Lecce (Italy) 10-12 Sept 2025

Smarter than a student? Testing AI performance in an engineering exam

Fantozzi I.C.*, Martuscelli L.*, Schiraldi M.M.*

** Department of Enterprise Engineering, University of Rome "Tor Vergata", Via del Politecnico 1- Rome – Italy*
(italo.cesidio.fantozzi@uniroma2.it ; luca.martuscelli@alumni.uniroma2.eu ; schiraldi@uniroma2.it)

Abstract: This study evaluates the performance of an artificial intelligence (AI) system trained on audio recordings of an engineering course at university. The final objective of the research is to verify the reliability of the AI in the role of teaching assistant, correctly answering students' questions; to this extent, the first step has been to analyze the AI's ability to pass the final exam of the course. The originality of this research lies in the fact that the AI was trained using only the audio recordings of the lectures given in the classroom by the lecturer, without supplementing them with textbooks, notes, or slides. This ensures that the knowledge-base on which the system works is absolutely and exactly consistent with the training objective of the course. The AI was then subjected to the exact same examination that the students faced at the end of the course, i.e. multiple choice tests. A key aspect in the system testing is that the AI was given the opportunity to reconsider all of its answers, mimicking the real exam conditions experienced by students. This allowed for a detailed analysis of the AI's accuracy rate as well as its decision-making process. This approach provides deeper insights into the AI's ability to reassess its choices and refine its understanding, further simulating human-like learning behavior. The findings contribute to the ongoing discussion on AI-assisted learning in higher education, offering valuable insights into the potential and limitations of these tools in supporting traditional teaching methodologies and assessment strategies in engineering education.

Keywords: Artificial Intelligence, Engineering Education, Manufacturing Systems Engineering, University 4.0, Google NotebookLM

1. Introduction

The integration of Large Language Models (LLMs) into higher education has sparked increasing interest in their potential to support or reshape traditional approaches to teaching, learning, and assessment. As artificial intelligence (AI) technologies continue to evolve, their application within academic settings is extending beyond conventional uses such as automated grading or tutoring, toward more sophisticated roles that involve direct interaction with course content and real-time student engagement. A particularly promising yet underexplored area involves the deployment of LLMs trained on course-specific instructional materials to support learning in domain-intensive fields such as engineering. This study investigates the extent to which an LLM, trained exclusively on the audio recordings of university lectures, can act as a virtual teaching assistant by accurately answering exam-style questions derived from the course content. The originality of this approach lies in its strict adherence to the naturalistic training environment: the model's knowledge is drawn solely from what was communicated during classroom instruction, without the inclusion of textbooks, notes, slides or other auxiliary resources. This design choice was motivated by the intention to replicate a purely naturalistic learning environment, where the AI receives only the same auditory information as a student attending the lectures. While multimodal learning may offer benefits, isolating the audio dimension allows us to evaluate its standalone potential and eliminates biases introduced by external materials such as structured slides or textbooks. Moreover, spoken lectures often include pedagogical nuances—emphasis, repetition, and elaboration—not captured in written documents, which may help AI systems simulate student comprehension more realistically. This ensures that the model's knowledge base is precisely aligned with the pedagogical intent of the course. To evaluate the LLM's performance and reliability, the model was subjected to the same assessment protocol as enrolled students—namely, multiple-choice exams—under controlled conditions that allowed it to revisit and revise its responses, mirroring real examination settings. This framework enabled an in-depth analysis not only of the model's accuracy but also of its iterative decision-making process, offering valuable insights into its capacity to emulate human-like learning behaviors such as self-correction and refinement. While previous studies have demonstrated that LLMs like ChatGPT can achieve passing scores in various standardized assessments—including the United States Medical Licensing Examination (USMLE) and the Fundamentals of Engineering (FE) exam —these often rely on general-purpose models with broad training corpora(Kung et al., 2023; Pursnani et al., 2023). By contrast, the present study focuses on a closed, domain-specific training context, thereby allowing for a more granular evaluation of content fidelity, contextual accuracy, and alignment with instructional goals. Complementary research also points to the growing use of LLMs as interactive learning aids, further reinforcing their potential role in personalized and adaptive education (Sánchez-Vera, 2025).

Nevertheless, concerns remain regarding the consistency and depth of reasoning that these models can demonstrate in complex or ambiguous scenarios. Previous work has

highlighted challenges such as factual inaccuracies, shallow logical inference, and limitations in multi-step problem solving (Jafarian and Kramer, 2025; Wei, 2025). These limitations are particularly relevant in engineering education, where the ability to model systems, apply abstract reasoning, and handle stochastic variability is essential. This study contributes to the broader discourse on AI-assisted learning by providing empirical evidence from a controlled, course-specific experiment. The findings aim to clarify the conditions under which LLMs can be integrated as meaningful instructional agents in higher education.

2. Literature background

The ongoing development of Large Language Models (LLMs) represents one of the most transformative shifts in the digitalization of education. From early rule-based systems to contemporary transformer-based architectures, these AI systems—exemplified by OpenAI's GPT series and Google's Gemini—have achieved unprecedented linguistic and cognitive capabilities. Through deep learning and self-attention mechanisms, LLMs are now able to generate human-like text, perform context-aware reasoning, and adapt across diverse knowledge domains. These advancements have enabled their application in a growing range of educational tasks, including content generation, real-time tutoring, automated assessment, and research assistance. What sets modern LLMs apart from previous educational technologies is their flexibility in adapting to unstructured inputs and providing nuanced responses, making them increasingly considered as intelligent collaborators rather than mere tools. Empirical studies, such as those by Kung et al. (2023) and Sánchez-Vera (2025) underscore this potential, showing that LLMs can effectively engage with knowledge-intensive exams and enhance student learning via interactive feedback mechanisms.

Despite their impressive capabilities, the performance of LLMs in formal evaluation settings remains variable and context-dependent. When tested under standardized conditions, LLMs have demonstrated a dichotomy between competence in factual recall and inconsistency in deep reasoning. For instance, Terwiesch (2023) reported that ChatGPT performed well in MBA-level analytical tasks but faltered in stochastic or multi-layered problem-solving. Similarly, Wei (2025) identified performance disparities in LLM responses depending on question type in medical board exams, and Pursnani et al. (2023) noted a significant drop in performance in open-ended engineering questions.

These inconsistencies raise important concerns about the role of LLMs in academic evaluation. One of the most pressing issues is the tendency of these models to produce convincing but factually incorrect outputs—commonly referred to as "hallucinations." In high-stakes academic contexts, where precision and trustworthiness are critical, such limitations call for cautious and context-specific deployment of LLMs.

Beyond their technical performance, the integration of LLMs into education raises critical ethical and pedagogical concerns(Bordt and von Luxburg, 2023). The ability of these models to generate high-quality responses has led to debates about academic integrity, with some scholars warning against the risks of AI-facilitated plagiarism and over-reliance on automated learning tools (Bae et al., 2024; Briatore et al., 2025). According to recent studies, students who excessively depend on LLM-generated content may experience a decline in independent problem-solving skills, as AI-driven assistance reduces cognitive effort in tackling complex tasks (Antaki et al., 2023; Hayes et al., 2024). Furthermore, issues of bias and fairness continue to pose challenges in AI-assisted assessments. Research has demonstrated that LLMs, trained primarily on publicly available datasets, may inadvertently reinforce existing biases in educational content. This raises concerns regarding equity in assessment outcomes, particularly for students from underrepresented backgrounds. As such, educators and policymakers must establish guidelines for the responsible use of LLMs in academia, ensuring that these technologies enhance learning without compromising fundamental educational principles(Gilson et al., 2023).

While LLMs have shown promise in augmenting educational experiences, their long-term role in academia remains an area of active research. Future advancements in AI alignment, domain-specific fine-tuning, and prompt engineering may help bridge existing gaps in reasoning and contextual understanding. Additionally, integrating LLMs with adaptive learning platforms could enable more personalized and responsive educational experiences, tailoring AI assistance to individual student needs. As the field evolves, ongoing interdisciplinary collaboration between AI researchers, educators, and policymakers will be essential in shaping the responsible deployment of LLMs in education. The challenge lies not only in refining these models for improved accuracy but also in fostering an academic culture that prioritizes critical thinking, ethical AI use, and equitable access to learning resources.

3. Methodology

This work aims to investigate how effectively a language model can respond when trained on information derived from a university-level course. While other studies have explored the impact of domain-specific fine-tuning on large language models within educational settings—such as the work by Gilson et al. (2023) this study represents a novel contribution by examining, for the first time, how an LLM performs when trained exclusively on audio recordings. Furthermore, it explores this within the context of highly technical subject matter, such as that addressed in the Industrial Plants course.

For this study, we utilized Google Notebook LM (henceforth, NLM), a state-of-the-art natural language model designed for complex linguistic tasks. The model was specifically trained on the lessons of the Manufacturing Systems Engineering course, part of the Bachelor's degree program in Management Engineering at a leading Italian University. NLM was provided with the audio recordings of the lectures from two different academic years, specifically 2020 and 2021, each course consisting of about 60 hours of lectures. It is important to emphasise that the course syllabus remained the same over the two years, so - apart from the natural variability of the lecturer's oral presentation - it can be assumed that each lecture was exactly replicated two times in the different academic years, for a grand total of approximately 120 hours of recordings.

The decision to use the recordings of these two academic years lies in the fact that these lessons were delivered remotely, through Microsoft Teams, given the permanence of the bans imposed by the COVID pandemic. This meant that the recordings from these two academic years were those with the best audio quality. In any case, the choice of using the recordings from two academic years lies also in the aim of avoiding the risk that some concepts would be difficult for the AI to understand for reasons of audio quality and speech-to-text transcription capabilities. Furthermore, again due to COVID restrictions, in those two academic years the final exam of the course was exceptionally turned into a test as multiple-choice questions. The comparability of the results on a test with multiple-choice questions is extremely easier than on the resolution of a business case, which is what the examination of this specific course consisted of before 2020 and after 2022. The type and structure of the exam has also remained the same in all the 13 exam sessions of the two academic years.

Five tests on the program of the Manufacturing Systems Engineering course were given to NLM, each consisting of different 50 multiple-choice questions (4 answer options), divided into categories according to the following Table 1.

Table 1 Categorization of topics covered

Category	Number of questions
Compressed Air	2
Electrical Systems	2
Water Systems	3
Material Handling	4
Pneumatic Systems	2
Thermal Systems	2
Warehousing	6
Cost-Volume-Profit Analysis	5
Plant Layout	2
Little's Law	2
Overall Equipment Effectiveness	5
Production Capacity Calculations	5
Production Capacity Theory	5
Production Company Classification	5

Five tests were administered to the AI because this was the maximum number of tests that allowed for the total differentiation of all 50 questions, based on the total size of the question database from which the questions for the construction of the students' examinations are randomly extracted. Large Language Models (LLMs) are inherently sensitive to the way prompts are formulated, as their performance can significantly vary depending on the structure, clarity, and contextual framing of the input. In this study, the tests were administered to NLM individually, with the interface refreshed after each session to avoid contextual interference. Each test contained questions grouped in sets of five, and before each group, a consistent prompt was introduced: "Think before answering and tell me the correct answer for each of these questions." This prompt was designed to encourage the model to reason through the questions more deliberately. Afterward, the same questions were presented again in a "second-round", asking the software to engage in deeper reasoning using the prompt: "Think carefully and base your answer on the information from the audio files in your archive before responding:"

This second round was conducted for all questions, regardless of whether the first-round answer was correct, This additional instruction aimed to simulate a more deliberate reasoning process. However, it may have also introduced unintended complexity, causing the model to second-guess its prior responses. A follow-up study might test whether using the reflective prompt in the first round changes overall performance or merely increases uncertainty. in order to analyze any change in response — whether from incorrect to correct (desirable) or, conversely, from correct to incorrect.

4. Results

The test results were analyzed using the same evaluation criteria as the student exams, namely:

- 1 point for each correct answer;
- 0 points for unanswered questions;
- -0.25 points for each incorrect answer.

The score out of 50 was then converted into a grade out of 30, the standard grading system in Italian universities, using a simple proportion. In several instances, the model opted not to provide a response. This was primarily due to the model indicating insufficient information within the provided sources; in other cases, it reported that none of the answer choices were correct. In all such cases, the corresponding question was treated as unanswered.

To benchmark the model's performance, data from 13 exam sessions were employed. The dataset includes the grades of 553 students A summary of these data is presented in the table below.

Table 2 Overall results from students' marks

Number of students	Mean	St. Dev	Min	Q1	Q2	Q3	Max
553	17.44	5.17	0.90	14.55	18.15	21	29.25

The test results achieved by the AI across the five administered tests are shown in the following table 3:

Table 3 Results obtained by the AI model across five multiple-choice tests – first round

Test No.	Correct Answers	Incorrect Answers	Unanswered Questions	Grade
1	37	11	2	20.55
2	38	10	2	21.30
3	40	8	2	22.80
4	39	10	1	21.90
5	39	8	3	22.20

It is important to note that in these tests all questions were different, i.e. there was no repetition of questions between the different tests. The box plot presented in Figure 1 illustrates a comparative analysis between the cumulative performance of students and the aggregate results obtained by the AI model across five administered tests. The distribution of student grades exhibits a wider variability, including several outliers, as well as a lower median and

interquartile range. In contrast, the AI's results are more concentrated, indicating higher consistency and overall performance, with all scores positioned in the upper quartile of the student distribution. This visual comparison highlights the model's ability to achieve strong and stable outcomes across technical multiple-choice assessments, despite the fact that AI does not perform as well as the best students. In fact, the top 13.9% of students exceed the maximum mark achieved by AI, i.e. 22.8/30, while 69.8% got a grade lower than NLM's worst performance 20.55/30. This discrepancy can be attributed to the AI's difficulty in handling abstract reasoning and problem-solving tasks that go beyond factual recall—skills that top-performing students typically excel in. It may also reflect the model's lack of experiential learning, which plays a key role in human understanding and decision-making during exams.

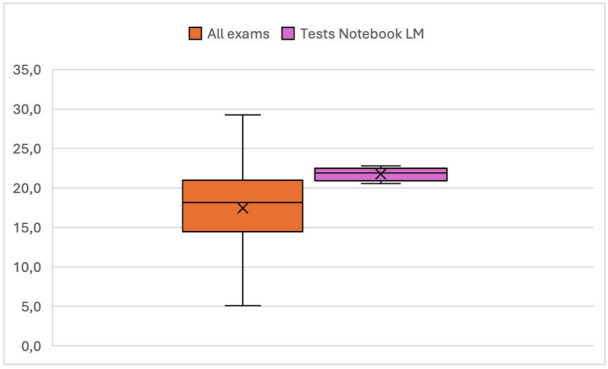

Figure 1: Comparison between student grades and the AI model performance.

Subsequently, the results obtained in the "second round" were analyzed, as shown in the following table 4:

Table 4 Results obtained by the AI model across five multiple-choice tests – second round

Test n.	Correct Answers	Incorrect Answers	Unanswered Questions	Grade
1	31	10	9	17.1
2	35	10	5	19.50
3	37	9	4	20.85
4	37	9	4	20.85
5	40	9	3	22.65

A comparison between the grades obtained in the first and second rounds of testing reveals a consistent decline in performance, with the exception of Test 5, where the second-round score slightly surpasses the first. The average grade in the second round is 20.19, showing a reduction of 1.56 points compared to the first round, and the sample standard deviation increases to 2.06, indicating greater variability. This result is somewhat counterintuitive, considering that the second round was explicitly designed to encourage more structured and reflective reasoning. This trend is clearly illustrated in the following figure 2 and figure 3. The bar chart highlights the score differences across each test, where the purple bars (first round) consistently exceed the grey bars (second round), except in the final test. The accompanying box plot further reinforces this pattern, showing a shift in the distribution of grades. While the first round is characterized by higher central values and less dispersion, the second round exhibits a lower median and greater variability. Together, these visualizations support the conclusion that prompting the model to reason more deeply did not lead to improved performance—in fact, the results became slightly more inconsistent.

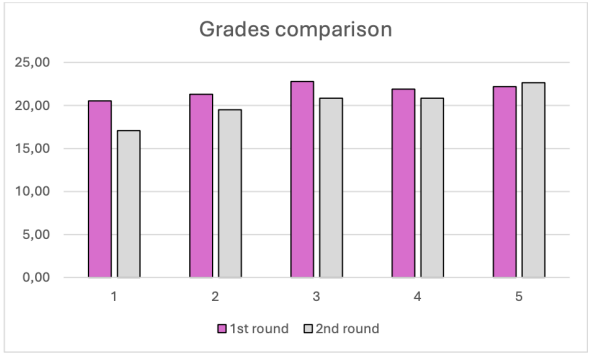

Figure 2 Comparison of the grades obtained by the AI in the first and second round across five tests.

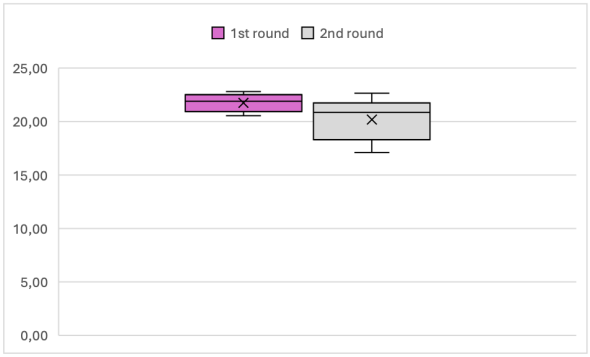

Figure 3 Comparison of the grade distributions for the first and second round.

5. Discussion

This methodology sets the study apart from other research works with similar goals and contributes to the emerging literature on LLMs trained on dynamic and spoken inputs. Moreover, it raises new questions about how training modality (audio vs. text) might impact the quality of AI comprehension, retention, and response reliability in academic contexts.

Another strength of the study is its application to a technical domain such as engineering, which has seen limited exploration in LLM testing. Many AI applications in education have focused on general topics, soft sciences, or language tasks. This research, however, applied rigorous testing to a domain characterized by numerical reasoning, industrial systems theory, and domain-specific terminology. The fact that the AI consistently passed all five tests with average scores well above the median of the human benchmark sample speaks to its robust recall abilities and structured reasoning performance, especially when dealing with theoretical or concept-based content.

However, the study is not without limitations. A primary challenge lies in the limited scope of test administration. Only five tests were administered, and while each consisted of 50 questions, the total number of samples (especially in the second round of testing) remains modest. Similarly, when comparing the AI results with those of the students, it is important to note that a proportion of the students

certainly took the exam more than once in the 13 sessions. In this analysis, each exam was treated as a single test, whereas in reality this was not the case. More interesting results could be obtained by nominally distinguishing the students and measuring only the performance on their first attempt, as well as differentiating the results for the 13 examination sessions of the two academic years.

A critical element highlighted by the study is the inconsistency in AI behavior between test rounds. Despite using the same prompts, the model often provided different answers to the same questions in the second round—sometimes downgrading from a correct answer to a blank or incorrect one. In fact, when asked different questions about the same concept, the system sometimes provided contradictory answers, indicating variability in its internal reasoning or interpretation of the training data. Furthermore, in several instances, the AI alternated between affirming and denying the presence of certain concepts within the provided sources—even when responding to the same question at different times. These inconsistencies highlight challenges in response reliability and semantic coherence, especially when dealing with nuanced or technical topics. This reveals a lack of conceptual consistency and suggests a challenge in how the model balances confidence and source attribution when prompted differently. This performance drop might stem from the increased cognitive load imposed by the reflective prompt, which asked the model to filter its responses through the specific constraint of the lecture archive. This additional step may have limited the model's generative confidence, prompting it to leave more questions unanswered rather than risk hallucination or misalignment. Moreover, even though the second round was designed to invoke deeper reasoning, results show that scores actually declined (with the exception of one test), and variability increased. This counterintuitive outcome—likely due to the model's decision to withhold answers rather than risk incorrect ones— suggests a potential limitation in reasoning under constraint, although the current design does not allow definitive conclusions about the effects of using restricted sources alone. Comparative studies with multiple input types are needed to substantiate this hypothesis

6. Conclusion

This study explored the potential and limitations of large language models (LLMs) trained on lecture audio recordings in supporting academic learning and assessment within a technical university context. By evaluating the performance of an AI model trained exclusively on content from the Manufacturing Systems Engineering course in the Management Engineering program at a leading Italian University, we have highlighted key insights into its capabilities in replicating student performance on structured examination tasks.

The results indicate that while the AI system is capable of achieving consistent and reliable performance—often comparable to the average student—it consistently falls short of matching the top-performing human students. This underscores the continuing value of human reasoning and critical thinking in navigating complex, technical subject matter. Nevertheless, the AI's ability to handle structured knowledge-based tasks, respond to a variety of prompts, and maintain a uniform level of performance makes it a promising supplementary tool for educational environments.

Importantly, this study provides preliminary evidence that audio-based training—without structured textual input—can serve as a potentially viable method for instructing LLMs, although further validation is necessary given the moderate maximum scores achieved. This opens up interesting prospects for building AI tools that are closely aligned with real-world teaching practices, potentially capturing verbal cues and pedagogical nuances that static documents often fail to convey.

Despite these strengths, limitations remain. The AI exhibited inconsistencies in responding to semantically similar questions, and its understanding of certain advanced concepts was limited. Moreover, the relatively small sample size and restricted scope of questions may constrain the generalizability of the results. Further research with broader datasets, more varied academic disciplines, and hybrid training inputs (audio + text) would help deepen our understanding of LLM capabilities in academic contexts.

Future research should include explicit comparative analyses using multiple input types—such as slides, transcripts, and textbooks—in conjunction with audio recordings. This would allow a more comprehensive evaluation of how input modality affects learning accuracy, semantic coherence, and response quality in LLMs. Looking ahead, AI systems like the one tested in this study may hold significant promise in student support roles, particularly as interactive chatbots capable of guiding learners through course content, offering clarifications, and referencing relevant materials from lectures. While not a replacement for traditional assessment or instruction, these models may become valuable assets in blended learning environments, helping to improve access, personalization, and efficiency in education.

In conclusion, this work contributes to the growing discourse on the integration of generative AI into higher education, offering a novel methodology and valuable empirical findings. From an operational perspective, educators could use such AI systems as supplementary teaching aids—for example, offering automated clarification of lecture content, generating practice questions, or summarising key concepts. However, when applied to technical courses, instructors should be mindful of the AI's tendency to underperform on complex reasoning tasks and design prompts or feedback mechanisms accordingly. With further development, fine-tuning, and validation, LLMs could evolve into key educational tools—supporting both students and instructors in shaping the future of digital learning.

7. References

Antaki, F., Touma, S., Milad, D., El-Khoury, J., Duval, R., 2023. Evaluating the Performance of ChatGPT in Ophthalmology: An Analysis of Its Successes and Shortcomings. Ophthalmology Science 3. https://doi.org/10.1016/j.xops.2023.100324

Bae, S.-M., Jeon, H.-R., Kim, G.-N., Kwak, S.-H., Lee, H.-J., 2024. Performance of ChatGPT on the Korean National Examination for Dental Hygienists. Journal of Dental Hygiene Science 24, 62–70. https://doi.org/10.17135/jdhs.2024.24.1.62

Bordt, S., von Luxburg, U., 2023. ChatGPT Participates in a Computer Science Exam.

Briatore, F., Vanni, F., Mosca, M.T., Mosca, R.N., Fruggiero, F., Mancusi, F., 2025. Exploring Industry 4.0's Role in Sustainable Supply Chains: Perspectives from a Bibliometric Review. Logistics. https://doi.org/10.3390/logistics9010026

Gilson, A., Safranek, C.W., Huang, T., Socrates, V., Chi, L., Taylor, R.A., Chartash, D., 2023. How Does ChatGPT Perform on the United States Medical Licensing Examination? The Implications of Large Language Models for Medical Education and Knowledge Assessment. JMIR Med Educ 9. https://doi.org/10.2196/45312

Hayes, D.S., Foster, B.K., Makar, G., Manzar, S., Ozdag, Y., Shultz, M., Klena, J.C., Grandizio, L.C., 2024. Artificial Intelligence in Orthopaedics: Performance of ChatGPT on Text and Image Questions on a Complete AAOS Orthopaedic In-Training Examination (OITE). J Surg Educ 81, 1645–1649. https://doi.org/10.1016/j.jsurg.2024.08.002

Jafarian, N.R., Kramer, A.W., 2025. AI-assisted audio-learning improves academic achievement through motivation and reading engagement. Computers and Education: Artificial Intelligence 8. https://doi.org/10.1016/j.caeai.2024.100357

Kung, T.H., Cheatham, M., Medenilla, A., Sillos, C., De Leon, L., Elepaño, C., Madriaga, M., Aggabao, R., Diaz-Candido, G., Maningo, J., Tseng, V., 2023. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. PLOS Digital Health 2. https://doi.org/10.1371/journal.pdig.0000198

Pursnani, V., Sermet, Y., Kurt, M., Demir, I., 2023. Performance of ChatGPT on the US fundamentals of engineering exam: Comprehensive assessment of proficiency and potential implications for professional environmental engineering practice. Computers and Education: Artificial Intelligence 5. https://doi.org/10.1016/j.caeai.2023.100183

Sánchez-Vera, F., 2025. Subject-Specialized Chatbot in Higher Education as a Tutor for Autonomous Exam Preparation: Analysis of the Impact on Academic Performance and Students' Perception of Its Usefulness. Educ Sci (Basel) 15. https://doi.org/10.3390/educsci15010026

Terwiesch, C., n.d. Would Chat GPT3 Get a Wharton MBA? A Prediction Based on Its Performance in the Operations Management Course.

Wei, B., 2025. Performance Evaluation and Implications of Large Language Models in Radiology Board Exams: Prospective Comparative Analysis. JMIR Med Educ 11. https://doi.org/10.2196/64284