Aerospace Engineering Education in the Era of Generative AI

Julie B. Coder, The Pennsylvania State University

Julie B. Coder is a Ph.D. candidate in the Department of Curriculum and Instruction in the College of Education at Penn State University.

Dr. James G Coder, Pennsylvania State University

Dr. Jim Coder is an Associate Professor of Aerospace Engineering at Penn State University, specializing in applied aerodynamics and computational fluid dynamics.

Dr. Mark D. Maughmer, The Pennsylvania State University

Dr. Mark D. Maughmer is a professor of Aerospace Engineering at Penn State University, specializing in applied aerodynamics and aircraft design.

Aerospace Engineering Education in the Era of Generative AI

Abstract

The proliferation of large-language model (LLM) generative artificial intelligence (AI) tools like ChatGPT raises an inevitable question of how it should impact student assessments in aerospace engineering. To evaluate this, a large sample of multiple-choice questions from undergraduate aerodynamics and aeronautics courses were input into ChatGPT-4 and Gemini and the accuracy evaluated. The cognitive level of each question was coded using Bloom's taxonomy based on consensus of the authors. It was found that that generative AI performs increasingly poorly as the cognitive level is increased. Chi-square analyses of the data show very strong association with ChatGPT and strong association with Gemini for these trends. Cursory analysis of questions where both tools gave different wrong answers are consistent with the pattern matching aspects of LLMs. Based on the authors' observations, recommendations are offered for writing multiple choice questions that actually assess human understanding.

Introduction

Aerospace education across the United States is being subjected to strong external pressures, including but not limited to rapidly growing enrollments [1] and the proliferation of generative artificial intelligence (AI) tools like ChatGPT [2]. These are not independent concerns: they intersect each other when determining student assessments.

Over the past four years at The Pennsylvania State University, entrance to the Aerospace Engineering major and subsequent enrollment in third-year core aerospace courses has grown from approximately 160 students per year to over 250 students per year. As a result, this has placed an emphasis on scalability of assessments so that grading can be completed within a reasonable time and so that quality control can be maintained between the instructor and multiple teaching assistants. Timely scoring/grading is crucial. For formative assessments, students need the ability to act on feedback while the material is still relevant in the course. For summative assessments, scores and grades allow students to make programmatic decisions, and at the end of the semester there are strict deadlines for final grade entry. A common tendency is to use question types that reduce the variability in possible answers, such as multiple choice or highly scaffolded free-response prompts. This is true whether there is one monolithic section of a course with a single instructor or multiple smaller sections that require consistency across multiple instructors. While reducing variability improves the efficiency of grading, it comes at the expense of appreciating subtleties and differences in the thought processes of engineering students. In turn, it also limits the ability of high-performing students to distinguish themselves.

In the design of engineering course assessments, instructors must also be cognizant of generative AI's impact on both academia and industry. On the academic side, it presents a potential paradigm shift for how students learn outside of the classroom and how they complete asynchronous assignments. Surveys of student attitudes towards the use of generative AI in academics have revealed that they are more likely to turn to ChatGPT instead of a textbook or another supplemental resource to answer questions because it's "just easier" to ask ChatGPT [3],[4]. Students can also ask ChatGPT for additional practice questions as a way to prepare for exams [5]. To the former

point, generative AI is poised to function like a teaching assistant answering student questions and helping with homework. However, there is no instructor oversight of this process, nor will it be possible for anyone to provide a meaningful warranty on the accuracy of AI responses. It is important to remember that AI operates without intent. While it does not intend to be wrong, it also cannot intend to be correct. To the latter point, generative AI is a clear and present danger to academic integrity. Previously, instructors could create entirely new questions for which answer keys and solution manuals didn't exist if they were concerned about copied work. With generative AI, seemingly plausible answers to homework questions can be instantly constructed.

On the industrial side, the increasing breadth of capabilities of generative AI raises the inevitable question of the value added by human engineers. While the authors believe firmly in the importance of human creativity in the engineering process, that discourse is beyond the scope of the present paper. Nevertheless, we must confront the reality that generative AI will be viewed favorably by for-profit entities if the costs remain sufficiently low. As such, it is the responsibility of engineering educators to ensure that our students are competitive in the workforce against AI. Therein lies the intersection with course assessments: are the assessments used to evaluate and grade engineering students sufficient Turing tests to assert that the students provide added value over generative AI?

We concede that not every student in a given course will become a daily practitioner of that course's material. In integrated aerospace engineering curricula, the future astrodynamicist must take fluid dynamics, and the future aircraft structural designer must take orbital mechanics. Nevertheless, most students will likely need to interact with the products of generative AI, and it is useful for instructors to know how students' performance compares.

Background

Large language models (LLMs) such as Chat Generative Pretrained Transformer (ChatGPT) have exploded in use in recent years. ChatGPT was developed by Open AI and launched to the public on November 30, 2022 [6]. Within one week, it reached over 1 million users. LLMs can understand text and produce human-like responses to text input. They are trained on massive data sets and aim to predict the next word in a sequence by identifying patterns [7].

The rapidly expanding use of ChatGPT and other generative AI has prompted response from academic entities. Journals are developing policies around Chat GPT authorship [8] - [10], and universities are creating policies around the use of generative AI both in and outside of the classroom [11] - [13]. While the exact impact of these tools has yet to be seen, they are certainly poised to be disruptive technologies in engineering, and we must critically evaluate their use in the aerospace engineering classroom.

Exams are an important tool for determining student mastery of course material. Many exams use multiple choice questions (MCQs) where students are asked to select the best answer from a selection of 4-5 options. MCQs are objective, have (usually) one correct answer, and are easy to grade, which can be a significant advantage in very large courses. If they are written appropriately, MCQs can assess both basic knowledge and higher cognitive levels, making them a valuable tool for assessing student learning.

Previous studies have evaluated ChatGPT's performance on exams in a variety of fields. Laskar et al. [14] performed a comprehensive evaluation of ChatGPT's performance across a variety of academic fields and found that ChatGPT performs best in social science fields and worst in STEM fields. Open AI reported that ChatGPT achieved high passing scores on AP Biology, AP Chemistry, AP Calculus BC, AP Physics 2, and SAT and GRE math and quantitative sections [15], significantly better than the average of actual test-taking students.

Significant work has been done looking at the performance of ChatGPT on exams in medical fields [16] - [20] demonstrating that ChatGPT can obtain passing scores on some licensing exams in the medicine-related fields. Others have looked at the performance of ChatGPT in science, math, and engineering fields. Frenkel and Emara [21] assessed the use of ChatGPT-3.5 and -4 on junior- and senior-level undergraduate mechanical engineering exams and found correct answers of 51% and 76%, respectively, on test questions. Shikarian et al. [22] found that the failure rate of ChatGPT on math questions increases as word problems get more complex, although there was an improvement in performance when the input prompt was written to request that work be shown, while Azaria et al. [23] also observed overconfidence in answers to math-related questions. ChatGPT was found by Pursnari et al. [24] to achieve 75% accuracy on the Environmental Fundamentals of Engineering exam but failed at complex multistep calculations, often selecting the correct formula but arriving at the wrong answer. Ogundare et al. [25] found similar issues in trouble answering challenging questions; in particular they found that ChatGPT failed at applying knowledge to novel or unusual situations.

In general, ChatGPT has several shortcomings when answering questions in engineering. It is unable to learn from experience, tends to fabricate information or hallucinate, and has a considerable lack of knowledge past September 2021 [26]. It also shows significant bias, especially in math and physics fields [27]. Responses provided by ChatGPT are often verbose and confident, and responses can be misleading with no real evidence of deep understanding of concepts presented. Calculations performed by ChatGPT are not truly calculated but are instead ChatGPT's best predictions at what number comes next, limiting its effectiveness at solving any calculation-related questions.

To assess the cognitive level of the MCQs in this study, the authors chose to use Bloom's taxonomy. Bloom's taxonomy is a widely accepted framework in education for developing educational objectives and assessing learning outcomes based on a hierarchy of cognitive levels [28],[29] The 2001 revision [29] categorizes cognitive skills into six levels: remembering, understanding, applying, analyzing, evaluating, and creating. Previous studies have examined the classification of multiple-choice questions using Bloom's taxonomy [30],[31]. While Bloom's taxonomy is not without its limitations, it was chosen based on its familiarity and ability to provide common language for the discussion of different cognitive levels required to answer various assessment questions.

The research question being investigated in this study is how do generative AI tools perform on different Bloom's taxonomy cognitive level multiple choice questions in undergraduate aerospace engineering courses and what are the implications of this for evaluating student performance. Focus is placed here on ChatGPT-4 and Gemini. ChatGPT-4 was selected based both on its brand-

recognition and availability. Gemini was selected as a readily available alternative to crosscheck whether observed behaviors are unique to ChatGPT-4. This investigation has implications for both assessment writing as well as for the use of generative AI platforms in aerospace engineering education.

Methods

The basis of this study was 104 randomly selected multiple-choice questions (MCQs) from midterm and final exams in a selection of undergraduate junior- and senior-level aerodynamics and aeronautics courses in Aerospace Engineering at The Pennsylvania State University a large public research university located in University Park, Pennsylvania. The questions were fully written by the authors Dr. Coder and Dr. Maughmer for aerospace engineering courses that they have taught. All MCQs are original without any copyright issues.

Initially, the intent was to include short answer questions in addition to MCQs. Many of those questions required sketches in their answer, which were not meaningfully producible by ChatGPT-4 and not at all producible by Gemini as of the writing of this paper in January 2025. Due to these limitations, the authors made the decision to exclude short answer questions for this study.

The MCQs were classified into four cognitive levels based on revised Bloom's taxonomy by consensus among the authors. No questions at the "Evaluate" or "Create" level were included, as these are difficult cognitive levels to assess with multiple choice questions. Examples of criteria used to determine Bloom's taxonomy level as well as example questions for each level can be found in Table 1.

Two generative AI models were selected for testing- ChatGPT-4 (Open AI, San Francisco, CA) and Gemini (Google, Mountain View, California). These models were chosen for their accessibility and availability. The following prompt was used for each set of questions: "Choose the best answer for the following aerospace engineering multiple choice test questions." Questions were entered in groups of 4-5 at a time in the order that they were presented on the exams that they were taken from, and answers were recorded and later scored as being correct or incorrect using answer keys from the exams from which the questions were taken.

To evaluate the performance of ChatGPT-4 and Gemini in relation to the cognitive levels of the multiple-choice questions, a chi-square analysis was conducted. The chi-square test was used to assess the null hypothesis that there is no correlation between the AI models' performance (correct or incorrect answers for MCQs) and the Bloom's cognitive level of the questions. For each cognitive level, the observed frequencies of correct and incorrect answers were compared against the expected frequencies under the assumption of no correlation. Additionally, p-values were calculated to determine the statistical significance of the results, with a significance level of α=.05. This analysis determined whether variations in performance across cognitive levels were statistically significant or occurred due to chance.

TABLE I

EXAMPLES OF QUESTIONS BY BLOOM'S TAXONOMY LEVEL

Bloom's Taxonomy Level	**Explanation of Bloom's Taxonomy Level**	**Example Question**
Remember	Remembering previously learned information; recalling facts and basic concepts	Subsonic profile drag a) includes induced and parasite drag. b) is composed of viscous and pressure drags. c) is not dependent on angle of attack. d) is determined primarily using potential flow. e) is not a component of the parasite drag.
Understand	Grasping the meaning of information; explaining ideas or concepts	The Kutta condition a) uniquely determines the value of the circulation. b) requires a stagnation point at airfoil trailing edges having a finite angle. c) requires a stagnation point at the trailing-edge location in the circle plane. d) requires smooth flow off the trailing edge. e) all of the above.
Apply	Using information in new situations	The major consequence of the incompressibility assumption is that a) it decouples conservation of mass from conservation of momentum. b) by satisfying the equation of state, it reduces the problem to that of solving the continuity and momentum equations. c) it makes the energy equation invalid. d) as density is constant, it reduces the problem from one of six equations and six unknowns to one of five equations and five unknowns. e) all of the above.
Analyze	Drawing connections among ideas; breaking down ideas into simpler parts and seeing how they are related	Consider the following PDE that is to be discretized using Godunov's finite-volume method with the HLL approximate Riemann solver. $\frac{\partial u}{\partial t} + \begin{bmatrix} 1 & 2 \\ 0 & 3 \end{bmatrix} \frac{\partial u}{\partial x} = 0 \quad \rightarrow \quad \frac{\partial u}{\partial t} + \frac{f_{j+1/2} - f_{j-1/2}}{\Delta x} = 0$ If $u_j = [1, 1]^T$ and $u_{j+1} = [3, 1]^T$, what is the value of the flux at j + ½? Note that the two eigenvalues of the system matrix are equal to +1 and +3, respectively. a) $f_{j+1/2} = [1, 1]^T$ b) $f_{j+1/2} = [3, 3]^T$ c) $f_{j+1/2} = [5, 3]^T$ d) $f_{j+1/2} = [2, 1]^T$

Results

Of the 104 MCQ questions selected, 100 were included in the analysis. Four questions containing images were excluded because they could not be analyzed by either ChatGPT-4 or Gemini as of January 2025.

The overall performance of ChatGPT-4 (69.0% of answers correct) was higher than that of Gemini (63.0% of answers correct). Performance differences were observed across the different cognitive levels of Bloom's taxonomy. Of the 100 analyzed questions, 37 of the items were classified as "Remember" (37.0%), 42 were "Understand" (42.0%), 18 were "Apply" (18.0%), and 3 were "Analyze" (3.0%). Figure 1 shows a breakdown of questions by Bloom's taxonomy level.

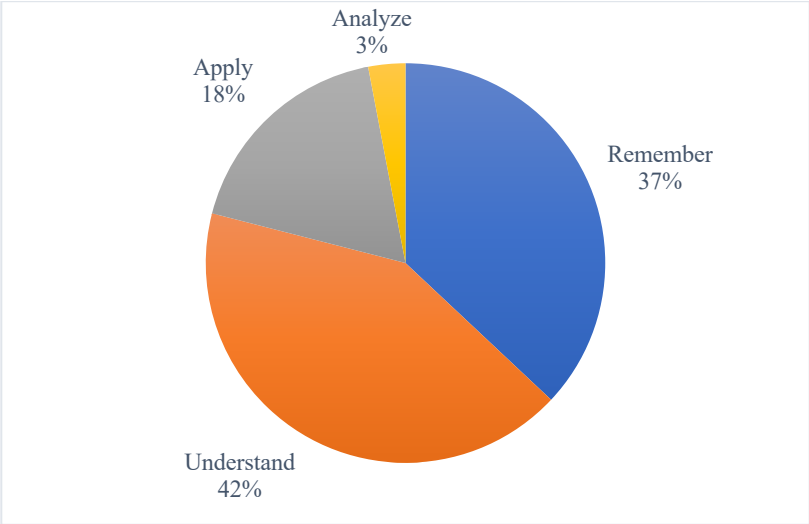

Fig. 1 Breakdown of questions analyzed by Bloom's Taxonomy level.

Performance of ChatGPT-4 and Gemini on each question type is summarized in Tables II and III. At the "Remember" level, both models showed relatively high accuracy. ChatGPT-4 outperformed Gemini by almost 10%. For questions classified as "Understand," which required comprehension and interpretation of previously known facts, ChatGPT-4 again outperformed Gemini, this time by a smaller margin of approximately 5%. At the "Apply" level, where the questions required applying learned principles to solve problems, both Chat GPT-4 and Gemini showed identical performance, each correctly answering only 50% of the questions with differences in which questions were correct and incorrect for each model. For the small subset of "Analyze" questions, both models were unable to answer any questions correctly, highlighting the challenges of higher-order cognitive tasks for AI models.

Bivariate analysis was performed to find the association of ChatGPT-4 and Gemini performance with Bloom's taxonomy level. Very strong association was found between the performance ChatGPT-4 and Bloom's taxonomy level ($p = .002$ $\alpha = .05$), indicating that performance varied significantly across the cognitive levels. Strong association was found between the performance of Gemini and Bloom's taxonomy level ($p = .03$, $\alpha = .05$), demonstrating that Gemini's ability to correctly answer questions was also correlated to the cognitive demands of the questions. These findings show that generative AI has differing capabilities at answering questions across cognitive

levels with performance declining at higher cognitive levels. The results indicate that while both models can handle basic recall MCQs relatively effectively, there are clear limitations when asked to perform more complex cognitive tasks. These observations provide important insights into the capabilities and constraints of generative AI models in the context of aerospace engineering education.

TABLE II
RELATIONSHIP OF BLOOM'S TAXONOMY LEVELS AND PERFORMANCE OF CHATGPT-4

Bloom's Taxonomy Level	**Chat GPT Correct**	**Chat GPT Incorrect**	**X^2**
Remember (n=37)	32 (86.5%)	5 (13.5%)	X^2 (3) = 15.08, p = .002
Understand (n=42)	28 (66.7%)	14 (33.3%)	
Apply (n=18)	9 (50.0%)	9 (50.0%)	
Analyze (n=3)	0 (0.0%)	3 (100.0%)	

TABLE III
RELATIONSHIP OF BLOOM'S TAXONOMY LEVELS AND PERFORMANCE OF GEMINI

Bloom's Taxonomy Level	**Gemini Correct**	**Gemini Incorrect**	**X^2**
Remember (n=37)	28 (75.7%)	9 (24.3%)	X^2 (3) = 8.97, p = .030
Understand (n=42)	26 (61.9%)	16 (38.1%)	
Apply (n=18)	9 (50.0%)	9 (50.0%)	
Analyze (n=3)	0 (0.0%)	3 (100.0%)	

Discussion

This study focused on the performance of ChatGPT-4 and Gemini on multiple choice questions in undergraduate aerospace engineering courses at varying levels of Bloom's taxonomy. The findings revealed that as cognitive levels of the questions posed increased, performance of both generative AI models decreased significantly, as seen in Figure 2, demonstrating the limitations of generative AI in solving problems in aerospace engineering.

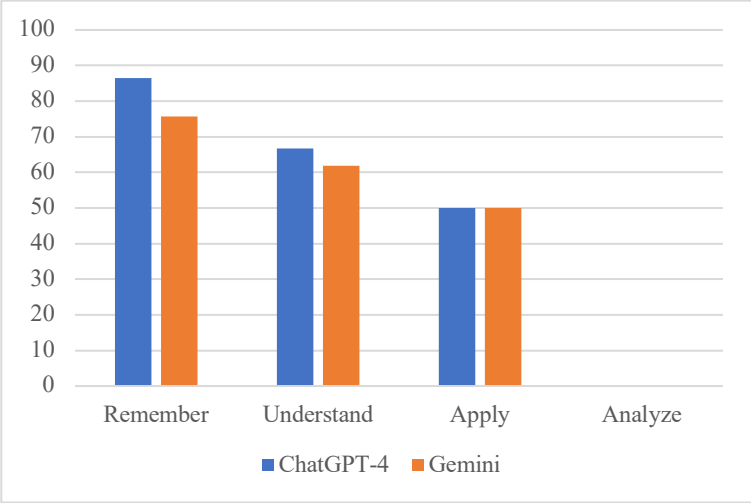

Fig. 2. Percent correct by ChatGPT-4 and Gemini for each question type.

The “Remember” level of Bloom’s taxonomy relies on recalling of previously learned facts and concepts. Chat GPT-4’s performance at this level (86.5% of questions correct) was higher than that of Gemini (75.7% correct) indicating that Gemini’s ability to answer even simple questions in aerospace engineering is less accurate than that of ChatGPT-4. This performance is lower than what has been observed in other fields [30],[31]. One reason for this could be that the questions were specifically written by the authors for their courses and, to the best of their knowledge, not published online. As a result, these questions were likely not part of the ChatGPT or Gemini training materials which likely contributes to the lower-than-expected performance.

Given their low performance on "Remember" questions, it's not surprising that ChatGPT and Gemini struggled at higher cognitive levels, which require synthesizing various pieces of information to form new conclusions. The higher levels of cognition such as apply and analyze require not only the recall of facts but also the ability to connect different concepts in new ways. Since both models showed limited abilities in basic recall of facts, their ability to perform at higher cognitive levels is also limited, highlighting the problems that generative AI faces in complex problem-solving tasks and deep understanding of content.

It was deemed of interest to further analyze multiple choice questions for which Chat-GPT 4 and Gemini provided different wrong answers for the same problem. The motivation for doing so is to help understand what qualities of a multiple-choice question and/or the distractors provides difficulty for AI. Here, for brevity, we focus on four different questions, one each that were coded Remember, Understand, Apply, and Analyze.

1. Remember

The selected Remember question came from an “Aeronautics” final exam,

The downwash for a uniformly loaded wing is
a. Always constant
b. Greatest at the midspan

c. *Greatest at the wing tips*
d. *Not a function of the spanwise location*
e. *Reduced as the lift coefficient is increased*

This question pertains to lifting-line theory, which is a prominent topic covered in the course, and contains the keywords "downwash" and "wing" that students can use to identify the topic. The correct answer to this question is (c), and students are expected to recall that a uniformly loaded wing may be represented as a single horseshoe vortex under lifting-line theory, with finite-strength vortex filaments trailing from the wingtips. It is repeatedly emphasized in the course that velocities vary with 1/r as they approach a vortex filament. This example is subsequently used in the construction of lifting-line theory for generalized lift distributions. Moreover, the distractors for this question invoke recall of other results from lifting-line theory. For instance, answers (a) and (d) are qualities of an elliptical lift distribution, answer (b) is the opposite of the correct answer, and answer (e) is not a typical behavior at positive lift coefficients (which can be reasonably assumed in the absence of further qualifications).

The answers from ChatGPT-4 and Gemini were (b) and (a), respectively. If a student would answer (a), this would be an expected wrong answer because it is associated with elliptical loading, which takes a prominent place in many discussions of lifting-line theory due to its property of minimum induced drag. Furthermore, if someone with mastery of lifting-line theory were to read the distractors without reading the questions, answer (a) might be the expected result. Conversely, it is not clear how expectations and patterns could lead to answer (b), and no additional speculation will be provided in this paper.

2. Understand

The selected Understand question came from an "Aeronautics" midterm exam,

The assumptions made in thin-airfoil theory
a. *Simplify the governing equations*
b. *Ensure that the velocity is continuous across the vortex sheet*
c. *Allow the governing equations to be linearized*
d. *Allow the boundary conditions of the problem to be linearized*

As stated in the question, it pertains to thin-airfoil theory, which is covered in-depth in the Aeronautics course. The correct answer is (d), but the question requires deeper understanding to identify why (a) and (c) are incorrect as these were the respective answers from ChatGPT-4 and Gemini. Thin-airfoil theory is derived assuming linear potential flow, which is governed by Laplace's equation. The simplification and linearization of the governing equation(s) from Navier-Stokes is independent of the derivation of thin-airfoil theory, hence why (a) and (c) are not correct. Nevertheless, these are requisite qualities that are exploited by thin-airfoil theory and included in its discussion, which may be why LLMs select these answers. Although Laplace's equation is linear, the difficulty in analytically solving potential flow for the general case is the nonlinearity of the boundary conditions, which is the piece directly addressed by the assumptions of thin-airfoil theory. Answer (b) is fully incorrect, as a vortex sheet is defined by the presence a discontinuous velocity field, irrespective of any other flow qualities ascribed to or enforced on the sheet.

3. Apply

The selected Apply question came from a "Stability and Control of Aircraft" midterm exam,

> *An underbalanced elevator (one which is hinged well forward) will tend to*
> a. *Float with increasing angles of attack depending on the pressure distribution which occurs over the elevator surface*
> b. *Float up with increasing angles of attack due to its inertia properties*
> c. *Float down with increasing angles of attack*
> d. *Float up causing a forward shift of the neutral point and an increase in the static stability*

This question relates to the behavior of an aircraft and its elevator when calculating stick-free behavior, such as the stick-free neutral point, and contains keywords such as "elevator" and "hinged" to indicate the specific topic. It requires the student to recognize the overall behavior of the aircraft/tailplane/elevator system and apply the correct understanding of both aerodynamics, flight dynamics, and elevator dynamics. Answer (a) is correct and is properly reflective of all subtleties of the problem, whereas (b) and (c), which were selected by ChatGPT and Gemini, are overgeneralizations of the system response that also include additional incorrect information to help the students. As an "Apply" question, (a) is written in a way that is unlikely to match lists of basic facts about stick-free elevator behavior. Conversely, (b) and (c) match up to available facts albeit with a lower degree of certainty and may have been selected with a simple plurality of confidence. Answer (d) is neither correct nor was chosen by the generative AI tools. While the stick-free neutral point can be forward of the stick-fixed neutral point, any forward movement of the neutral point decreases the static stability as commonly measured by the static margin. Hence, (d) is self-contradictory.

4. Analyze

The selected Analyze question was taken from an "Introduction to Numerical Methods in Fluid Dynamics" final exam,

> *Consider the finite-difference scheme given by*

$$\frac{df}{dx} \approx \frac{f_{j-2} - 4f_{j-1} + 4f_{j+1} - f_{j+2}}{4\Delta x}$$

> *Which of the following expressions is its modified wavenumber response?*

a. $\tilde{k}\Delta x = \sin k\Delta x$
b. $\tilde{k}\Delta x = \sin k\Delta x - \frac{1}{4}\sin 2k\Delta x$
c. $\tilde{k}\Delta x = 2\sin k\Delta x - \frac{1}{2}\sin 2k\Delta x$
d. $\tilde{k}\Delta x = \frac{4}{3}\sin k\Delta x - \frac{1}{6}\sin 2k\Delta x$

Modified wavenumber analysis is a standard concept in the analysis of numerical schemes, and one that was covered fully in the course. The finite-difference formula given in the prompt is one that is atypical and would not appear verbatim in any study material. Thus, students are expected to apply the concepts of both Taylor and Fourier analysis to derive the correct answer, which is (c). ChatGPT and Gemini answered (b) and (d), respectively. Interestingly, neither selected (a), which is a common result from using modified wavenumber analysis with a standard second-order finite-difference expression. Answer (b) is fully incorrect because it is not a consistent result for a valid finite-different scheme, though it does include elements of both the common response as well as a correction for the question asking about a five-point stencil rather than a three-point stencil. Hence, a pattern matching approach could reasonably obtain this result. On the other hand, answer (d) is the modified wavenumber response for a different five-point stencil, but one that is more commonly used.

Based on what was observed in this study, we offer a few ways to increase the likelihood that assessments are actually testing human understanding of aerospace engineering concepts instead of student ability to ask questions of generative AI. The first recommendation is to incorporate questions that are higher in Bloom's taxonomy, as accuracy of generative AI at the "Apply" and "Analyze" levels was greatly decreased. The second is include images in the questions and require sketches in the responses, which truly assesses human understanding. Third, selection of distractors in MCQs should not include "throwaway" answers but instead should include common keywords, phrases, and concepts that are prominent in potential training material. Finally, avoid using questions that come directly from textbooks or are readily available online. This ensures that ChatGPT and other platforms do not have the questions in their training databases.

Study Limitations

The study limitations include the limited number of courses and relatively small sample size represented in the selection of MCQs which limits the generalizability of the results of the study to other aerospace engineering courses. The rapid evolution of LLMs also presents risk that the results may not be relevant to the capabilities of generative AI models in the future.

Conclusions and Future Work

This study demonstrates that both ChatGPT-4 and Gemini fail to accurately answer aerospace engineering MCQs at higher cognitive levels of Bloom's taxonomy. This brings into question the use of generative AI as a supplementary learning tool in aerospace education. Students claim that using ChatGPT has deterred them from reading textbooks, accessing the library, or using Google and YouTube to search for supplemental learning resources because of the low effort required to access ChatGPT. They can ask ChatGPT basic or detailed questions anywhere and any time. Even though ChatGPT has been shown to have quality and accuracy issues and can produce verbose and confident wrong answers that sound realistic to a novice learner, the ease of use outweighs any risks of inaccuracy from the perspective of many students. This presents a real danger to education in general if students are not prepared to assess the accuracy of the content presented to them in the answer to a question.

When creating questions for assessment, especially assessment that is done outside of class, it is important to incorporate higher-level cognitive skills from Bloom's taxonomy as this decreases the chances that students can effectively use ChatGPT to accurately answer the questions. Choice of distractors also influences the cognitive level of the question and can make it more difficult for ChatGPT to answer.

The limitations of ChatGPT are particularly concerning when one considers the introduction of virtual teaching assistants and the use of ChatGPT to create lesson plans, curriculum, and perform grading tasks. If it cannot be trusted to answer MCQs at higher cognitive levels in aerospace engineering, should we trust it to tutor students or to grade their work? Can it effectively design lessons and projects that authentically assess a student's knowledge and engineering skills? If students become overreliant on ChatGPT, how will it affect the creativity that is needed in engineering? These are questions that will need to be grappled with in the upcoming years as generative AI becomes more pervasive and are areas that would benefit from future study.

References

[1] "Facts and Figures : About Us," *College of Engineering - Purdue University*. https://engineering.purdue.edu/Engr/AboutUs/FactsFigures/school_facts.html

[2] Vazquez, R. (2024) 'Experiences and insights from a mini-course on responsible generative AI use in aerospace engineering', *IFAC-PapersOnLine*, 58(16), pp. 35–40. doi:10.1016/j.ifacol.2024.08.458.

[3] A. Shoufan, "Exploring Students' Perceptions of ChatGPT: Thematic Analysis and Follow-Up Survey," in *IEEE Access*, vol. 11, pp. 38805-38818, 2023, doi: 10.1109/ACCESS.2023.3268224

[4] C. C. Tossell, N. L. Tenhundfeld, A. Momen, K. Cooley and E. J. de Visser, "Student Perceptions of ChatGPT Use in a College Essay Assignment: Implications for Learning, Grading, and Trust in Artificial Intelligence," in *IEEE Transactions on Learning Technologies*, vol. 17, pp. 1069-1081, 2024, doi: 10.1109/TLT.2024.3355015.

[5] "Association of Women Surgeons Need extra board exam preparation? Try Chat GPT…," *Womensurgeons.org*, 2024. https://blog.womensurgeons.org/medical-students/need-extra-board-exam-preparation-try-chat-gpt/ (accessed Jan. 15, 2025).

[6] OpenAI, *Introducing ChatGPT*, 2022. https://openai.com/index/chatgpt

[7] S. Wolfram, "What Is ChatGPT Doing … and Why Does It Work?," *writings.stephenwolfram.com*, Feb. 14, 2023. https://writings.stephenwolfram.com/2023/02/what-is-chatgpt-doing-and-why-does-it-work/

[8] "Change to policy on the use of generative AI and large language models," *Science.org*, 2023, doi: https://doi.org/10.1126/science.adm9531.

[9] "Ethical Standards for Publication of Aeronautics and Astronautics Research," *www*, 2018. https://www.aiaa.org/publications/Publish-with-AIAA/Ethical-Standards-for-Publication-of-Aeronautics-and-Astronautics-Research#3-use-of-artificial-intelligence-in-aiaa-publications (accessed Jan. 15, 2025).

[10] "American Society of Mechanical Engineers - Authorship and AI Tools," *Secure-platform.com*, 2025. https://vvs.secure platform.com/a/page/author_resources/authorship_and_AI_tools (accessed Jan. 15, 2025).

[11] "Guidelines," *Official Site of the Penn State AI Hub*, Aug. 29, 2024. https://ai.psu.edu/guidelines/

[12] "Considering Generative AI and ChatGPT at Virginia Tech," *tlos.vt.edu*.

https://tlos.vt.edu/resources/generative-ai.html

[13] "Acceptable Use of ChatGPT and Similar AI Tools | UT Austin Information Security Office," *security.utexas.edu*. https://security.utexas.edu/ai-tools

[14] Laskar, B. M. Saiful, M. Rahman, Bhuiyan, S. Joty, and J. X. Huang, "A Systematic Study and Comprehensive Evaluation of ChatGPT on Benchmark Datasets," May 2023, doi: https://doi.org/10.48550/arxiv.2305.18486.

[15] OpenAI, "GPT-4," *Openai.com*, 2023. https://openai.com/index/gpt-4-research/

[16] S. Wójcik, A. Rulkiewicz, Piotr Pruszczyk, Wojciech Lisik, Marcin Poboży, and Justyna Domienik-Karłowicz, "Reshaping medical education: Performance of ChatGPT on a PES medical examination," *Cardiology Journal*, Oct. 2023, doi: https://doi.org/10.5603/cj.97517.

[17] K. M. Surapaneni, "Assessing the Performance of ChatGPT in Medical Biochemistry Using Clinical Case Vignettes: Observational Study," *JMIR Medical Education*, vol. 9, p. e47191, Nov. 2023, doi: https://doi.org/10.2196/47191.

[18] L. Birkett, T. Fowler, and S. Pullen, "Performance of ChatGPT on a primary FRCA multiple choice question bank," *British Journal of Anaesthesia*, vol. 131, no. 2, pp. e34–e35, Aug. 2023, doi: https://doi.org/10.1016/j.bja.2023.04.025.

[19] K. Ishida and E. Hanada, "Potential of ChatGPT to Pass the Japanese Medical and Healthcare Professional National Licenses: A Literature Review," *Cureus*, Aug. 2024, doi: https://doi.org/10.7759/cureus.66324.

[20] Mehmet Buldur and B. Sezer, "Evaluating the accuracy of Chat Generative Pre-trained Transformer version 4 (ChatGPT-4) responses to United States Food and Drug Administration (FDA) frequently asked questions about dental amalgam," *BMC Oral Health*, vol. 24, no. 1, May 2024, doi: https://doi.org/10.1186/s12903-024-04358-8.

[21] M. E. Frenkel and Hebah Emara, "ChatGPT-3.5 and -4.0 and mechanical engineering: Examining performance on the FE mechanical engineering and undergraduate exams," *Computer Applications in Engineering Education*, Jul. 2024, doi: https://doi.org/10.1002/cae.22781.

[22] P. Shakarian, Abhinav Koyyalamudi, N. Ngu, and Lakshmivihari Mareedu, "An Independent Evaluation of ChatGPT on Mathematical Word Problems (MWP)," *arXiv (Cornell University)*, Feb. 2023, doi: https://doi.org/10.48550/arxiv.2302.13814.

[23] A. Azaria, R. Azoulay, and S. Reches, "ChatGPT is a Remarkable Tool—For Experts," Data intelligence, vol. 6, no. 1, pp. 1–49, Nov. 2023, doi: https://doi.org/10.1162/dint_a_00235.

[24] V. Pursnani, Y. Sermet, M. Kurt, and I. Demir, “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*, vol. 5, p. 100183, Jan. 2023, doi: https://doi.org/10.1016/j.caeai.2023.100183.

[25] O. Ogundare, S. Madasu and N. Wiggins, "Industrial Engineering with Large Language Models: A Case Study of ChatGPT's Performance on Oil & Gas Problems," *2023 11th International Conference on Control, Mechatronics and Automation (ICCMA)*, Grimstad, Norway, 2023, pp. 458-461, doi: 10.1109/ICCMA59762.2023.10374622.

[26] 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.

[27] K. Abramski, S. Citraro, L. Lombardi, G. Rossetti, and M. Stella, “Cognitive network science reveals bias in GPT-3, ChatGPT, and GPT-4 mirroring math anxiety in high-school students,” *arXiv (Cornell University)*, May 2023, doi:10.48550/arxiv.2305.18320.

[28] B. Bloom, *Taxonomy of Educational Objectives: The Classification of Educational goals. Handbook 1, Cognitive Domain*. New York: Longman, 1956.

[29] L. W. Anderson and D. Krathwohl, *A taxonomy for learning, teaching, and assessing: A revision of Bloom’s*. Essex: Pearson, 2001.

[30] D. Shay *et al.*, “Could ChatGPT-4 pass an anaesthesiology board examination? Follow-up assessment of a comprehensive set of board examination practice questions,” *British Journal of Anaesthesia*, vol. 132, no. 1, pp. 172–174, Nov. 2023, doi: https://doi.org/10.1016/j.bja.2023.10.025.

[31] B. Z. Dymond, M. Swenty, and J. C. Carroll, “Comparing Exam Performance in a Reinforced Concrete Design Course with Bloom’s Taxonomy Levels,” *Journal of Civil Engineering Education*, vol. 146, no. 1, p. 04019001, Jan. 2020, doi: https://doi.org/10.1061/(asce)ei.2643-9115.0000002.