The Future of Engineering Education
2024 Annual Conference & Exposition
Oregon Convention Center
Portland, OR . June 23 - 26, 2024
ASEE
Paper ID #43635

WIP: Traditional Engineering Assessments Challenged by ChatGPT: An Evaluation of its Performance on a Fundamental Competencies Exam

Trini Balart, Pontificia Universidad Católica de Chile

Trinidad Balart is a PhD student at Texas A&M University. She completed her Bachelors of Science in Computer Science engineering from Pontifical Catholic University of Chile. She is currently pursuing her PhD in Multidisciplinary Engineering with a focus in engineering education and the impact of AI on education. Her main research interests include Improving engineering students' learning, innovative ways of teaching and learning, and how artificial intelligence can be used in education in a creative and ethical way.

Dr. Jorge Baier, Pontificia Universidad Católica de Chile

He is an associate professor in the Computer Science Department and Associate Dean for Engineering Education at the Engineering School in Pontificia Universidad CatÃ³lica de Chile. Jorge holds a PhD in Computer Science from the University of Toronto in Ca

Martín Eduardo Castillo, Pontificia Universidad Católica de Chile

Martín Castillo is currently pursuing a Bachelor of Science in Robotics Engineering at the Pontifical Catholic University of Chile. His interests lie in the intersection of artificial intelligence, robotics, control systems and applications of AI in education.

WIP: Traditional Engineering Assessments Challenged by ChatGPT: An Evaluation of its Performance on a Fundamental Competencies Exam

Introduction

The evolution of artificial intelligence (AI) technologies, particularly in natural language processing, has brought forth transformative changes across various areas, including engineering education [1]. One of the most prominent manifestations of these advancements is ChatGPT, a large language model (LLM) developed by OpenAI, which has demonstrated remarkable capabilities in text generation and comprehension [2].

Indeed, the GPT-3.5 and GPT-4 models have raised important questions about the effectiveness and relevance of current engineering assessment methodologies [3]. The ability of these models to solve complex engineering problems, traditionally reserved for highly trained professionals, challenges the conventional pedagogical and evaluative approaches in engineering education. This paper explores the implications of such technologies on traditional engineering education, specifically focusing on the Fundamental Competencies Exam (FCE) used in a selective Latin American engineering school. This exam covers diverse subjects ranging from calculus to ethics, segmented into three modules: mathematics, science, and general engineering.

This study aims to empirically evaluate the performance of GPT-3.5 and GPT-4 on the FCE, encompassing a range of topics from calculus to engineering ethics. The literature review explores the trajectory of AI in educational settings, highlighting previous studies that have AI technology to solve academic problems and exams. It also discusses the theoretical underpinnings of AI in learning environments, examining how AI tools like ChatGPT can augment or, conversely, undermine traditional educational paradigms. Additionally, the review addresses the broader implications of AI in engineering education, including potential shifts in curriculum design, teaching methodologies, and assessment strategies.

The results presented in this paper represent more evidence in support of the fact that significant changes in engineering education are necessary. ChatGPT challenges assessment strategies and teaching methods, advocating for a shift towards exercise-centric learning to foster more meaningful experiences [4]. The ability of ChatGPT to redefine assignments underscores the need for educational institutions to establish clear guidelines for its use, ensuring it aids rather than hinders the learning process [3]. The emphasis on critical thinking and problem-solving skills remains crucial, as these enable students to effectively analyze and utilize information provided by AI tools like ChatGPT [5].

By scrutinizing the performance of ChatGPT on a standardized engineering exam, this paper contributes to the ongoing discourse on the role of AI in education and its impact on future learning and assessment models. The findings and discussions presented here may offer insights for educators, policymakers, and AI developers.

Methodology and findings

The Fundamental Competence Exam (FCE) is a prerequisite to obtain a Bachelor of Engineering degree and its objective is to assess students' fundamental engineering competences. To give the test students need to first pass a list of courses that are part of a common access plan that all the engineering undergraduate students take in the first two years of studies. This is because these courses are then assessed in the FCE.

The subjects that FCE aims to assess range from mathematics, science, ethics, programming, and economics. The FCE is formed by three modules: “Mathematics and probability” (Module 1), “Natural sciences, physics and chemistry” (Module 2) and “Engineering” (Module 3). The first module consists of 30 multiple choice questions with 4 possible answers and tests the subjects of 6 math related courses. The second module is made of 48 multiple choice questions with topics of courses of natural sciences. The third module tests topics of three general engineering courses and spreadsheets usage. The summary of the 3 modules, the numbers of questions, time given to students and the courses that each modulus covers are shown in Table 1. In order to pass the FCE the student has to answer 60% of the questions correctly. Finally, it's possible to pass a module individually with a score greater than 60% if the student meets certain criteria of scores in the exam. The criteria being scoring 50% or greater in the totality of the exam and scoring higher than 33% in each module individually

Table 1. Summary of FCE questions

Module	Time	Questions	Courses
1	2 hours	30	*Calculus 1, Calculus 2, Calculus 3, Differential equations, Linear algebra and Probability and statistics*
2	2 hours 50 minutes	48	*Dynamics, Electricity and magnetism, chemistry for engineering and thermodynamics*
3	1 hour 55 minutes	32	*Introduction to economy, Introduction to programming and Ethics for engineering*

We used a total of 8 versions of the FCE originally in MS Word, which is the format used to create the questions. Images in questions were saved in separate files. For each question, we recorded whether or not such a question contained mathematical expressions or images. The 8 exams were used from 2020 to 2023, with one version each semester (2020-2 being the same version as 2020-1) adding up to a total of 880 questions. From the total of questions, 23 questions were invalid, all belonging to module 1. This is due to the ambiguity of the mathematical expressions in some questions. In particular, because of a conversion made in the file format of the exams, matrices in linear algebra related questions were flattened and the shape of matrices is not clear. In each exam the number of invalid questions of module 1 was under 10% of the total number of questions. For the above, the dataset consisted of 857 questions with a total of 23 invalid questions.

It's should be noted that students are given a copy of the reference manual “Fundamentals of Engineering Supplied - Reference Handbook" by the National Council of Examiners for Engineering and Surveying (NCEES) for the resolution of the exam and in this study information from this manual was not directly included in any prompt to ChatGPT.

Questions with tables, mathematical or chemical expressions [6] were translated to LaTeX format, a well-known typesetting system that facilitates the structured representation of mathematical expressions, due to the difficulty of representing this type of content in UTF-8 plain text [7][8]. The aim of this strategy was to reduce human bias in the responses when assessing the model's fitness for the exam and to ensure the objectivity and consistency of the formulas and tables.An example of expressions in LaTeX language and the visualization of the expressions is in Table 2.

Table 2. Examples of FCE expressions and their LaTeX representation

Expression	LaTeX
$V_x = k \cdot e\,(\frac{6}{x} + \frac{16x}{x^2 - d^2})$	V_x = k \cdot e (\frac{6}{x} + \frac{16 x}{x^2 - d^2})
$2MnO_4^- + Br^- + 2H_2O \rightarrow MnO_2 + BrO_3^- + 2H^+ + 3e$	\text{Mn O_4}^- + \text{Br}^- + 2\text{H}_2\text{O} \rightarrow \text{Mn O_2} + \text{Br O_3}^- + 2\text{H}^+ + 3 e

The FCE questions were solved with ChatGPT using GPT-3.5 and GPT-4 with and without images. Questions that have images or diagrams were entered along with the text prompt with the recent inclusion of multimodality to the GPT-4 model [9]. For each question a single prompt to ChatGPT was made with the text corresponding to the question. The prompt, response and associated alternative were registered and then the average accuracy percentage per year, per model and per module is calculated. The accuracy percentage is of questions whose alternative coincided with the correct answer over the total of valid questions. The score obtained with each model, out of a total of 100%, per semester is shown in table 4, and the average accuracy per model is summarized in table 5.

Table 4. Accuracy percentage per semester of each model

Semester	2020 - 1	2020 - 2	2021 - 1	2021 - 2	2022 - 1	2022 - 2	2023 - 1	2023 - 2
GPT-3.5	43,08	-	52,96	46,25	45,38	47,22	43,85	52,96
GPT-4	61,39	-	62,14	64,56	54,72	67,90	64,17	66,54
GPT-4/ with images	62,24	-	69,40	66,38	60,17	63,41	63,50	65,57

Table 5. Average score obtained by each model

GPT-3.5	GPT-4	GPT-4 with image prompt
47,38%	63,06%	64,38%

From the above it is clear that the model GPT-4 scores higher overall than GPT-3.5, and that a slight increase in accuracy is achieved with images in the prompts for GPT-4. Also, a summary of the average score per module of each model is shown in Figure 1. From the graph it is clear that the highest score of each model is achieved in the module 3 and if the criteria for individual module passing was met, all of the models would pass the module 3. Also GPT-4 would barely pass module 1, and GPT-4 with images would pass individually modules 1 and 2 as well.

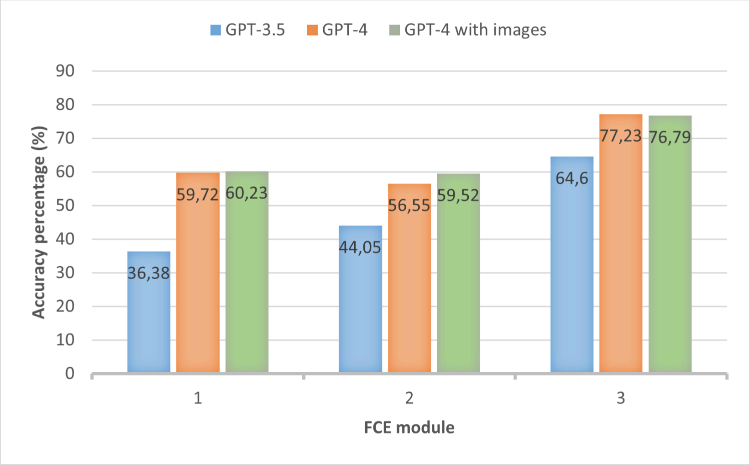

Figure 1. Average score obtained by module for each model

Discussion

When a student passes the FCE after passing all bachelor-level courses, our current assessment system regards the student as attaining the fundamental competencies of an engineering bachelor's. In our study, we observe that ChatGPT passes our FCE. The question that naturally arises is: does this mean ChatGPT attains the fundamental competencies we would associate with engineering bachelor's degree? The answer, as we argue below, is that this is very unlikely, which leads us to question current assessment practices, at the very least.

Various versions of ChatGPT have been shown unable to solve very simple reasoning problems. For example, Valmeekam et al. [10] show it performs very poorly in solving instances of the blocksworld problem, which is a toy AI problem developed in the 70s to evaluate AI planning systems. In an instance of blocksworld an agent is given an initial configuration, in which towers of labeled blocks are located on a table. The agent is also given a set of goal conditions, each of which expresses a statement of the form "block A is directly on top of block B", or "block C is directly on the table". The objective is to compute a sequence of actions that allows the agent to transform the initial configuration into one that satisfies the goal conditions. In every state, the agent, which has a robotic hand which can hold at most one block, may (1) pick a block from the table, (2) unstack a block at the top of a tower – both actions resulting in the agent having the block in her hand –, (3) put the block in hand down on the table, and (4) stack the block in hand on top of any of the towers. On a benchmark that can be solved by humans with an accuracy of over 70%, ChatGPT (in various configurations) cannot achieve more than 6% accuracy.

One of our FCE's competencies is "*to model simple situations from reality and basic physical phenomena using differential equations*", which is associated with the course *Differential Equations*. Given ChatGPT's inability to reason about (simple) blocksworld, it is unlikely that it attains such a competency, and unlikely that it actually attains other related competencies related to reasoning about the real world. Therefore, another natural question that arises is: if a system like ChatGPT passes the FCE but may not attain the fundamental competencies, do *students* that pass the FCE attain such competencies? A complete answer to this question remains out of the scope of the paper. However, the evidence that we provide in this paper seems to point to the fact that our assessment tools may be inadequate, as passing these exams does not clearly mean that the entity passing the exam (human or AI) definitely attains the fundamental competencies.

Finally, the redesign of the way we educate engineers in the future goes beyond how we actually assess competencies, but should also consider that the engineer of the future will likely use even more advanced AI technologies during their studies and professional activity. Future developments in engineering education should therefore consider the way these technologies will redefine competencies, assessment, and teaching.

Conclusions and future work

The findings from our study, which highlights the performance of ChatGPT on the FCE, signal a need for a comprehensive reevaluation of assessment strategies within engineering education. This reevaluation is not merely about adapting to AI technologies like ChatGPT but also about ensuring that the fundamental competencies expected of engineering graduates align with the rapidly evolving technological landscape. The future work of this research will focus on several key areas to further our understanding of these challenges and opportunities.

First, we plan to compare the performance of ChatGPT on the FCE with historical data from human test-takers. This comparison will help us to understand how the abilities of AI models like ChatGPT align with human competencies over time. By analyzing trends in performance, we can identify whether certain areas of the exam have become more or less challenging for students, providing insights into potential shifts in educational focus or gaps in knowledge.

A detailed examination of the types of questions answered correctly and incorrectly by ChatGPT will be conducted. This analysis will differentiate between computational, theoretical, and application-based questions to determine where ChatGPT excels and where it struggles. Understanding these patterns will inform the development of more effective teaching and assessment methods that can better differentiate between rote memorization and deep understanding. Based on our findings, we will propose new assessment methodologies that are designed to evaluate the competencies essential for future engineers. These methodologies will aim to be robust against the capabilities of sophisticated AI tools, focusing on critical thinking, problem-solving, and the application of knowledge in novel situations.

Recognizing the potential of AI to both enhance and disrupt traditional education models, we will develop recommendations for integrating AI tools into engineering education in a way that supports learning objectives. This includes using AI for personalized learning, as a tool for exploring complex engineering problems, and for automating routine tasks to free up time for more creative and analytical activities. Ethical implications of using AI in educational settings will also be addressed, including issues of fairness, accessibility, and the potential for misuse. We will propose policy recommendations for educational institutions, accrediting bodies, and policymakers to navigate these challenges, ensuring that the integration of AI into engineering education enhances learning outcomes without compromising integrity or equity.

References

[1] J. Qadir, "Engineering Education in the Era of ChatGPT: Promise and Pitfalls of Generative AI for Education," 2023 IEEE Global Engineering Education Conference (EDUCON), Kuwait, Kuwait, 2023, pp. 1-9, doi: 10.1109/EDUCON54358.2023.10125121

[2] P. P. Ray, “ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope,” Internet of Things and Cyber-Physical Systems, vol. 3, pp. 121–154, Jan. 2023, doi: 10.1016/j.iotcps.2023.04.003.

[3] S. Nikolic et al., “ChatGPT versus engineering education assessment: a multidisciplinary and multi-institutional benchmarking and analysis of this generative artificial intelligence tool to investigate assessment integrity,” European Journal of Engineering Education, vol. 48, no. 4, pp. 559–614, May 2023, doi: 10.1080/03043797.2023.2213169.

[4] P. P. Ray, “ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope,” Internet of Things and Cyber-Physical Systems, vol. 3, pp. 121–154, Jan. 2023, doi: 10.1016/j.iotcps.2023.04.003

[5] D. Cotton, P. A. Cotton, and J. R. Shipway, “Chatting and cheating: Ensuring academic integrity in the era of ChatGPT,” Innovations in Education and Teaching International, pp. 1–12, Mar. 2023, doi: 10.1080/14703297.2023.2190148

[6] X. Tang, “Struc-Bench: Are large language models really good at generating complex structured data?,” arXiv.org, Sep. 16, 2023. https://arxiv.org/abs/2309.08963

[7] W. Yeadon and D. P. Halliday, “Exploring Durham University Physics exams with Large Language Models,” arXiv (Cornell University), Jun. 2023, doi: 10.48550/arxiv.2306.15609.

[8] R. M. Taylor et al., “Galactica: a large language model for science,” arXiv (Cornell University), Nov. 2022, doi: 10.48550/arxiv.2211.09085.

[9] OpenAI, “GPT-4 Technical Report,” arXiv.org, Mar. 15, 2023. https://arxiv.org/abs/2303.08774

[10] K. Valmeekam, S. Sreedharan, M. Marquez, A. Olmo, and S. Kambhampati, “On the Planning Abilities of Large Language Models (A Critical Investigation with a Proposed Benchmark),” arXiv.org, Feb. 13, 2023. https://arxiv.org/abs/2302.06706