PNAS RESEARCH ARTICLE | COMPUTER SCIENCES OPEN ACCESS

Could ChatGPT get an engineering degree? Evaluating higher education vulnerability to AI assistants

Beatriz Borges[a,1], Negar Foroutan[a,1], Deniz Bayazit[a,1], Anna Sotnikova[a,1], Syrielle Montariol[a], Tanya Nazaretzky[a], Mohammadreza Banaei[a], Alireza Sakhaeirad[a], Philippe Servant[a], Seyed Parsa Neshaei[a], Jibril Frej[a], Angelika Romanou[a], Gail Weiss[a], Sepideh Mamooler[a], Zeming Chen[a], Simin Fan[a], Silin Gao[a], Mete Ismayilzada[a], Debjit Paul[a], Philippe Schwaller[a], Sacha Friedli[a], Patrick Jermann[a], Tanja Käser[a], Antoine Bosselut[a,2], EPFL Grader Consortium[a,3], and EPFL Data Consortium[a,4]

Affiliations are included on p. 8.

Edited by Jeffrey Ullman, Stanford University, Stanford, CA; received August 9, 2024; accepted October 22, 2024

AI assistants, such as ChatGPT, are being increasingly used by students in higher education institutions. While these tools provide opportunities for improved teaching and education, they also pose significant challenges for assessment and learning outcomes. We conceptualize these challenges through the lens of vulnerability, the potential for university assessments and learning outcomes to be impacted by student use of generative AI. We investigate the potential scale of this vulnerability by measuring the degree to which AI assistants can complete assessment questions in standard university-level Science, Technology, Engineering, and Mathematics (STEM) courses. Specifically, we compile a dataset of textual assessment questions from 50 courses at the École polytechnique fédérale de Lausanne (EPFL) and evaluate whether two AI assistants, GPT-3.5 and GPT-4 can adequately answer these questions. We use eight prompting strategies to produce responses and find that GPT-4 answers an average of 65.8% of questions correctly, and can even produce the correct answer across at least one prompting strategy for 85.1% of questions. When grouping courses in our dataset by degree program, these systems already pass the nonproject assessments of large numbers of core courses in various degree programs, posing risks to higher education accreditation that will be amplified as these models improve. Our results call for revising program-level assessment design in higher education in light of advances in generative AI.

LLM | education | generative AI | education vulnerability

Significance

Universities primarily evaluate student learning through various course assessments. Our study demonstrates that AI assistants, such as ChatGPT, can answer at least 65.8% of examination questions correctly across 50 diverse courses in the technical and natural sciences. Our analysis demonstrates that these capabilities render many degree programs (and their teaching objectives) vulnerable to potential misuse of these systems. These findings call for attention to assessment design to mitigate the possibility that AI assistants could divert students from acquiring the knowledge and critical thinking skills that university programs are meant to instill.

ChatGPT, a system using a large language model (LLM), GPT-3.5, as its foundation, was released in November 2022 to broad adoption and fanfare, reaching 100M users in its first month of use and remaining in the public discourse to this day. As arguably the most hyped AI system to date, its release has prompted a continuing discussion of societal transformations likely to be induced by AI over the next years and decades. Potential changes in modern educational systems have remained a core topic in this discussion, with early reports highlighting the risk of these AI systems as tools that would allow students to succeed in university coursework without learning the relevant skills those courses are meant to teach. Despite this concern, there has yet to be a comprehensive empirical study of the potential impact of LLMs (and derivative tools) on the assessment methods that educational institutions use to evaluate students. A few studies have explored the interesting subtask of how well models perform on problems related to topics typically taught in many university courses and aggregated relevant question sets for this purpose (1–5). However, none of these works extrapolate these findings to assess the downstream impact of these tools on degree programs, where the risk of these technologies relative to their pedagogical benefits must actually be measured.

In this work, we conduct a large-scale study across 50 courses from EPFL to measure the current performance of LLMs on higher education course assessments. The selected courses are sampled from 9 Bachelor's, Master's, and Online programs, covering between 33% and 66% of the required courses in these programs. From these courses, we compile a bilingual dataset (English and French) of 5,579 textual open-answer and multiple-choice questions (MCQ). All questions were extracted from real exams, assignments, and practical exercise sessions used to evaluate students in previous years. The course distribution is presented in Fig. 1, and the dataset statistics are shown in Table 1 (stratified by particular dataset attributes).

[2]To whom correspondence may be addressed. Email: antoine.bosselut@epfl.ch.

[3]Grader Consortium: Alexandre Schöpfer, Andrej Janchevski, Anja Tiede, Clarence Linden, Emanuele Troiani, Francesco Salvi, Freya Behrens, Giacomo Orsi, Giovanni Piccioli, Hadrien Sevel, Louis Coulon, Manuela Pineros-Rodriguez, Marin Bonnassies, Pierre Hellich, Puck van Gerwen, Sankalp Gambhir, Solal Pirelli, Thomas Blanchard, Timothée Callens, Toni Abi Aoun, Yannick Calvino Alonso, and Yuri Cho.

[4]Data Consortium: Aleksandra Radenovic, Alexandre Alahi, Alexander Mathis, Anne-Florence Bitbol, Boi Faltings, Cécile Hébert, Devis Tuia, François Maréchal, George Candea, Giuseppe Carleo, Jean-Cédric Chappelier, Nicolas Flammarion, Jean-Marie Fürbringer, Jean-Philippe Pellet, Karl Aberer, Lenka Zdeborová, Marcel Salathé, Martin Jaggi, Martin Rajman, Mathias Payer, Matthieu Wyart, Michael Gastpar, Michele Ceriotti, Ola Svensson, Olivier Lévêque, Paolo Ienne, Rachid Guerraoui, Robert West, Sanidhya Kashyap, Valerio Piazza, Viesturs Simanis, Viktor Kuncak, Volkan Cevher, Akhil Arora, Alberto Chiappa, Antonio Sclocchi, Étienne Bruno, Florian Hofhammer, Gabriel Pescia, Geovani Rizk, Leello Dadi, Lucas Stoffl, Manoel Horta Ribeiro, Matthieu Bovel, and Yueyang Pan.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2414955121/-/DCSupplemental.

Published November 26, 2024.

Using this dataset, we subsequently test two commonly used models, GPT-4 (6), the system widely considered to be the most performant (7) among public AI assistants* and GPT-3.5 (8), a highly performant freely available system. We generate responses from these systems using a range of prompting strategies (9–16), each of which varies in complexity, but all of which could easily be applied by a lay practitioner with minimal training in prompt engineering (17). We evaluate these systems using both automatic and manual grading, where manual grading of open-answer questions is performed by the same faculty staff that designed these problems and who have experience grading student answers to them. Subsequently, we conduct a detailed analysis of the generated outputs and their assessment results, considering factors such as the number of courses that would be passed, their distribution across university programs, as well as the effects of the question difficulty and language.

Our results show that AI systems are relatively capable of answering questions used in university assessments. GPT-4 responds correctly to ∼65.8% of questions when aggregating responses across the different prompting strategies using a simple majority vote (i.e., a *knowledge-free* setting that assumes the user would use this tool with no subject knowledge). Furthermore, across the eight prompting strategies, GPT-4 can generate at least one correct response for 85.1% of questions (maximum performance), indicating even greater assessment vulnerability in a setting where a user may have enough subject knowledge to *select* a correct answer even if they cannot produce it. This performance is relatively stable across courses in various scientific disciplines, impacting courses regardless of their subject and size. Importantly, we find that these results indicate that large numbers of university degree programs are highly vulnerable to these tools, with the nonproject components of many core courses being passed in multiple degrees offered by our institution.

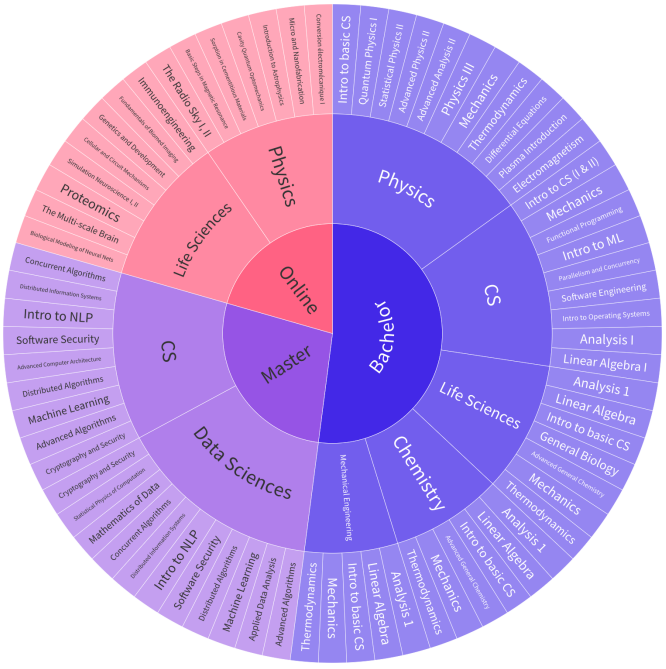

Fig. 1. Overview of Courses. Courses represented in our dataset, grouped by program and degree. Courses may belong to multiple programs, in which case their partition is split into chunks of equal size, with one chunk assigned to each program.

*As of November 2023.

Table 1. Dataset statistics

	Category	Total questions
Level	Bachelor's courses	2,147 (38.5%)
	Master's courses	1,631 (29.2%)
	Online programs	1,801 (32.3%)
Language	English	3,933 (70.5%)
	French	1,646 (29.5%)
Question type	MCQ	3,460 (62%)
	Open-answer	2,119 (38%)

Finally, we observe that while these systems are capable of reaching passing grades in many university assessments, they struggle with more complex question types where students also tend to perform most poorly. Taken together, these results indicate a possibility that these systems could be used to achieve passing marks in university courses while circumventing the process by which students acquire basic domain knowledge and learn to extend it to more complex problems. We conclude with a discussion on mitigations to university assessment settings, an outlook on how university systems should adapt to the increased use of these tools, and a discussion of limitations of our study, specifically with respect to how it diverges from exact assessment and grading policies at our institution.

Data Collection

We compile a dataset of assessment questions from 50 courses offered at our institution from both on-campus and online classes. Following data preprocessing and filtering steps, this dataset consists of a total bank of 5,579 textual multiple-choice (MCQ) and open-answer questions in both English and French. These questions span various levels (e.g., Bachelor, Master), and cover a broad spectrum of STEM disciplines, including Computer Science, Mathematics, Biology, Chemistry, Physics, and Material Sciences. Table 1 and Fig. 1 provide an overview of the dataset's main statistics and the distribution of questions across different topics. Additionally, we have collected course-specific attributes such as the year when the course is first offered in our institution's degree programs (e.g., *Master's year 1*), the program designation (e.g., *Physics*), the language of instruction (e.g., *French*), and the average student enrollment over recent years. Finally, certain questions have been labeled by the instructor who designed the question with a subjective annotation of the question's difficulty.

Experimental Findings

In our study, we investigate the vulnerability of university programs to generative AI systems using our question bank of 5,579 evaluation questions from 50 courses. We consider two models, GPT-4 and GPT-3.5, selected due to their popularity and usage rates. GPT-4 is considered the most performant model among all publicly accessible LLMs but is only available through a premium subscription, impeding its use by many students. GPT-3.5 is a less performant alternative, but free to use. We generate responses to questions from these models using eight relatively easy-to-apply prompting methods (implementation details are described in *SI Appendix*, section 2). For multiple-choice questions, we assess whether a response is correct by comparing the selected choice with the annotated correct answer option. For open-response questions, we use GPT-4 to rate the quality of the response on a four-point scale: Correct, Mostly

Correct, Mostly Incorrect, Incorrect, which we map to scores of 1, 0.66, 0.33, and 0, respectively, for calculating performance.[†]

Can LLM Systems Pass University-Level Courses? We begin our analysis by assessing model performance in a setting where the user has zero knowledge about the question topic. In the simplest scenarios, where we use the most straightforward prompting strategies such as directly asking a question (zero-shot) or asking the model to provide a reasoning chain before answering the question (zero-shot chain-of-thought), GPT-4 achieves average accuracies of 55.9% and 65.1%, respectively. With a slightly more complex zero-knowledge strategy, such as majority voting over the eight answers generated by the different prompting strategies, they would receive a correct answer to 65.8% (on average) of questions using GPT-4 (and 52.2% using GPT-3.5). We observe that this performance trend holds regardless of the language of the assessment, with English exhibiting slightly better results than French. Further experimental results for assessments in English and French are detailed in *SI Appendix*, section 5.C.

Beyond overall performance across the question bank, Fig. 2 presents the proportion of passed courses for our sample of 50 courses based on varying passing thresholds. Alarmingly, GPT-4 can easily be used to reach a 50% performance threshold (which could be sufficient to pass many courses at various universities) for 89% of courses with MCQ-based evaluations and for 77% of courses with open-answer ones. While not performing quite as well, GPT-3.5, the freely available model, can reach a 50% threshold for 70% of courses with MCQ-based assessments and for 50% of courses with open-answer questions. As passing thresholds may not be set to 50% for all institutions, we vary this threshold and find that GPT-4 still passes 68% of courses at a 60% passing threshold, and 38% of courses at a 70% passing threshold for MCQ. Similar results are found for open-answer questions.

Our results suggest that users with no knowledge of a particular subject could solve enough questions to pass nonproject assessments in a majority of the courses in our dataset. While these observations make a compelling argument that AI assistants could potentially augment student learning as support tools, they simultaneously indicate a credible short-term risk to educational systems if institutions are not adapted to protect against the misuse of these technologies. Finally, we expect this problem to only grow worse over time, as continual model improvements in the years to come will make these tools even more performant in academic contexts.

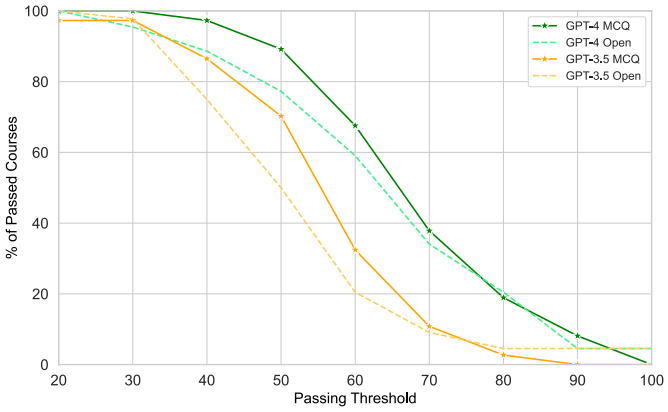

Fig. 2. Course Pass Rate of Generative AI Assistants. Proportion of 50 courses that models successfully pass at various performance thresholds. Results are presented independently for multiple-choice (MCQ) and open-answer (Open) question types for both GPT-3.5 and GPT-4. Model responses are aggregated using the majority vote strategy.

[†]Analysis of the quality of this automated grading is provided in *Materials and Methods* and *SI Appendix*, section 4. Importantly, we note that GPT-4 gives slightly higher average grades (on average ~2.75%) than humans for responses to a sample of questions graded by both.

Table 2. Program results

Program	% Courses passed $\tau = 50\%$	$\tau = 60\%$	$\tau = 70\%$	Max	Question count
Engineering	80.0	60.0	40.0	0.83	1,343
Chemistry	83.3	66.7	50.0	0.85	1,417
Life science	85.7	71.4	57.1	0.85	1,477
Physics bachelor	100.0	55.6	33.3	0.86	958
CS bachelor	91.7	66.7	50.0	0.87	1,487
CS master	100.0	83.3	50.0	0.87	1,514
Data science master	90.0	70.0	30.0	0.86	1,576
Physics online	100.0	63.6	27.3	0.84	837
Life science online	85.7	71.4	57.1	0.75	996

For each program, the first three columns show the percentage of courses for which GPT-4 surpasses the thresholds $\tau = 50, 60, 70\%$ correctly answered questions using the majority vote strategy. "Max" represents the proportion of questions in this degree correctly answered by at least one prompting strategy. Program levels are specified as Bachelor, Master, or Online. The first three programs (Engineering, Chemistry, Life Science) are first-year Bachelor course programs.

How Do These Results Affect University Programs? The average performance across courses demonstrates each course's potential vulnerability to generative AI tools, which is particularly important if considerable portions of degree programs can be completed using these tools. To evaluate this program vulnerability, we aggregate the questions in our datasets according to the study programs in which they are core courses. Specifically, we include four program types: first-year Bachelor, Full Bachelor, Full Master, and Online courses. We separate the first year of the Bachelor's degree because, at many institutions (including ours), the first year of the Bachelor's may have a fairly standardized curriculum that serves a special purpose (e.g., replacing or complementing entrance exams). Full Bachelor's and Master's correspond to regular Bachelor's and Master's programs. We also include online courses, as official certifications can often be awarded for completing a sequence of these courses. For each program, our dataset contains a sample of courses that cover from 33% to 66% of the required courses for that program. For more program statistics, see *SI Appendix*, section 3.A.

We consider the same two aggregation strategies across the responses provided by the eight prompting methods: majority vote and maximum performance. For the majority vote, given the eight prompting strategies we have, the final answer to the question is the one that is the most frequent across all strategies. In the maximum performance aggregation, only a single prompting strategy is required to answer correctly for the model to be deemed correct in its response, approximating a pseudo-oracle setting that remains contextually realistic, as a user might be able to distinguish the answer when presented with it, even if they could not find it on their own.

In Table 2, we present the number of courses that would be passed by GPT-4 across the 9 tested degree programs for various course passing thresholds τ (i.e., the performance that must be achieved to consider the course passed). Our results show that the general course vulnerability observed in the previous section extends to program vulnerability. At the $\tau = 50\%$ threshold for

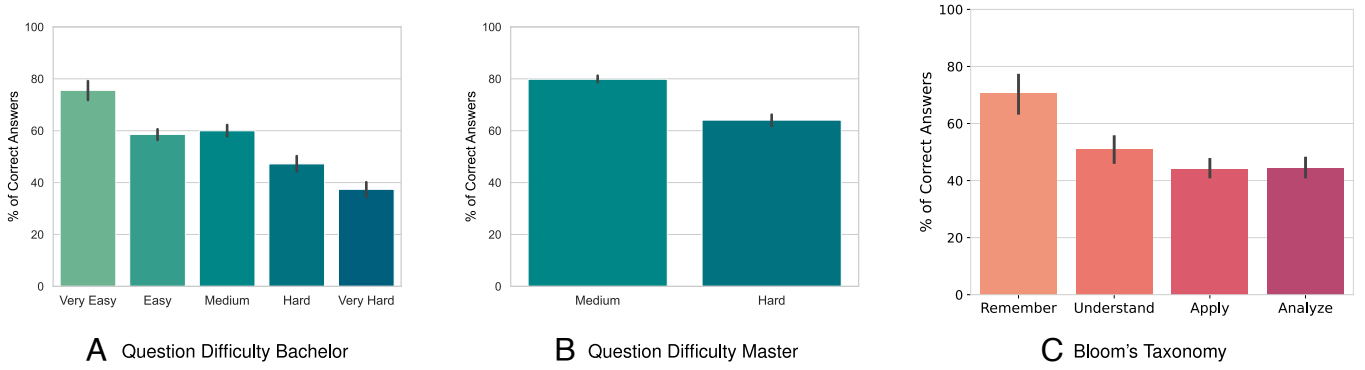

Fig. 3. Model Performance Stratified by Question Difficulty. (*A* and *B*) 376 Bachelor's and 693 Master's questions, respectively, annotated using instructor-reported difficulty levels. (*C*) 207 questions annotated using Bloom's taxonomy by two researchers in the learning sciences. Across all categorization schemes, GPT-4 performance slightly degrades as the questions become more complex and challenging. Performance is aggregated by the majority vote strategy. Error bars represent 95% CIs using the nonparametric bootstrap with 1,000 resamples.

passing a course, at least 83% of courses are passed in each of the evaluated programs. In certain programs, such as the Physics Bachelor and Computer Science Master, all tested courses are passed. While this number drops as we raise the passing threshold τ, the maximum performance for each program typically remains above 80%, indicating that a combination of clever prompting and partial subject knowledge may be sufficient to achieve high marks on assessment questions.

Topically, we find that the models consistently exhibit lower performance on assessments of courses involving mathematical derivations. Conversely, the model demonstrates strong performance on problems that have straightforward generation formats, such as text or code. For example, models consistently yield high performance in subjects such as *Software Engineering* and *Intro to Machine Learning*. This observation is further supported by the difference in performance between Master's and Bachelor's level courses (shown across Fig. 3 *A* and *B*). In our dataset, Bachelor's courses feature more mathematical derivations, while Master's courses have more text and code-based problems. In *SI Appendix*, section 5.A, we provide further performance comparisons between the courses representing each program. In *SI Appendix*, section 5.B, we analyze model performance across all prompting strategies and both question types.

Finally, we highlight that these models are effective in courses that large portions of the student body must take, increasing the overall vulnerability of course programs. Fig. 4 demonstrates that some of the largest classes on campus, with upward of 300 students, are also some of the most vulnerable, with GPT-4 achieving (using the majority vote strategy) an average performance of 69.9% in these classes (while hovering around 60% for other class sizes). This result is particularly problematic because larger courses are often limited in terms of the monitoring and mitigation strategies they can implement due to the number of students. While smaller courses may more easily be able to combat the misuse and unethical use of generative AI tools, larger courses, which are often mandatory for degree completion, must ensure fair and scalable solutions for a larger student population.

More Challenging Assessments Are Only a Half-Solution. One possible solution to mitigate assessment vulnerability would be to increase their difficulty beyond what generative AI systems are capable of solving, as we observe that the performance of these systems does decrease on more challenging assessment questions (Fig. 3). We measure the difficulty using a subsample of our question bank that is annotated according two different categorizations of their difficulty: 1) *instructor-reported question difficulty*, a five-point difficulty categorization for Bachelor courses and two-point for Master's courses provided by the course instructors, and 2) *Bloom's taxonomy* (18), a six-point scale that measures the cognitive complexity of a question.[‡]

For the instructor-reported difficulty categorization, we collect annotations from course instructors for a subset of 376 questions from the Bachelor's program (n.b., the instructors that designed the questions). The instructor-reported rating ranges from "Very Easy" to "Very Hard" on a 5-point scale. We also collect 693 questions from the Master's program annotated on a 2-point scale, ranging from "Medium" to "Hard" (the original scale was meant to be 3-point, but no instructor reported an "Easy" question). In Fig. 3 *A* and *B*, we show the model's performance on questions stratified by their difficulty rating and observe that GPT-4 performs worse on questions that instructors deem harder. For example, in Bachelor courses, there is a 38% difference in accuracy between "Very Easy" and "Very Hard" questions. However, the differences between Bachelor's "Easy"

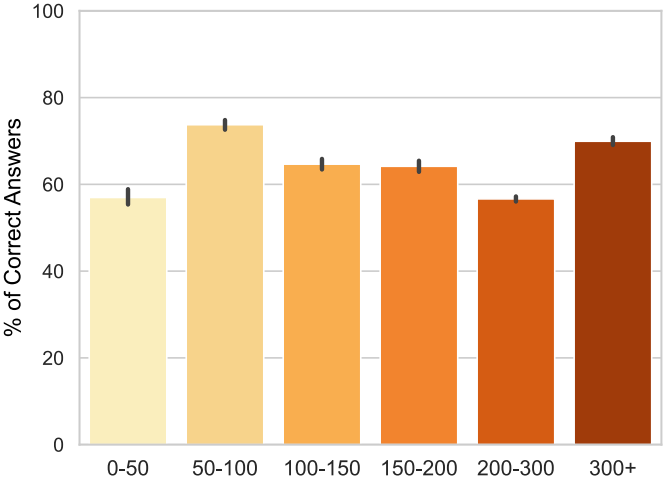

Fig. 4. Course Performance by Course Size. Average course performance of GPT-4 with the majority vote strategy stratified by the course size, measured by the number of enrolled students. GPT-4 successfully answers questions for assessments in some of the largest courses by enrollment, amplifying the potential impact of assessment vulnerability. Error bars represent 95% CIs using the nonparametric bootstrap with 1,000 resamples.

[‡]More details about Bloom's Taxonomy can be found in *SI Appendix*, section 3.B.

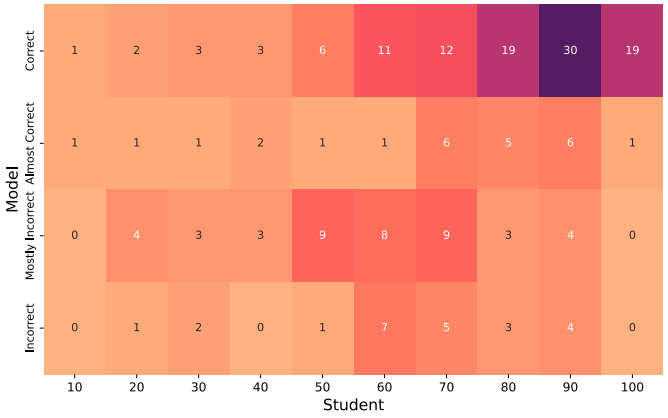

Fig. 5. Comparison of Student Performance and GPT-4. Average student performance for a subset of 197 questions is computed and stratified along 10-point intervals from 0 to 100. The model's performance with the majority vote strategy is assessed by human graders using a 4-point scale. We observe the model typically correctly answers questions that students also answer correctly.

and "Hard" questions or Master's "Medium" and "Hard" questions are only 11.5% and 15.75%, respectively.

This pattern is also observed in our assessment of question difficulty performed using Bloom's taxonomy, which classifies educational learning objectives into levels of complexity and specificity: remember, understand, apply, analyze, evaluate, and create. Two researchers in the learning science manually annotated 207 questions from our dataset according to Bloom's taxonomy (18). While the taxonomy typically associates questions into six categories, we found that most course assessment questions were covered by the first four categories. The results, grouped by taxonomy category in Fig. 3*C*, illustrate that GPT-4 performance is negatively correlated with the cognitive complexity of the question, with higher performance on lower-level tasks compared to higher-level analysis and application tasks.

However, harder assessments may not necessarily be a suitable solution for this vulnerability as they would also likely lead to lower student performance. When comparing the performance of students and GPT-4 on problem sets from a subset of questions for which student performance statistics were collected (Fig. 5), we note that the model tends to excel on questions where students also perform well. This pattern perhaps exacerbates fairness as GPT-4 (and similar models) could be used to achieve average results on an assessment while providing few benefits to students who can already complete the easier portion of assessments but struggle with harder questions. Notably, however, we observe that for a subset of problems, the model typically struggles, receiving "Mostly Incorrect" or "Incorrect" marks, while students demonstrate relatively strong performance, with scores ranging from 0.5 to 0.9. These problems typically require mathematical derivations and extensive computations.

Discussion

Summary. In this work, we tested the ability of LLMs to solve assessment questions for a large number of courses from technical and natural sciences academic programs at EPFL. We find that LLMs are generally capable of answering 50 to 70% (depending on the model) of questions correctly given no subject-related knowledge, and up to 85.1% of questions correctly when some subject-specific knowledge is assumed (i.e., the ability to recognize the correct answer). Most importantly, when considering performance across programs, GPT-4 can achieve greater than 50% performance for 83% to 100% of courses (depending on the program) with an average program pass rate of 91.7%. While a higher per-course passing threshold of 70% would only result in 23% to 50% of courses being passed across our programs (with an average of 37%), this would also lead to higher student fail rates as passing thresholds would similarly affect them. Given that continuous advancements in LLM technology will likely further improve these LLM performance numbers in the future, we conclude that higher-education assessment schemes are immediately vulnerable to student exploitation of generative AI tools, specifically in the engineering and natural sciences.

Assessment Vulnerability. Our results indicate that the broad availability of generative AI tools should urgently trigger discussion on the design and implementation of assessments. Naturally, our results must be placed in the context where they would normally be observed. In many educational institutions, student assessments are frequently closed-book, thereby precluding the direct use of generative AI tools. Many course assessments (e.g., assignments), though, are completed at home without

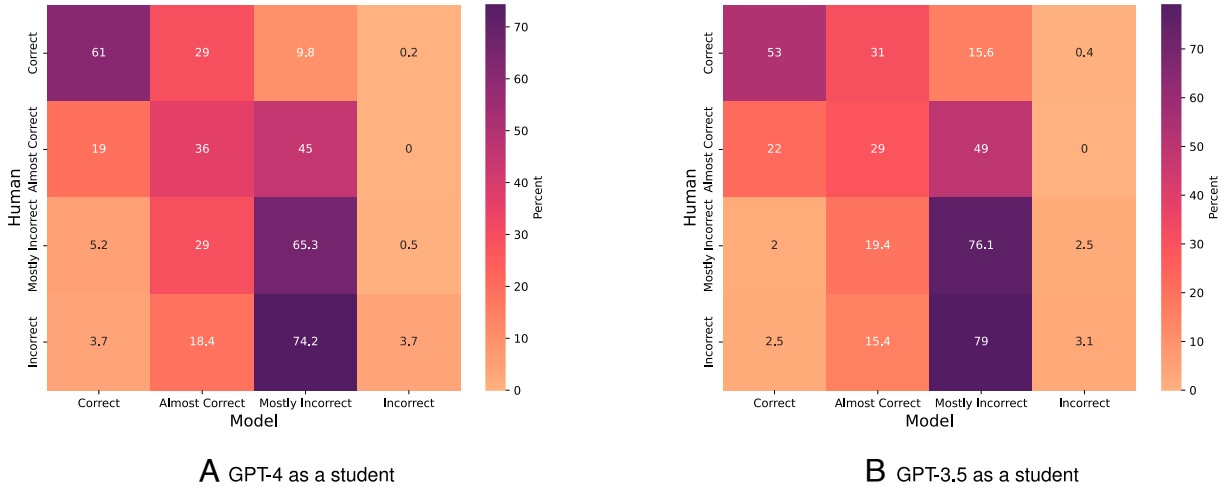

Fig. 6. Comparison of Human and GPT-4 grading. Average model and human performance for a subset of 933 questions and answers from (*A*) GPT-4 and (*B*) GPT-3.5 generated with the metacognitive prompting method.

supervision. In the same vein, most online courses typically evaluate students without any supervised, in-person assessment. In these unsupervised settings, the availability of generative AI tools for aiding in the completion of assessments poses risks for many commonly used student evaluation methods.

One general trend that we observe from our data (Fig. 3 *A*–*C*) is that models perform well on basic learning tasks, such as memorizing factual knowledge. In courses where memorization of factual knowledge is a core evaluation objective, students should not be allowed to use these tools in nonproctored settings, and these courses should perhaps return to traditional in-person examination (19). Barring this possibility, in the case of nonproctored assessments, we recommend that their design should not only consider the possibility of assistance from generative AI but actively assume its use. At the very least, assessments should be designed with generative AI in the loop to design AI-adversarial evaluations that remain fair to students.

At the same time, these findings provide an opportunity to improve and diversify student learning through assessments. Students acquire relevant skills when assessments emphasize analytical and applied knowledge settings (20). Rather than using proctored exams, then, or limited practical works, such as assignments, students should be evaluated using assessments requiring a more composite application of course concepts, such as broader class projects. Project settings more closely assess students on problems resembling real-world challenges, would provide students with more opportunities to make problem-solving decisions, such as problem simplification, decomposition, and planning (21), and would mitigate the impact of generative AI tools (Fig. 3*C*).

Education Vulnerability. While our results point to an urgent need to mitigate assessment vulnerabilities in higher education, a longer-term view requires considering how education as a practice should evolve alongside the availability of generative AI tools. Since the release of ChatGPT, ongoing discussions have revolved around this topic with both negative and optimistic views. Although numerous studies explore the ways AI can enhance learning, ethical concerns related to plagiarism, biases, and overreliance on technology have also been highlighted (22–28).

An important dimension of these discussions emphasizes the need to revisit evaluation procedures to ensure that students acquire necessary skills and critical thinking abilities in the face of AI adoption (29–32). For instance, observations from various works (and our study) show that models excel in generating code to solve software problems (33–37). While this capability reduces the burden on professional (and hobbyist) software developers, it also poses a risk for learners by potentially offering shortcuts that impede the acquisition of fundamental coding and engineering skills (38). Coding tools such as GitHub's Copilot or OpenAI's Codex may lead novices to overrely on autosuggested solutions. This overreliance may diminish their engagement with computational thinking (29, 30), which is arguably the most important skill that is learned in any computer science course or program.

Beyond this example, many studies underscore the significance of adapting course materials and assessments to promote critical thinking, encourage student collaboration, develop practical skills, enhance soft skills, and promote interdisciplinary learning, all with the aim of cultivating graduates equipped with a diverse range of competencies (32, 39–41). In particular, reinforcing our conclusions above, open-ended assessments are proposed to promote originality and creativity, potentially discouraging reliance on generative AI tools and fostering unique ideas and critical analysis (41, 42). One example of program reconsideration is teaching students at computer science courses prompt engineering, which would be essentially programming in natural language (43). This would prioritize problem-solving and higher-level concepts over the technical syntax of programming languages. Ultimately, many of these studies suggest the greater risk of generative AI may be its potential to enable the unintentional circumvention of the frameworks by which learners are taught the foundations to learn later skills, and that teaching and assessment should be adapted for this risk.

Finally, assuming that students will use and become acquainted with the capabilities of these technologies, we recommend that students should not only be educated on the technical and ethical challenges of generative AI systems but also on the critical thinking required to successfully engage with such tools (44). One such measure could involve establishing committees for ethical oversight and adding classroom discussions on the use of AI tools. Such discussions would clarify what constitutes plagiarism and address potential ethical concerns, ensuring that students understand the importance of academic integrity and discern the boundaries between legitimate assistance and academic misconduct (31, 38–42).

Limitations

While our study offers preliminary insights into the vulnerability of degree programs to student use of AI assistants for assessments, we acknowledge several limitations in our study.

First, our study excluded any multimodal inputs, such as questions containing diagrams, figures, or graphs, which were omitted from our dataset. Approximately 57% of the initially collected data did not qualify for inclusion in the final dataset of 5,579 questions. Consequently, models were solely evaluated with text-only questions. This approach likely resulted in performance outcomes that are higher than what these models would attain when tested on question sets that include these other modalities, though we also note rapid growth in the multimodal capabilities of these models (45).

Our results for GPT-4's performance on open-answer questions may have also been slightly overestimated because we also used GPT-4 model as a grader As this dual use of GPT-4 could introduce potential grading bias (46), we compared the grades provided by GPT-4 to human scores on a subset of the questions. When comparing the alignment between human-assigned and model-assigned grades for responses from both GPT-4 and GPT-3.5, our results show minimal bias toward GPT-4's responses relative to GPT-3.5.

Simultaneously, our findings might underestimate the performance potential that students could attain through collaboration with these systems. Although we conducted a thorough examination of prompting strategies, our methods are limited by the fact that they 1) rely solely on published prompting strategies, 2) are generally noninteractive, and 3) are tailored for scalability across all questions to facilitate a comprehensive study. Students aiming to address individual questions could devote more time and devise more interactive, less standardized prompting strategies to collaboratively guide the models toward improved solutions.

Finally, we acknowledge certain gaps between our evaluation of AI systems in this study, and how students are normally evaluated in these courses. First, our study standardizes system evaluation across all course assessments, removing course-specific assessment policies for questions. For example, certain courses, beyond not giving points for correct answers to multiple-choice

questions, might also penalize incorrect answers more than leaving a question unanswered, while our study simply gives zero points for incorrect answers. Second, our dataset of questions for each course is not necessarily balanced according to a course's grading rubric. As an example, our dataset may contain a balanced mixture of questions from assignments and exams for a particular course, while the overall evaluation of a student in this same course would compute their grade as a 10% mixture of assignments, and 90% mixture of exam questions. Similarly, many courses at our institution also include lab or project components as part of their final grade. Since these parts of the assessment do not have a "correct answer" that can be easily marked, they are not included in our dataset.

As we do not consider these course-specific assessment policies when computing the course pass rates of our tested AI assistants, these design decisions introduce a gap between our evaluation and the actual assessment rubrics by which students are graded in our institution's courses. Despite this divergence, however, we note that other institutions may implement course assessments and grading rubrics in different ways. As a result, while our study is not an exact simulation of our institution's diverse assessment schemes among its courses, it remains a suitable testbed for providing insights into how course assessments are vulnerable to AI assistants, and how this vulnerability would extend to full university programs without mitigations.

Materials and Methods

In this section, we provide further details on our data collection process, the prompting strategies used for response generation, and our pipeline for automated grading.

Dataset Collection. Our data collection was approved by the Human Research Ethics Committee at EPFL. Data were voluntarily submitted by members of the Data Consortium, and no materials were used without the permission of the data owner.

Dataset Preprocessing. To preprocess our data, we collect assessments from participating faculty, extract questions and answers from these assessments, and standardize them into a uniform format. After compiling an initial question bank from the raw data, we filter unsuitable data points by 1) removing questions that lack the question body or solution, 2) eliminating duplicate questions, and 3) removing questions that require information that cannot be parsed by LLMs in a textual format (e.g., diagrams, images, plots). In cases where a joint context is provided for multiple questions, we augment each question individually with this context.

Prompting Strategies. To generate answers to questions, we employ various prompting strategies requiring only familiarity with relevant literature and minimal adaptation. We selected eight distinct prompting strategies that we broadly categorize into three types: direct, rationalized, and reflective prompting. Under direct prompting, we use zero-shot, one-shot (9), and expert prompting (10), where models are directly prompted for an answer without encouraging any underlying rationale. For rationalized prompting, three strategies are implemented: zero-shot (12) and four-shot chain-of-thought (11), and tree-of-thought (13) prompting. Here, language models are prompted to generate a rationale before providing the final answer. Last, reflective prompting includes self-critique (14, 15) and metacognitive prompting (16), where models are asked to reflect on a previously provided answer and adjust their response according to this reflection. In our experiments, we noted that the prompting strategy substantially influences model performance, with at least one strategy consistently producing the correct answer in the majority of cases. A detailed description of all prompting strategies, along with prompts, is provided in *SI Appendix*, section 2.

Evaluation. In this section, we outline the grading strategies used to evaluate the model's performance across two question types: multiple-choice (MCQ) and open-answer questions. For MCQ, grading is automated by comparing against the annotated answer. Answers receive a binary score of 0/1 if only one correct option exists, or a proportional score based on the number of correct choices made for multianswer questions (with no penalty for wrong choices). *SI Appendix*, section 4.A provides more details for grading MCQs. For open-answer questions, we constructed a multistep evaluation pipeline using GPT-4 as a grader (7), which we describe below. For both types of results, we report error bars representing 95% CIs (Figs. 3 and 4). These intervals were computed using the nonparametric bootstrap with 1,000 resamples. We also tasked human experts with independently grading a subset of model responses to measure alignment between model and human grading and establish a confidence level for model-based grading.

Automated grading. A significant portion of the questions we extracted are open-answer questions, which are challenging to evaluate manually due to the sheer volume of responses (a total of 33,904 answers from 2,119 questions, answered by 2 models using 8 prompting strategies). As a result, we use a state-of-the-art LLM, GPT-4, as a grader. To automate the grading of open answers, we provide the model with the question, the correct solution from an answer sheet of the assessment, and the generated answer text, prompting it to assign a rating based on the quality of the response. We provide the model with a 4-point grading scale: Correct, Mostly Correct, Mostly Incorrect, Incorrect. The model is first tasked with assessing the accuracy and completeness of the answer before assigning the final grade. Although we do not use these interim accuracy and completeness scores, we manually observe that these stages enhance the quality of overall grading. The specific prompting strategy is detailed in *SI Appendix*, section 4.B. As an answer was produced for each question using eight distinct prompting strategies, we obtained eight different answers and corresponding grades. To present a cohesive performance score for both GPT-4 and GPT-3.5 for a given question, we employ two aggregation methods: 1) the *maximum* approach, which selects the answer with the highest grade for each question as a representation of model performance, and 2) the *majority* approach, which considers the grade that appears most frequently among the eight prompting strategies. As an example, for a hypothetical question whose generated answers for the eight prompting strategies received 2 Correct, 2 Mostly Correct and 4 Mostly Incorrect grades, the *maximum* grade would be Correct and the *majority* grade would be Mostly Incorrect. To report dataset-level performance, we map grade ratings for each example to a score from a discrete range between 0 and 1: {Correct: 1.0, Mostly Correct: 0.66, Mostly Incorrect: 0.33, Incorrect: 0.0} and average the scores.

Human grading. To assess how well model grading aligned with human grading on open-answer questions, we enlisted 28 expert annotators from the teaching faculty of 11 courses to evaluate 933 questions. The courses chosen for human expert grading are listed in *SI Appendix*, section 4.C. Specifically, we requested graders to assign scores to open-ended responses generated by GPT-4 and GPT-3.5. Responses for human grading for both models were generated using two prompting strategies: zero-shot chain-of-thought prompting (11) (a simple prompting method at the disposal of any student) and metacognitive prompting (16) (one of the most effective strategies across all courses). We anonymized the outputs to prevent graders from knowing which model and prompting strategy they were evaluating. To maintain consistency, we instructed graders to use the same grading scale as GPT-4's direct grading. The number of graders per course varied from 1 to 10, and a total of 3,732 answers were evaluated. On average, graders spent approximately 5 min assessing each answer.

Fig. 6 indicates a general alignment between human graders and GPT-4 when categorizing answers into a simplified correct/incorrect quadrant. Out of the examples identified as Correct by graders, the model assigned the same grade to 61% of them. Similarly, for examples graded as Almost Correct by graders, the model's grade matched in 36% of cases. Additionally, in instances where graders labeled examples as Mostly Incorrect, the model's grade aligned with the grader's assessment 65% of the time. However, we note certain patterns of discrepancy. For instance, GPT-4 as a grader tends to avoid explicitly labeling solutions as Incorrect, and instead opts for Mostly Incorrect (i.e., in 74% of cases that humans annotated a solution as Incorrect, the model identified it as Mostly Incorrect), potentially due to the practice of aligning

models for harmlessness (47). We find a few instances where the model rates an answer as Correct while humans assign a lower score.

Interestingly, upon comparing average grades assigned by human graders and GPT-4 across 11 courses, we find a difference in average grade of only 2.75%. However, we observe variations between courses, with an average course grade deviation of 8.5% (and the largest deviation for a course being 26%) between human and model graders. Finally, we also note the performance correlation between MCQ and open-answer questions in Fig. 2, providing a comparison point for the rationality of our model-based open-answer grading results. While scores for open-answer questions are typically lower than MCQ, the patterns exhibited by both curves are similar across both models. Overall, we note that the grades provided by humans and models are moderately correlated and that the summary statistics across courses tend to have a high correlation. Additional results, such as further comparison of human and model grades for additional prompting strategies, pairwise agreement scores between human and model graders, and qualitative human assessments of the responses for both GPT-3.5 and GPT4, can be found in *SI Appendix*, section 4.

Automated grading in prior work. A substantial body of research leverages LLMs for response evaluation. Traditionally, automated assessment has necessitated high-quality reference data obtained through human grading, which is both costly and time-intensive. Consequently, there has been considerable exploration into the potential of LLMs to serve as evaluators (48). Recent research has found LLMs to be capable of generating quality feedback (15, 49–55), a trend also reflected in investigations into LLM-based evaluation (7, 56–59).

Automated solutions for student grading have been explored in the field of learning science, as well, some of which now use LLMs (60). Intelligent Tutoring Systems, such ALEKS (61), ASSISTments (62), Cognitive Tutor (63), and MATHia (64) are widely employed to automatically assess student performance in closed-ended questioning. These systems cater to several hundred thousand students annually (62, 65). Meanwhile, AES platforms such as e-Rater (66), IntelliMetric (67), and Intelligent Essay Assessor (68) have emerged as useful tools for evaluating numerous student essays and responses to open-ended questions each year (67–71).

Data, Materials, and Software Availability. The data and code are available under an open-source license to facilitate further research and collaboration within the community. The course data, model responses, and code can be accessed at the GitHub repository (https://github.com/epfl-nlp/nlp4education) (72). The dataset has been anonymized to ensure compliance with privacy regulations.

ACKNOWLEDGMENTS. A.B. gratefully acknowledges the support of the Swiss NSF (No. 215390), Innosuisse (PFFS-21-29), Sony Group Corporation, and the Allen Institute for AI. P.S. acknowledges support from the NCCR Catalysis (grant number 180544), a National Centre of Competence in Research funded by the Swiss NSF. T.K. is partially funded by the Swiss State Secretariat for Education, Research, and Innovation (No. 591711).

Author affiliations: [a]École Polytechnique Fédérale de Lausanne (EPFL), Lausanne 1015, Switzerland

Author contributions: B.B., N.F., D.B., A.S., S. Friedli, P.J., T.K., and A.B. designed research; B.B., N.F., D.B., A.S., S.M., T.N., M.B., A.S., P.S., S.P.N., J.F., and P.S. performed research; E.D.C. contributed new reagents/analytic tools; B.B., N.F., D.B., A.S., S.M., A.R., G.W., S.M., Z.C., S. Fan, S.G., M.I., D.P., P.S., S. Friedli, P.J., T.K., A.B., and E.G.C. analyzed data; E.D.C contributed data for the research; and B.B., N.F., D.B., A.S., S.M., and A.B. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

[1]B.B., N.F., D.B., and A.S. contributed equally to this work.

1. D. Hendrycks *et al.*, Measuring massive multitask language understanding. arXiv [Preprint] (2021). https://arxiv.org/abs/2009.03300 (Accessed 12 January 2021).
2. Y. Huang *et al.*, C-eval: A multi-level multi-discipline Chinese evaluation suite for foundation models. arXiv [Preprint] (2023). https://arxiv.org/abs/2305.08322 (Accessed 6 November 2023).
3. X. Wang *et al.*, Scibench: Evaluating college-level scientific problem-solving abilities of large language models. arXiv [Preprint] (2023). https://arxiv.org/abs/2307.10635 (Accessed 28 June 2024).
4. W. Zhong *et al.*, "AGIEval: A human-centric benchmark for evaluating foundation models" in *Findings of the Association for Computational Linguistics: NAACL 2024*, K. Duh, H. Gomez, S. Bethard, Eds. (Association for Computational Linguistics, Mexico City, Mexico, 2024), pp. 2299–2314.
5. D. Arora, H. Singh, Mausam, "Have LLMs advanced enough? A challenging problem solving benchmark for large language models" in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, H. Bouamor, J. Pino, K. Bali, Eds. (Association for Computational Linguistics, Singapore, 2023), pp. 7527–7543.
6. OpenAI, GPT-4 technical report. arXiv [Preprint] (2023). https://arxiv.org/abs/2303.08774 (Accessed 4 March 2024).
7. L. Zheng *et al.*, Judging LLM-as-a-judge with MT-bench and chatbot arena. *Adv. Neural Inf. Process. Syst.* **36**, 46595–46623 (2024).
8. OpenAI, GPT-3.5. https://platform.openai.com. Accessed 10 November 2023.
9. T. Brown *et al.*, "Language models are few-shot learners" in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M Balcan, H. Lin, Eds. (Curran Associates, Inc., 2020), vol. 33, pp. 1877–1901.
10. B. Xu *et al.*, Expertprompting: Instructing large language models to be distinguished experts. arXiv [Preprint] (2023). https://arxiv.org/abs/2305.14688 (Accessed 24 May 2023).
11. J. Wei *et al.*, "Chain-of-thought prompting elicits reasoning in large language models" in *Proceedings of the 36th International Conference on Neural Information Processing Systems (NIPS '22)*, S. Koyejo *et al.*, Eds. (Curran Associates Inc., Red Hook, NY, 2024), pp. 24824–24837.
12. T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, Y. Iwasawa, Large language models are zero-shot reasoners. arXiv [Preprint] (2022). https://doi.org/10.48550/arXiv.2205.11916 (Accessed 29 January 2023).
13. S. Yao *et al.*, "Tree of thoughts: Deliberate problem solving with large language models" in *Proceedings of the 37th International Conference on Neural Information Processing Systems (NIPS '23)*, A. Oh *et al.*, Eds. (Curran Associates Inc., Red Hook, NY, 2024), pp. 11809–11822.
14. R. Wang *et al.*, "Enhancing large language models against inductive instructions with dual-critique prompting" in *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, R. Cotterell, M. Sap, L. Huang, Eds. (Association for Computational Linguistics, Mexico City, Mexico, 2024), pp. 5345–5363.
15. A. Madaan *et al.*, Self-refine: Iterative refinement with self-feedback. arXiv [Preprint] (2023). https://arxiv.org/abs/2303.17651 (Accessed 25 May 2023).
16. Y. Wang, Y. Zhao, "Metacognitive prompting improves understanding in large language models" in *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, K. Duh, H. Gomez, S. Bethard, Eds. (Association for Computational Linguistics, Mexico City, Mexico, 2024), pp. 1914–1926.
17. P. Sahoo *et al.*, A systematic survey of prompt engineering in large language models. Techniques and applications. arXiv [Preprint] (2024). https://arxiv.org/abs/2402.07927 (Accessed 5 February 2024).
18. B. Bloom, D. Krathwohl, *Taxonomy of Educational Objectives: The Classification of Educational Goals* (Longmans, Green, 1956), vol. 1.
19. T. Wang, D. V. Díaz, C. Brown, Y. Chen, "Exploring the role of AI assistants in computer science education: Methods, implications, and instructor perspectives" in *2023 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC)* (IEEE, 2023).
20. L. Zhang, Y. Ma, A study of the impact of project-based learning on student learning effects: A meta-analysis study. *Front. Psychol.* **14**, 1202728 (2023).
21. B. J. Montgomery, A. M. Price, C. Wieman, "How traditional physics coursework limits problem-solving opportunities" in *Physics Education Research Conference 2023* (Sacramento, CA, 2023), pp. 230–235.
22. L. Yan *et al.*, Practical and ethical challenges of large language models in education: A systematic scoping review. *Br. J. Educ. Technol.* **55**, 90–112 (2023).
23. S. Chen, Generative AI, learning and new literacies. *J. Educ. Technol. Dev. Exch.* **16**, 1–19 (2023).
24. A. Pinto, A. Abreu, E. Costa, J. Paiva, How machine learning (ML) is transforming higher education: A systematic literature review. *J. Inf. Syst. Eng. Manag.* **8**, 21168 (2023).
25. T. Alqahtani *et al.*, The emergent role of artificial intelligence, natural learning processing, and large language models in higher education and research. *Res. Soc. Adm. Pharm.* **19**, 1236–1242 (2023).
26. G. Currie, Academic integrity and artificial intelligence: Is chatGPT hype, hero or heresy? *Semin. Nuclear Med.* **53**, 719–730 (2023).
27. Y. K. Dwivedi *et al.*, Opinion paper: "so what if chatGPT wrote it?" Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. *Int. J. Inf. Manag.* **71**, 102642 (2023).
28. Y. Lan *et al.*, Survey of natural language processing for education: Taxonomy, systematic review, and future trends. arXiv [Preprint] (2024). https://arxiv.org/abs/2401.07518 (Accessed 15 March 2024).
29. J. Prather *et al.*, "It's weird that it knows what i want": Usability and interactions with copilot for novice programmers. *ACM Trans. Comput. Hum. Inter.* **31**, 1–31 (2023).
30. B. A. Becker *et al.*, "Programming is hard–Or at least it used to be: Educational opportunities and challenges of AI code generation" in *Proceedings of the 54th ACM Technical Symposium on Computer Science Education V. 1 (SIGCSE 2023)*, M. Doyle, B. Stephenson, B. Dorn, L.-K. Soh, L. Battestilli, Eds. (Association for Computing Machinery, New York, NY, 2023), pp. 500–506.

31. J. Finnie-Ansley, P. Denny, B. A. Becker, A. Luxton-Reilly, J. Prather, "The robots are coming: Exploring the implications of openAI codex on introductory programming" in *Proceedings of the 24th Australasian Computing Education Conference, ACE '22* (Association for Computing Machinery, New York, NY, USA, 2022), pp. 10–19.
32. R. Nowrozy, D. Jam, Embracing the generative AI revolution: Advancing tertiary education in cybersecurity with GPT. arXiv [Preprint] (2024). https://arxiv.org/abs/2403.11402 (Accessed 18 March 2024).
33. P. Vaithilingam, T. Zhang, E. L. Glassman, "Expectation vs. experience: Evaluating the usability of code generation tools powered by large language models" in *Chi Conference on Human Factors in Computing Systems Extended Abstracts*, S. Barbosa, C. Lampe, C. Appert, D. A. Shamma, Eds. (Association for Computing Machinery, New York, NY, 2022), pp. 1–7.
34. F. F. Xu, U. Alon, G. Neubig, V. J. Hellendoorn, "A systematic evaluation of large language models of code" in *Proceedings of the 6th ACM SIGPLAN International Symposium on Machine Programming*, S. Chaudhuri, C. Sutton, Eds. (Association for Computing Machinery, New York, NY, 2022), pp. 1–10.
35. Y. Li *et al.*, Competition-level code generation with alphacode. *Science* **378**, 1092–1097 (2022).
36. W. Hou, Z. Ji, A systematic evaluation of large language models for generating programming code. arXiv [Preprint] (2024). https://arxiv.org/abs/2403.00894 (Accessed 5 October 2024).
37. J. Liu, C. S. Xia, Y. Wang, L. Zhang, Is your code generated by chatGPT really correct? Rigorous evaluation of large language models for code generation *Adv. Neural Inf. Process. Syst.* **36**, 21558–21572 (2024).
38. P. Denny *et al.*, Computing education in the era of generative AI. *ACM* **67**, 56–67 (2024).
39. M. Alier, F. García-Peñalvo, J. D. Camba, Generative artificial intelligence in education: From deceptive to disruptive. *Int. J. Interact. Multimed. Artif. Intell.* **8**, 5 (2024).
40. I. Chaudhry, S. Sarwary, G. El-Refae, H. Chabchoub, Time to revisit existing student's performance evaluation approach in higher education sector in a new era of chatGPT - A case study. *Cogent Educ.* **10**, 1–30 (2023).
41. D. Cotton, P. Cotton, R. Shipway, Chatting and cheating: Ensuring academic integrity in the era of chatGPT chatting and cheating: Ensuring academic integrity in the era of chatGPT. *Innov. Educ. Teach. Int.* **61**, 228–239 (2023).
42. M. Liu *et al.*, Future of education in the era of generative artificial intelligence: Consensus among chinese scholars on applications of chatGPT in schools. *Future Educ. Res.* **1**, 72–101 (2023).
43. B. N. Reeves *et al.*, Prompts first, finally. arXiv [Preprint] (2024). https://doi.org/10.48550/arXiv.2407.09231 (Accessed 12 July 2024).
44. K. D. Wang, E. Burkholder, C. Wieman, S. Salehi, N. Haber, Examining the potential and pitfalls of ChatGPT in science and engineering problem-solving. arXiv [Preprint] (2023). https://arxiv.org/abs/2310.08773 (Accessed 28 October 2023).
45. X. Yue *et al.*, A massive multi-discipline multimodal understanding and reasoning benchmark for expert AGI. arXiv [Preprint] (2023). https://arxiv.org/abs/2311.16502 (Accessed 13 June 2024).
46. A. Panickssery, S. R. Bowman, S. Feng, LLm evaluators recognize and favor their own generations. arXiv [Preprint] (2024). https://doi.org/10.48550/arXiv.2404.13076 (Accessed 15 April 2024).
47. Y. Bai *et al.*, Constitutional AI: Harmlessness from AI feedback. arXiv [Preprint] (2022). https://arxiv.org/abs/2212.08073 (Accessed 15 December 2022).
48. C.-H. Chiang, H.-y. Lee, "Can large language models be an alternative to human evaluations?" in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, A. Rogers, J. Boyd-Graber, N. Okazaki, Eds. (Association for Computational Linguistics, 2023), pp. 15607–15631.
49. J. Scheurer *et al.*, Training language models with language feedback. arXiv [Preprint] (2022). https://arxiv.org/abs/2204.14146 (Accessed 17 November 2022).
50. S. Welleck *et al.*, Generating sequences by learning to self-correct. arXiv [Preprint] (2023). https://arxiv.org/abs/2211.00053 (Accessed 31 October 2022).
51. N. Tandon, A. Madaan, P. Clark, Y. Yang, "Learning to repair: Repairing model output errors after deployment using a dynamic memory of feedback" in *Findings of the Association for Computational Linguistics: NAACL 2022*, M. Carpuat, M. C. de Marneffe, I. V. Meza Ruiz, Eds. (Association for Computational Linguistics, Seattle, United States, 2022), pp. 339–352.
52. W. Saunders *et al.*, Self-critiquing models for assisting human evaluators. arXiv [Preprint] (2022). https://arxiv.org/abs/2206.05802 (Accessed 14 June 2022).
53. D. Paul *et al.*, "REFINER: Reasoning feedback on intermediate representations" in *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, Y. Graham, M. Purver, Eds. (Association for Computational Linguistics, St. Julian's, Malta, 2024), pp. 1100–1126.
54. T. Schick *et al.*, PEER: A collaborative language model. arXiv [Preprint] (2022). https://arxiv.org/abs/2208.11663 (Accessed 24 August 2022).
55. X. Chen, M. Lin, N. Schärli, D. Zhou, Teaching large language models to self-debug. arXiv [Preprint] (2023). https://arxiv.org/abs/2304.05128 (Accessed 5 October 2023).
56. J. Fu, S. K. Ng, Z. Jiang, P. Liu, GPTScore: Evaluate as you desire. arXiv [Preprint] (2023). https://arxiv.org/abs/2302.04166 (Accessed 13 February 2023).
57. T. Kocmi, C. Federmann, "Large language models are state-of-the-art evaluators of translation quality" in *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, M. Nurminen *et al.*, Eds. (European Association for Machine Translation, 2023), pp. 193–203.
58. J. Wang *et al.*, "Is ChatGPT a good NLG evaluator? A preliminary study" in *Proceedings of the 4th New Frontiers in Summarization Workshop*, Y. Dong, W. Xiao, L. Wang, F. Liu, G. Carenini, Eds. (Association for Computational Linguistics, 2023), pp. 1–11.
59. Y. Liu *et al.*, "G-Eval: NLG evaluation using Gpt-4 with better human alignment" in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, H. Bouamor, J. Pino, K. Bali, Eds. (Association for Computational Linguistics, Singapore, 2023), pp. 2511–2522.
60. H. Hasanbeig, H. Sharma, L. Betthauser, F. V. Frujeri, I. Momennejad, ALLURE: Auditing and improving LLM-based evaluation of text using iterative in-context-learning. arXiv [Preprint] (2023). https://arxiv.org/abs/2309.13701 (Accessed 27 September 2023).
61. J. C. Falmagne, E. Cosyn, J. P. Doignon, N. Thiéry, "The assessment of knowledge, in theory and in practice" in *Formal Concept Analysis*, R. Missaoui, J. Schmidt, Eds. (Springer, Berlin, Heidelberg, 2006), pp. 61–79.
62. N. T. Heffernan, C. L. Heffernan, The assistments ecosystem: Building a platform that brings scientists and teachers together for minimally invasive research on human learning and teaching. *Int. J. Artif. Intell. Educ.* **24**, 470–497 (2014).
63. S. Ritter, S. E. Fancsali, "Carnegie learning's cognitive tutor" in *Educational Data Mining*, O. C. Santos *et al.*, Eds. (International Educational Data Mining Society (IEDMS), 2015).
64. S. Ritter, *The Research Behind the Carnegie Learning Math Series* (Carnegie Learning, Pittsburgh, PA, 2011).
65. V. Aleven, B. Mclaren, I. Roll, K. Koedinger, Toward meta-cognitive tutoring: A model of help seeking with a cognitive tutor. *Int. J. Artif. Intell. Educ.* **16**, 101–128 (2006).
66. Y. Attali, J. Burstein, Automated essay scoring with e-rater. *J. Technol. Learn. Assess.* **4**, 1–31 (2006).
67. L. Rudner, V. Garcia, C. Welch, An evaluation of intellimetric^{SM} Essay scoring system. *J. Technol. Learn. Assess.* **4**, 1–22 (2006).
68. P. Foltz, D. Laham, T. Landauer, The intelligent essay assessor: Applications to educational technology. *Interact. Multimed. Electron. J. Comput. Learn.* **1**, 1–6 (1999).
69. M. Shermis, J. Burstein, *Handbook of Automated Essay Evaluation: Current Applications and New Directions* (Taylor & Francis, 2013).
70. D. Ramesh, S. K. Sanampudi, An automated essay scoring systems: A systematic literature review. *Artif. Intell. Rev.* **55**, 2495–2527 (2022).
71. B. Beigman Klebanov, N. Madnani, "Automated evaluation of writing – 50 years and counting" in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, D. Jurafsky, J. Chai, N. Schluter, J. Tetreault, Eds. (Association for Computational Linguistics, Online, 2020), pp. 7796–7810.
72. B. Borges, N. Foroutan, D. Bayazit, A. Sotnikova, A. Bosselut, Could ChatGPT get an engineering degree? Evaluating higher education vulnerability to AI assistants. NLP4Education. https://github.com/epfl-nlp/nlp4education. Deposited 1 November 2024.