International Journal of Artificial Intelligence in Education (2025) 35:367–397
https://doi.org/10.1007/s40593-024-00403-3

ARTICLE

# GPT-4 in Education: Evaluating Aptness, Reliability, and Loss of Coherence in Solving Calculus Problems and Grading Submissions

**Alberto Gandolfi**[1]

Accepted: 26 March 2024 / Published online: 5 May 2024


## Abstract

In this paper, we initially investigate the capabilities of GPT-3 5 and GPT-4 in solving college-level calculus problems, an essential segment of mathematics that remains under-explored so far. Although improving upon earlier versions, GPT-4 attains approximately 65% accuracy for standard problems and decreases to 20% for competition-like scenarios. Overall, the models prove to be unreliable due to common arithmetic errors.

Our primary contribution lies then in examining the use of ChatGPT for grading solutions to calculus exercises. Our objectives are to probe an in-context learning task with less emphasis over direct calculations; recognize positive applications of ChatGPT in educational contexts; highlight a potentially emerging facet of AI that could necessitate oversight; and introduce unconventional AI benchmarks, for which models like GPT are untrained. Pertaining to the latter, we uncover a tendency for loss of coherence in extended contexts. Our findings suggest that while the current ChatGPT exhibits comprehension of the grading task and often provides relevant outputs, the consistency of grading is marred by occasional loss of coherence and hallucinations. Intriguingly, GPT-4's overall scores, delivered in mere moments, align closely with human graders, although its detailed accuracy remains suboptimal.

This work suggests that, when appropriately orchestrated, collaboration between human graders and LLMs like GPT-4 might combine their unique strengths while mitigating their respective shortcomings In this direction, it is imperative to consider implementing transparency, fairness, and appropriate regulations in the near future.

**Keywords** AI · Large Language Models · ChatGPT · Natural Language Processes · Calculus · Automated Grading

---

✉ Alberto Gandolfi
ag189@nyu.edu

[1] Division of Science, New York University Abu Dhabi, Abu Dhabi 129188, UAE

## Introduction

The natural language processing tool ChatGPT, particularly the one based on the multimodal transformer (Vaswani et al. (2017)) GPT-4, has demonstrated the ability to interpret text input and generate natural language text that addresses a variety of requests (OpenAI (2023)). ChatGPT can produce convincing college papers and fake research-paper abstracts that are difficult for instructors and scientists to detect (Gao et al. (2022); Crothers et al. (2022)), and its human-like capabilities have even inspired the idea of subjecting ChatGPT to cognitive psychological assessment, see Binz and Schulz (2023).

In particular, ChatGPT performs quite well and often outscores the vast majority of human test takers in various domains, see OpenAI (2023); Ibrahim et al. (2023); Srivastava et al. (2022); Suzgun et al. (2022); Chung et al. (2022). This raises concerns about the potential uses and misuses of ChatGPT in education, especially in high schools and colleges.

Despite the quality of ChatGPT's answers in many domains, its output in the field of Mathematics has been less human-like. A recent study based on GPT-3.5 (Frieder et al. (2024)) examined a large number of mathematical problems across all levels (with limited calculus content) and found that ChatGPT's responses, if graded, would receive scores insufficient for passing standard exams. A similar pattern was found for high school math problems in Vietnam (Dao and Le (2023)). Additionally, a comparison across multiple fields showed that, for GPT-3.5, Mathematics is the only field in which ChatGPT's outputs would consistently score worse than human students' answers (Ibrahim et al. (2023)).

To complete the range of topics for which LLM have been tested, we first analyze the potential of GPT-3.5, but more importantly the more recent upgrade GPT-4, in solving Calculus exercises. Our focus is on GPT among other possible LLM's as it appears to be the best performer in Calculus (Calonge et al. (2023)). As some degradation in time has been noted in GPT-4 abilities in solving problems (Chen et al. (2023)), we have repeated the tests six months apart.

Next, we analyze another potential use of ChatGPT that has received little attention so far but may soon be considered in many contexts: the possibility of asking ChatGPT to grade submissions such as solutions to exercises.This is largely an in-context learning task (Garg et al. (2022)), as we assign both the text of the exercise and the solution and ask for an evaluation of other proposed solutions.

Several reasons exist for testing ChatGPT on grading. One reason is the investigation of a less demanding in-context learning task; another is the possibility of envisioning some profitable integration of ChatGPT in the education process: while we consider Calculus here, grading activity can be tested and extended to other areas; another reason is that grading has not been considered yet as a potential activity for LLMs, but it could become widespread as soon as the potential is realized and input becomes more versatile. Finally, we aim to develop nonstandard AI tests (Bubeck et al. (2023)) that have not been considered during training or preliminary testing by AI developers (OpenAI (2023)). One of the results is that GPT-4 tends to exhibit loss of coherence within a single answer when faced with a long list of numerical inputs.

After the first draft of this work several other researches have investigated the potentials of LLM's in grading students' submissions: Nilsson and Tuvstedt (2023), for instance, considers programming tests, Mizumoto and Eguchi (2023); Fiacco et al. (2023) analyze the ability to evaluate linguistic papers.

No work has yet investigated the possibility to grade Calculus solutions, which is the main focus of the present paper.

## Background

### Testing LLM's

Since their recent growth in abilities and popularity LLMs and, in particular, ChatGPT have been subjected to numerous evaluations across several domains (Tamkin et al. (2021); Chen et al. (2021); Ouyang et al. (2022); Bubeck et al. (2023); Frieder et al. (2024); Ibrahim et al. (2023); Binz and Schulz (2023)). Several competing models are now available for free or with a moderate fee, and their relative performances have been evaluated in various forms. While in general which LLM performs better depends on the context (Dao (2023)), it appears that in Math ChatGPT and even more GPT-4 are currently the best performers in many respects (Calonge et al. (2023)). Potentialities and challenges of GPT in education have been reviewed by several studies (Kasneci et al. (2023)).

### GPT and Math

In Mathematics, ChatGPT based on GPT-3.5 appears to grasp the problem at hand, retrieve the appropriate information, and devise a suitable solution strategy. However, GPT-3.5 often stumbles when it comes to even basic calculations, as documented in previous studies (Frieder et al. (2024); Ibrahim et al. (2023)). As a result, the performance of GPT-3.5 as a math problem solver rarely approaches that of human solvers, including students.

LLMs fare better on math problems that are relatively simpler. For example, Minerva, another LLM, achieved a 78.5% success rate on the GSM8k problem dataset (Lewkowycz et al. (2022); Dao and Le (2023)) whereas ChatGPT's performance on datasets such as pre-algebra, pre-calculus, and basic topology is around or above 70%, see Frieder et al. (2024). However, GPT-3.5's abilities falter significantly when computations are involved (Frieder et al. (2024)): this pattern is identified in Frieder et al. (2024) via the frequency of error code e4 (calculation errors) in datasets at the pre-calculus, integration, and other similar levels. Although a comprehensive analysis of mathematically related features is included in Frieder et al. (2024), Calculus is not included, nor is the exploration of the possibility of asking ChatGPT to assess the correctness of a solution.

ChatGPT's performance can be moderately improved for datasets simpler than Calculus through prompt engineering or verification training (Liu et al. (2023); Frieder et al. (2024); Lewkowycz et al. (2022)). However, for more advanced problems than Calculus, the performance of GPT-3.5 drops significantly, achieving only between 40 and 60% of the scores, see Frieder et al. (2024).

Integrating various expertise is a promising idea that may enhance ChatGPT's performance. One approach is to combine LLMs such as ChatGPT with formally structured computing systems, such as Coq (Chlipala (2022)) or Isabelle (Wenzel et al. (2008)) for logic, and WolframAlpha (Dimiceli et al. (2010); Davis and Aaronson (2023)) for computations. Some existing integrated systems include pal-math (Gao et al. (2022)) and ChatGPT-LangChain (Chat-GPT-LangChain), which allow the chat to access tools like WolframAlpha or pal-math to improve its problem-solving abilities.

The potential use of LLMs, such as ChatGPT, in violation of academic integrity has raised the issue of detectability of the use of a language model to generate test submissions. In this direction, classifiers specifically designed to detect such forgeries in other fields have shown limited effectiveness. For example, two of the most popular classifiers, including one developed by OpenAI itself, have a false negative rate ranging from 31 to 49%, and can be obfuscated to make detection nearly impossible, with detection rates as low as 6%, see Ibrahim et al. (2023).

Another emerging topic of interest is the use of AI for tutoring and personalized learning. LLMs such as ChatGPT have shown potential in providing customized assistance to learners, tailoring their responses to individual students' needs and levels of understanding (Bhutoria (2022)). Research in this area focuses on evaluating the effectiveness of AI-driven tutoring and exploring potential improvements by combining LLMs with other AI technologies or human expertise (Johnson (2023); Crompton and Burke (2023)).

## GPT for Automated Scoring

Automated essays scoring has a long history (Gao et al. (2023)) and has been more extensively explored recently, showing that GPT has a certain level of accuracy and reliability (Mizumoto and Eguchi (2023); Fiacco et al. (2023)). It has been concluded (Mizumoto and Eguchi (2023)) that GPT can be effectively utilized as automated essays scoring tools, potentially revolutionizing methods of writing evaluation and feedback in both research and practice in the areas of humanities. A wider study of simple question–answer pairs (Schneider et al. (2023)) including multiple languages and diverse fields using a specific hyperparameter-tuned model finds an accuracy of about 86.5%. This is close to being acceptable for unsupervised automated scoring.

In computer science, the grading abilities of GPT-4 have been investigated on exercises from the foundations of programming first year courses of the Computer Science: Nilsson and Tuvstedt (2023) finds accuracy and precision ranging from about 55% to 100% depending on the observed year. As an effect of such variability the overall performance of GPT at grading in Nilsson and Tuvstedt (2023) is not satisfactory enough to deploy as a grading tool on its own today; GPT it could plausibly be used as part of a workflow together with unit tests and manual TA inspections.

In the related area of scientific paper reviewing, the idea of automated review has been thoroughly examined in Yuan et al. (2022). The authors carefully investigate the potential criteria for evaluating reviews and conclude that while natural language processing models, which are the basis for LLMs, are useful for general aspects, they lack attention to detail.

Overall, the potential use of GPT for automated grading of specific Math tests has not been explored in details so far. An analysis of various AI tools, including neural networks and large language models pre-GPT, for automated grading has been carried out in Erickson et al. (2020) for high school math problems in the ASSISTments platform; other statistical methods for high school math problems have been developed in Baral et al. (2021), Baral et al. (2022), Baral et al. (2023), Botelho et al. (2023), Zhang et al. (2022). With ad hoc methods developed in close consideration of the specific task of instructor's grade prediction, these works achieve around 80% prediction, based on various metrics, of the grade assigned by instructors. While this is a remarkable achievement these works do not directly address the possibility to use GPT nor discuss calculus level problems.

## Some Limitations of GPT

Several limitations have been noted on ChatGPT and GPT-4. The two below are relevant to our discussion.

Despite not having a persistent memory of individual prompts, these AI models use the context window to track the current conversation, akin to a"short-term memory." This allows the AI to recall previous responses and prompts within the token limit, aiding in the generation of contextually relevant and coherent responses. The ability to reference prior tokens helps counter some of the context loss by forming responses based on learned patterns, grammatical rules, vocabulary, and simulated common sense reasoning (Jacob (2023)). Token limitation can however lead to lack of coherence in long sentences or exchanges.

In addition, a recent work has observed that the performance of GPT-4 seems to have been degrading over time, in particular in Math, while GPT3.5 seems to have improve its performance (Chen et al. (2023)).

## The Present Study

The current research landscape suggests that while LLMs such as ChatGPT have made significant progress in various domains, they still face challenges in solving complex mathematical problems: we aim to extend this analysis to problems from Calculus which, in spite of this being one of the most attended math courses, have not been considered in details so far.

The second research question involves the ability of ChatGPT or GPT-4 to automatically grade calculus submissions in comparison to that of human expert graders.

In addition, we aim to follow-up on a series of issues which have surfaced in other studies or while investigating the previous questions. In particular, we want to highlight a phenomenon which has become apparent during this study, and can be identified as a type of loss of coherence.

Finally, we test whether the claimed degradation over time of GPT can be traced in the tasks that we are analyzing.

We hope hereby to contribute to the ongoing discussion about the potential applications and limitations of LLMs in education.

## Methods

We utilized the publicly available ChatGPT provided by OpenAI at https://chat.openai.com/

based on GPT-3.5 and the subscription version GPT-4. To simulate the simplest user experience, we used the GPT playground with default temperature, usually set at 0.7.

### Exercises and Solutions by ChatGPT

We have collected Calculus problems from three sources:

1. standard calculus exercises from Stewart (2020) whose solution consists only of a numerical answer; sulutions can be found in the appendix of Stewart (2020);
2. high school competition problems from the 2021 SMT Team Round Stanford Math Tournament from AoPSOnline (2022), with solutions published in the website;
3. exercises from a real Calculus course held at New York University Abu Dhabi in the Summer of 2017; problems were chosen by randomly selecting 10 questions from homework and final exams; the questions were selected based on the following criteria: (i) the questions were limited to two paragraphs, although the answers could be of any length; (ii) multiple-choice questions were not allowed; (iii) no images, sketches, or diagrams were allowed in either the question or the answer; (iv) no attachments were allowed in either the question or the answer. Solutions are provided by the instructor of the course.

The Stanford problems are published in TeX format, while all other problems are typewritten in Math format. To prepare for submission to ChatGPT the other problems have been rewritten in a Math plain-text format available in Keely (2009); notice that this format is essentially the only one accepted by Wolfram-Alpha in its Natural-Language mode. The format is ambiguous for large formulas, but consistent for the calculus exercises we are considering when enough brackets are used, see Keely (2009).

All exercises are reproduced in the online material, in the format which has been presented to ChatGPT, including the exact prompt, and can be directly resubmitted. All exercises have been submitted to ChatGPT-3.5 and GPT-4, asking for a solution; for the real calculus exercises each problem had been submitted three times in a double-blind previous experiment (Ibrahim et al. (2023)); the solutions returned by the chat are included in the online material. If submitted again, the answers will be different due to the internal randomness of ChatGPT, but even when the same exercise was submitted multiple times, the overall results seemed to be consistent.

This data is the basis of the evaluation of calculus solving capabilities of ChatGPT.

### Submissions

Next, in order to prepare a submission to be graded, all exercises have been submitted to both GPT-3.5 and GPT-4 several times, asking for solutions. All answers are recorded in the online material.

In addition, for the exercises from Stewart and Stanford, we created plausible synthetic answers which, starting from one correct solution, included one correct answer equal to the correct solution, one correct answer in a form different from the correct solution, one partially correct answer, and one completely incorrect, in random order.

For the exercises from the actual calculus class, answers by actual students of the course, selected among those who have given consent so as to represent a variety of overall class grades, were collected in Ibrahim et al. (2023); such data collection is blind on the side of students, who clearly were unaware of the current test at the time of solving the exercises; the answers by the students have been converted in digital text format. Students' consent for the usage of their anonymized submissions in the study has been obtained; and students' confirmation that they were 18 years of age or older has been obtained.

### Grading Rubrics

Submissions to be graded were prepared by stating again each problem, and the relative correct solution.

Next, grading rubrics have been established for the study. Due to the extensive experience in grading calculus papers, the development of the grading system did not require complicated considerations (see Yuan et al. (2022) for challenges in establishing standards for paper reviewing).

The rubric for problems from Stewart (2020) has been chosen to be the simplest one: Exact solution: 10 pts. Incorrect solution: 0 pts.

For the other problems, the rubric follows traditional academic standards, with partial credit awarded for meaningful descriptions of procedures that would lead to the solution even if the calculations are incorrect. The rubric assigns only 8 points out of 10 for correct solutions without explanation. Since ChatGPT provides more explanations than the students, the rubric was designed to favor the chatbot. The grading rubric is as follows:

- Correct calculations and final result with correct explanation: 10 pts.
- Correct calculations and final result without explanation: 8 pts.
- Incorrect calculations, but with some sensible explanation: 2 pts.
- Incorrect calculations, but with a detailed and correct explanation: 4 pts.
- Almost everything correct, but with only a minor calculation error: 8 pts.
- Blank or no sensible explanation and incorrect results: 0 pts.

### Prompts

The grading activity requires a more elaborate prompt than problem-solving, as we provide the text of the exercise, the solution, the grading rubric suggesting partial credit, and a certain number of submissions to be graded. In addition, prompt engineering can significantly help in obtaining more appropriate answers, see Liu et al. (2023).

We also need to carefully consider the format, as the solutions are supposed to be produced by Calculus students but accessible to ChatGPT; we have opted for two possibilities: LaTeX and an intuitive Math plain-text format. Neither of these is very practical, as it would require extra effort from students and would have to be done on a computer, exposing the possibility of improperly accessing ChatGPT itself to get a solution. However, the input process will likely evolve rapidly, especially if handwriting recognition is systematically incorporated, so we consider this a rather technical but temporary issue.

### Grading Submissions

For each proposed solution to an exercise, the author of this research has provided a golden standard of scores in accordance with the rubric for each exercise. In about half of the exercises, when ample leeway was present, the golden standard has been provided as an interval of possible scores rather then a single value.

For the Stewart and Stanford problems, a prompt was prepared with the text of the exercise, the rubric, the correct solution and the synthetic solutions and this was submitted to ChatGPT and GPT-4 for grading.

For the real calculus problems, the answers provided by the chat, and the solutions of the students were labeled Submission 1, 2, … and for each problem, a prompt was prepared with the text of the exercise, a correct solution, the corresponding grading rubric, and the labeled submissions. Some prompt engineering was added, such as reminding that $-x$ is not correct when x is a solution, and the prompt was submitted to ChatGPT three times, and to GPT-4 for grading.

In addition, for the problems from the actual calculus class, we recruited three graders who were senior students that had previously taken the course and obtained an A grade (Ibrahim et al. (2023)). The graders were provided with the same material as the chat, and were asked to grade the submissions The test was conducted in a controlled environment to ensure that the human graders were unaware that some of the responses had been machine generated. Each grader was compensated at a rate of $15 per hour for their time. Participants were given debrief information after the research task was complete, which explained the full purpose of the research.

All the material submitted to chat and graders and their scores, redacted to remove handwriting, are available in the supplementary material.

Appendix A shows two examples of the prompts that we submitted to the chat and the corresponding scores provided by ChatGPT.

## Testing Loss of Coherence and Hallucinations

Several tests were conducted after noticing that GPT seemed to become less reliable along a test or when many analogous tasks were required. In particular, we focused on recognizing whether sequences of numbers contained some in a given interval. The task is easy for humans, but also time consuming and repetitive; it is also one of the basic tasks involved in grading. We report of the most significant tests.

In one test we generated 100 sequences of 10 random integers each picked uniformly in [1, 100]. We used the default random generator in Mathematica. GPT-4 was asked to identify in which sequences there appear numbers in the interval [16,24]. The test was repeated many times, each time with independent random numbers, and number of incorrect assignments were averaged. We report one typical list of the positions in the sequence of 100 in which GPT-4 made an incorrect identification.

In the second test, we similarly generated 100 sequences of 10 random integers each picked uniformly in [1,100]; and GPT-4 was asked to identify in which sequences there appear numbers in the interval [32,34]. The test was repeated 10 times, each time with independent random numbers using a Python random number generator, and a suitable prompt engineering was used:"Which of the following lists contains a number which is greater than or equal to 32 and less than or equal to 34? Check each list manually, do not use plugins. Make a list of the list numbers, but only the numbers. Write the answer as a list of the list numbers like [2,5,…]." The number of incorrect answers were averaged for each position in the 100 sequences. Many other similar tests were performed with different intervals, all with the approximately the same outcomes.

In the final test, the chat was asked to identify in which lists there appear the number 620 within the sequences of the previous test. In this final experiment, exploiting the interactive capabilities of the chatbots, we replied to the chat alerting of mistakes if its reply contained some, or even if it did not contain any to test possible hallucination.

### Testing Degrading of ChatGPT and GPT-4 and Potential Alternatives

To test the apparent degrading of ChatGPT3.5 and GPT-4 in the realm of Calculus problems, the solutions to all problems tested in March 2023 have been resubmitted in September for solution.

We also tested ChatGPT-LangChain (Chat-GPT-LangChain), which has the ability to leverage WolframAlpha for specific calculations by submitting all of the calculus problems in the same format.

### Statistical Analysis

For the performance in solving calculus problems average percentage scores have been computed and p-values of two-sample t-test are reported to compare abilities as problems solvers of ChatGPT3.5 and GPT-4 in March and September, and the students.

For the test of potential degrading of the chatbots, besides the p-values of two-sample t-test we have computed the correlations between the scores received in the various exercises.

For the evaluation of the chatbots as graders, four statistics have been computed by comparison with the golden standard determined by the author.

1. The average score assigned to each submission: this is relevant both since it is the only figure contributing to students' records and since it averages out possible discrepancies in opposite directions.
2. Scores deviating from the golden standard have been classified as"close to range" if the distance is less than or equal to 2 points (out of 10) or"completely off range" for larger deviations. The fraction of completely off range scores has been determined for each grader, representing the fraction of cases in which the grader has completely misunderstood the content of the submission.
3. For a finer evaluation of the grading activity, a grader's aptness index in the form of mean quadratic deviation from the golden standard has been computed. This value ranges potentially from 0 to 10, with 0 representing perfect scoring and larger values indicating some inefficiency. Grader's aptness indices have been compared for all pairs of graders to compare relative aptness, and p-values of two-sample t-tests have been reported.

For the evaluation of ability of the chatbots to identify sequences with integers in given intervals, the percentage error has been determined, both for the overall performance and for the performance at the tests in corresponding order of submission. For the answers which halted, the average number of items which have been checked by GPT-4 before halting has also been computed.

## Results

**Calculus problem-solving** In standard calculus exercises with definite, numerical answers from Stewart (2020), which the chatbot might have seen in its training, GPT-3.5 scored 30% ± 14.5% while GPT-4 scored 60% ± 15.5%. This is an improvement (p-value 0.078), but still below a passing grade in college. For the Calculus exercises from the 2021 SMT Team Round Stanford Math Tournament, GPT-3.5 scored 14% ± 5.2%, only from partial credit with no correct answers, while GPT-4 scored 24% ± 7.8%, with one correct answer in 10 exercises. In the selection of homework and in-class test exercises from the 2021 calculus class, GPT-3.5 scored 43% ± 3.9%, in a double-blind experiment which was part of a previous study (Ibrahim et al. (2023)), while GPT-4 scored 72% ± 11.6%: again, GPT-4 has shown a marked improvement, even if not statistically significant, due to large variance in the samples (p-value 0.25), and the GPT- 4 score is comparable to that of a sample of real students from the class. In the double-blind experiment, the human students received on average 70.88% ± 4.7% when graded by human graders who did not know about the authors of the submissions, see Fig. 1.

Notably, ChatGPT, even GPT-4, made many elementary errors (see Cobbe et al. (2021)), such as claiming that $3(-1)^4 - 4(-1)^3 - 12(-1)^2 + 1 = -8$ while the correct value is $-4$ (see Question 4.1.55 from Stewart (2020) in the Supplementary Material). Since calculus problems often involve several simple calculations, ChatGPT frequently produced incorrect intermediate results, leading to incorrect final answers. Nonetheless, ChatGPT's justifications for its reasoning were generally appropriate.

Our findings are consistent with related studies on ChatGPT-3.5 performance in general math (Frieder et al. (2024); Ibrahim et al. (2023)). The ability of GPT-4 in Calculus has considerably improved see OpenAI (2023), but, even with a lenient grading policy giving partial credit for explanations without correct results and point deductions for lack of explanations, its scores barely approach sufficiency for standard calculus problems, and the chat offers very little reliability even for simple calculations.

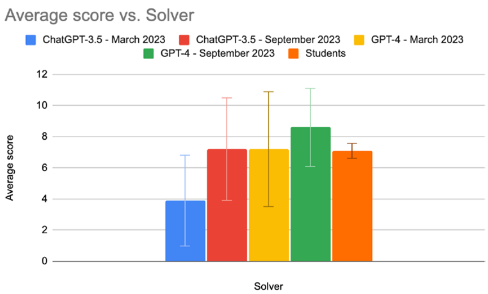



**Fig. 1** Average scores of ChatGPT and GPT-4 submissions and students' submissions for standard calculus problems. Except for ChatGPT in March of 2023, the performance of the chatbots is comparable to that of the students

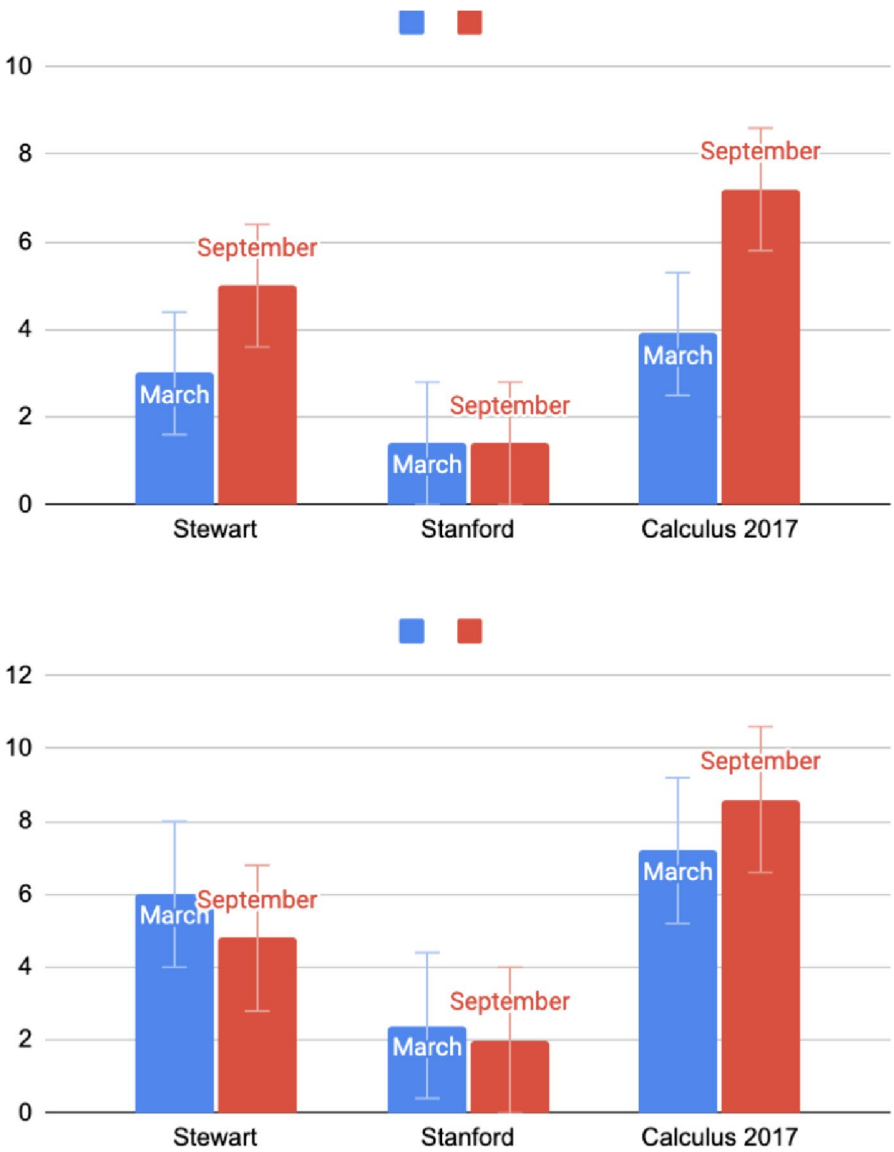



**Fig. 2** Changes in average scores of ChatGPT3.5 (above) and GPT-4 (below) in solving exercises between March and September 2023 for the indicated sets of exercises

The test about over time changes or degrading has shown that indeed ChatGPT3.5 seems to have improved its mathematical abilities, especially in the exercises from the calculus class (p-value 1.3%), while in all other cases, including the performance of GPT-4, there has been no significant change on average, see Fig. 2. The fluctuation has been substantial, with correlations ranging from 0.19

**Table 1** Percentage errors

| *Answers to be graded* | GPT-3.5 close to range | GPT-3.5 off range | GPT-4 close to range | GPT-4 off range |
|---|---|---|---|---|
| *Synthetic answers / Steward* | | $10 \pm 0.49$ | | $6.67 \pm 0.42$ |
| *Synthetic answers / Stanford* | 32.5 | 5 | 30 | 0 |

**Table 2** Average scores by solver and grader in percentage of total score Calculus class exercises

| | **Calculus class exercises** | | | | **Stewart** | **Stanford** |
|---|---|---|---|---|---|---|
| *Solver* | GPT-3.5 | GPT-4 | Human grader | Author | Author | Author |
| *GPT-3.5* | $70 \pm 4.1$ | $46.7 \pm 6.1$ | $46.7 \pm 6.1$ | | $30 \pm 14$ | $14 \pm 5$ |
| *GPT-4* | | | | $72 \pm 11.6$ | $60 \pm 15$ | $24 \pm 7.8$ |
| *Students* | $75.6 \pm 4.4$ | $70 \pm 7$ | $70.9 \pm 4.7$ | | | |

to 0.69 for GPT-4 and from 0.21 to 0.82 for ChatGPT-3.5. This testifies of the intrinsic randomness with the default temperature, and of some over time fluctuations of the quality of the outcomes. However, we did not identify any relevant change in GPT-4, while ChatGPT-3.5 has shown a detectable improvement.

Our findings with ChatGPT-LangChain were not encouraging. In almost all of the Calculus problems in our study, ChatGPT either did not utilize WolframAlpha or, when it did, the prompts were often too linguistically complex for WolframAlpha to comprehend. In one instance, for the prompt $\int_{-1}^{1} \frac{\sin x}{x^4+x^2+2} dx$, ChatGPT erroneously reported that WolframAlpha returned a non-zero value of $-\cos(1)/4 + \cos(-1)/2 - 1$ og(2). Going directly to WolframAlpha and reworking the prompt multiple times, WolframAlpha was finally able to understand the input and correctly returned a value of 0 (easily obtainable by symmetry) and a complex integral expression. We did not conduct further testing with ChatGPT-LangChain. While integration with Wolfram Alpha holds great promise, the chatbot's ability to rephrase queries in a way that WolframAlpha can use remains a key challenge (a more thorough investigation with similar conclusions is in Davis and Aaronson (2023)).

**Grading Calculus Submissions** For the Stewart's problems, ChatGPT had an error rate of $10 \pm 0.49\%$ and GPT-4 $6.67 \pm 0.42\%$, a significant difference (p-value $< 10^{-6}$).

When asked to grade solutions to the Stanford competition problems, GPT-3.5 graded about 62.5% of the submissions correctly, with close to range scores in about 32.5% of cases, and completely completely off range scores in around 5% of the cases. GPT-4 performed similarly but with no completely off range scores. See Table 1.

When asked to score real students submissions in a Calculus class, GPT-3.5 assigned an average score of 75.6% and GPT-4 assigned an average of 70.0% while the human graders assigned an average of 70.9%, see Table 2 and Figs. 3 and 4. As

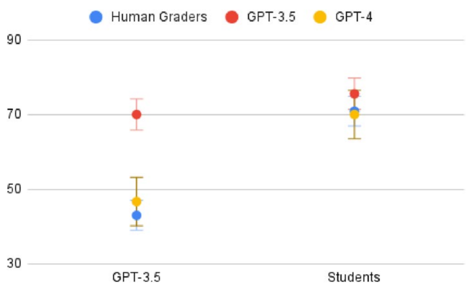



**Fig. 3** Average scores of ChatGPT submissions and students' submissions and error bars, as graded by human graders (blue), the GPT-3.5 (red), and GPT-4 (yellow) using the provided rubric. Human graders are more accurate compared to golden standard, based on their aptness index in Table 3, and correctly assign a higher passing grade to students, and a lower, and non passing, grade to the chat. GPT-3.5 is rather incorrect, while GPT-4 reproduces the averages quite accurately

an example, the average scores for Student D where 70, 70, 72 from GPT-3.5, 56 from GPT-4 and 58,62,58 from the three human graders. Regarding incorrect scoring, GPT-3.5 had an average of 10 scores incorrect but close to ranges, with about 3 completely off range, while GPT-4 has 10 close to range, but 1 off range; for comparison, human graders have about 1 close to range, but none off range.

ChatGPT-3.5 failed severely when grading its own work, altogether GPT-3.5 has an average of 7 extra incorrect scores close to the expected ranges, with about 10 extra completely off range, while GPT-4 has 6 more close to range, and 5 off range. Human graders also got confused, to a lesser degree, with about 5 extra close to range, and 1 off range.

**Table 3** Table of Grader's Aptness Indices. HG stands for Human Grader

| Aptness index | GPT-4 | GPT3.5–1 | GPT3.5–2 | GPT3.5–3 | HG 1 | HG 2 | HG 3 |
|---|---|---|---|---|---|---|---|
| AVE | 0.867 | 1.800 | 1.267 | 1.800 | 0.000 | 0.067 | 0.067 |
| SD | 1.358 | 2.578 | 2.309 | 2.578 | 0.365 | 0.365 | 0.507 |
| SE | 0.248 | 0.471 | 0.422 | 0.471 | 0.067 | 0.067 | 0.093 |

**Table 4** p-value two-sample t-test for pairwise comparison of scores using Two-sample t-test. HG stands for Human Grader

| p-value | GPT-4 | GPT3.5–1 | GPT3.5–2 | GPT3.5–3 | HG 1 | HG 2 | HG 3 |
|---|---|---|---|---|---|---|---|
| GPT-4 | | 0.0449 | 0.1223 | 0.0449 | 0.0008 | 0.0017 | 0.0022 |
| GPT3.5–1 | | | 0.2050 | 0.5000 | 0.0002 | 0.0003 | 0.0004 |
| GPT3.5–2 | | | | 0.2050 | 0.0025 | 0.0038 | 0.0042 |
| GPT3.5–2 | | | | | 0.0002 | 0.0003 | 0.0004 |
| HG 1 | | | | | | 0.2448 | 0.2840 |
| HG 2 | | | | | | | 0.5000 |

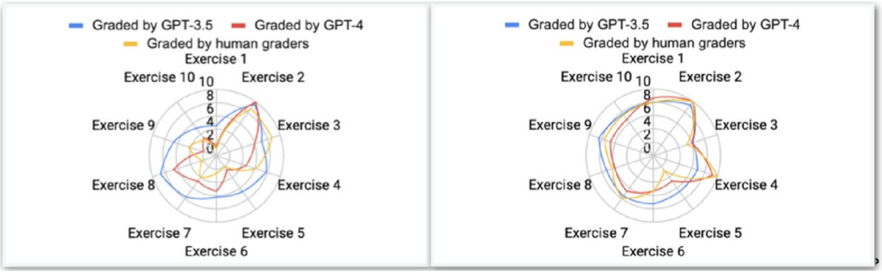



**Fig. 4** Scores assigned to single exercises, averaged over the number of grading iterations when there are more then one. The right table corresponds to grading students' submission; while GPT-3.5 scores diverge in several of the exercises, the once assigned by GPT-4 are very close to those of the human graders. The left table is for GPT-3.5 submissions; there is less agreement, likely due to the length of the answers, and the related difficulty in assigning partial score; while GPT-3.5 as grader is unreasonably lenient, GPT-4 and the human graders end up having about the same overall average

The grader's aptness index based on the grading of real students' submissions gave the values reported in Table 3. The human grader all have indices which are statistically indistinguishable from 0, reflecting their preparation. GPT-4 has an index of $0.867 \pm 0.248$, significantly different from zero but better than ChatGPT which has an index substantially greater than 1. p-values for two-sample t-test for pairwise comparison of index values are reported in Table 4. It can be seen that human graders have a very high degree of homogeneity, that GPT-4 performs better than most GPT3.5, but worse than the human graders, and that GPT3.5 has also a high degree of consistency, see also Fig. 4.

In summary, some of the grading outputs are indeed impressive: ChatGPT quickly returns appropriate scores for each submission, often providing motivation interpreted in terms of the grading rubric. Other times, the output contains minor misinterpretations of the submissions. There are, however, blunders and other issues which, as we discuss in detail, make the scoring scarcely reliable. The appropriateness of the scores varies depending on the input format, the complexity, or the length of the exercise or the solutions.

GPT-4 has improved upon the previous version and generates scoring of Calculus exercises which are generally appropriate, although with some errors and misunderstanding of equivalent solutions. It is remarkable that the average scores from GPT-4 are au-pair with those of humans, having been produced in a few seconds each, instead of the hours needed by human graders. On the other hand, at a more accurate aptness investigation, GPT-4 shows the persistence of (mostly moderate) inaccuracies. In a sense, the grading errors average out when computing the overall score.

With all of this taken into account, automated grading of calculus exercises by GPT-4 alone is not feasible at the moment, but it can certainly be of great support to a human grader, and blending scoring has enormous potentials.

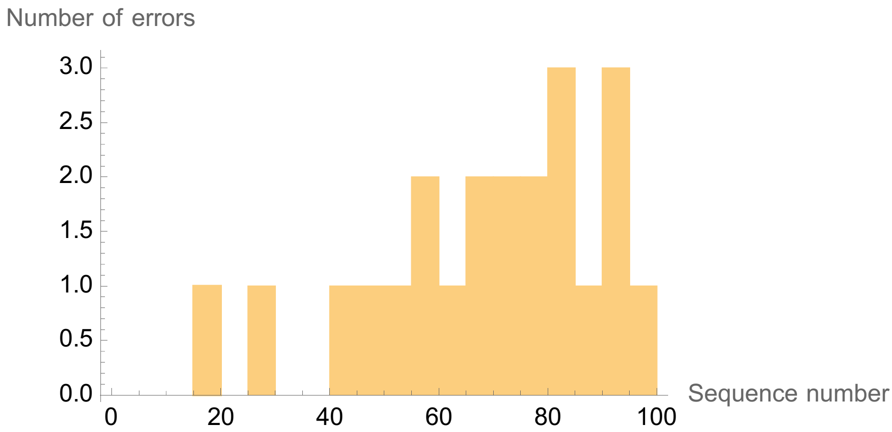
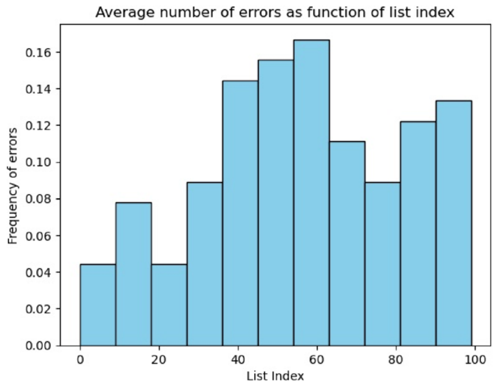



**Fig. 5** Error rate by iteration in identifying numbers in [16,24] among 100 sequences of 10 random numbers uniformly chosen in {0,100} each, in a typical experiment. Each cell has size 5, which indicates that error rate is about 50% for large sequence numbers

**Fig. 6** Average error rate by list index in identifying numbers in [32–34] among 100 sequences of 10 random numbers uniformly chosen in {1,100} each, in a typical experiment




**Loss of coherence and forced hallucination** Already in the simple task of identifying which lists contain integer numbers in a certain range GPT-4 makes a substantial number of mistakes, although this is improving.

In a first experiment with 50 sequence out of the 100, the fraction of errors was around 30% (28.6% ± 3.34%) in June 2023, but is it now down to 6.15% ± 2.18%. This is still below human level, for which recognition in adult age is considered essentially error free (Rhodes et al. (2019)). The error rate is even greater with 100 sequences: only the first sequences are analyzed by GPT-4, with an observed error rate which was 55.3 ± 6.41% and now is still 32% ± 3.55%. After a certain number of sequences, GPT-4 stops producing outcomes: the mean number of sequences analyzed before the output stops was 59.4 ± 0.5 out of 100 input sequences, and has now improved to 78 ± 3.64. Despite the improvements, the error rates remain very high.

In addition, we noticed a peculiar pattern of the errors. While the initial sequences are quite carefully analyzed and errors are rare, the performance degrades after the first close-by errors, and if the list is too long it stops altogether.

Figure 5 illustrates how in a typical experiment very few errors occur in analyzing the first sequences, while the identification degenerates in the longer run. Notice that since each cell has size 5, from about Sequence 50 the outcome is very close to random guessing. Figure 6 reports the average error rate in 10 experiments: one can notice that the error rate is around 5% among the 30 or so lists, and then moves up to 10–15% in the later experiments. This behavior seems to be quite consistent over many experiments.

Apparently, the identification of a shorter interval is "easier", so the error rate remains around 15%, while for a longer interval like [16–24] it goes up to 50%, corresponding to an essentially random guess.

We interpret this phenomenon as an instance of the loss of coherence which has been noted both in dialogue systems and code prediction tasks (Liu et al. (2020)): transformer models encounter specific challenges in maintaining context over extended sequences, and have difficulty referring to earlier parts of a conversation or when code is long and complex. Our results highlight that similar mechanisms take place also in Mathematical context, even for very simple tasks: while these are carried out well at the beginning, and few errors appear occasionally, when the errors pile up too much GPT is much less able to focus on the correct inputs and its output becomes much more random.

The mechanism of loss of coherence seems analogous to the inability to self-correct which has been noted in some researches: as without external ground truth LLM's are essentially unable to correct themselves (Huang et al. (2023)), it appears that the incorrect classifications are then taken as possible truths. In one case only GPT has been able to correct one mistake, indicating that one list number should be included, but then immediately commenting that that was a mistake.

These observations can lead to benchmark experiments for the attempts to improve upon the loss of coherence by modifications of the attention mechanism which is at the base of the transformer architecture (Ghaith (2024)). Checking the correctness of potential results of exercises when compared to a provided solution remains the aim of automated grading, but we identified here a more elementary process which can be used as initial benchmark.

We briefly attempted to use plugins to exploit the rigor of programs like WolframAlpha in the identification of sequences with numbers in a range, but, consistently with the findings in Davis and Aaronson (2023), GPT-4 does not seem to be able to input the correct instructions in the plugins and all of the attempts have crashed before generating any output.

We also forced hallucination. In the experiment asking to indicate which rows contain the number 620 when 100 rows each with 10 random integers are provided and the rows containing the number 620 are Row 17 and Row 98, GPT-4 correctly identifies only the first row, and loses coherence about the second. When prompted it finds the second one, completing the task. But then, if prompted again, it finds a non existent Row 110, even listing its supposed 10 numbers, all different from 620.

**Prompt:** Which rows contain the number 620. Check carefully one row at a time. [100 rows are provided].

**GPT-4:** I have found the number 620 in the following row: 621 34 440 789 499 620 919 839 425 23 [this is Row 17, one of the two containing the requred number].

**Prompt**: Are you sure?

**GPT-4:** Apologies for the confusion earlier. Upon re-examination, I have found the number 620 in the following row: 29 599 395 752 490 970 184 620 649 65 [this is Row 98, the other containing the requred number].

**Prompt:** It is also in another row, no?

**GPT-4:** You are correct. I apologize for the oversight. The number 620 appears in three rows:

Row 17: 621 34 440 789 499 620 919 839 425 23.

Row 98: 29 599 395 752 490 970 184 620 649 65.

Row 110: 892 150 522 605 225 947 660 440 112 742.

Thank you for pointing that out, and I apologize for any confusion earlier.

Loss of coherence and hallucinations are ubiquitous in GPT-4 output, often in tiny details which can be hard to detect. In Appendix B, we report the scoring of a competition level exercise by GPT-4: the scoring is essentially correct, but hallucinates in one detail. Later on, there is another peculiar mistake: GPT-4 seems to accept as correct one result which it had previously graded as incorrect; somehow, the mere presence of the incorrect result becomes a supporting element for a later statement. We consider this an instance of loss of coherence, in the sense that GPT-4 has lost track of what can be used and what not (see end of Appendix B).

## Discussion

### Summary

We have confirmed that, as expected from tests on other math problems, ChatGPT-3.5 in March 2023 was not capable of providing satisfactory answers to calculus problems. Both GPT-4 and ChatGPT-3.5 in September 2023 show substantial improvement, but can barely reach a passing score on standard problems and are not capable of handling competition-level questions. Although identification and description of mathematical tools are appropriate, there remain many arithmetical mistakes. Instructors should be aware that, at this stage, calculus students could use the chat to set up the solution, but would have to check every single calculation. In the end, this could be rather beneficial to the learning process.

A possibility insufficiently explored is that of asking ChatGPT to grade calculus assignments. At the moment, this is limited by the requirement to submit queries to ChatGPT in text form, necessitating submissions in an intuitive but unnatural math textual form. If image submission becomes possible, grading by ChatGPT might rapidly become a widespread option. Our tests show that while ChatGPT-3.5 used to make more mistakes, the more recent chat GPT's or GPT-4's grading are rather close to that of a human grader. It could be easily implemented, with appropriate prompt engineering (Lewkowycz et al. (2022). Errors are mostly due to arithmetical mistakes. In addition, GPT seems to be able to appropriately follow a grading rubric, including partial credit when appropriate. Considering the complementarity of the specific abilities, merging human and GPT grading has the potentialities of sharply improving grading.

Exploring the possibility of asking ChatGPT to grade calculus assignments also constitutes a nonstandard test in an operation, the grading, which the chat is unlikely to have seen during training and for which it has not been specifically trained. The fact that GPT-4"comprehends" the request in all its details and can perform it rather satisfactorily can definitely be considered a trace of intelligence (Bubeck et al. (2023)).

Although more testing is needed to fully evaluate the grading capabilities of GPT, the results of this study suggest that the chatbot has the potential to be used as a starting point for grading students' submissions of solutions to calculus exercises. This possibility presents both opportunities and issues. The examples in Appendix A and B offer a glance on the aptness as well as the issues which appear when GPT is asked to grade calculus submissions.

Lastly, we have witnessed several instances of a phenomenon which we called"loss of coherence" in long tasks in the mathematical setting: when prompted to repeat many analogous tasks with a clearly defined answer, it appears that GPT-4 tends to loose focus after a while. While this requires further investigation, our findings have brought to light the necessity to be alert about this possibility. The loss of coherence, or inability to stay focused in psychological terms, is likely due to the attention mechanisms that govern transformers like GPT-4 (Vaswani et al. (2017)): in repetitive tasks it seems as if the focus on the prompt slowly vanishes. We also confirmed the presence of hallucination. Studying and containing loss of coherence mechanisms and hallucination is likely to be of great importance in improving the performance of LLMs.

In the tests about identifying integers in lists, GPT-4 recognized which, among 100 lists of ten integers each, contains numbers in a given interval with an error rate of about 35%, see Fig. 5.

While the size of this error even in the simple task of identifying relations between integers explains the unreliability of the output of the solutions to calculus exercises, the experiment also highlights a possible cause. As GPT models employ a unique kind of attention mechanism referred to as transformer attention, characterized by its self-attention layers (Vig (2019)), it is the appearance of random mistakes with a certain density which derails the ability of GPT to continue an appropriate detection. Figure 5 illustrates the location of the errors in a typical experiment: while the first outputs are correct, when errors begin to become more frequent the detection becomes essentially random. This is another form of loss of coherence: the first mistakes prompt an overwhelming inability to stay focused on the original task.

## Artificial Intelligence in Education

In the continually expanding field of Artificial Intelligence Applied to Education, the advancements and challenges posed by the recent iterations of ChatGPT offer intriguing insights into the interplay between machine learning and pedagogy. Historically, attempts to apply AI for educational purposes have seen fluctuating levels of success, contingent upon the specificity and complexity of tasks involved (Chen et al. (2020); Ilkka (2018)). Our recent findings reveal a consistent trend observed across various AI evaluations: while ChatGPT-3.5 shows limitations in delivering satisfactory answers to calculus problems, its successor, GPT-4, demonstrates discernible improvements, albeit with lingering issues. These results had been observed in other studies (Calonge et al. (2023); Dao and Le (2023); Okonkwo and Ade-Ibijola (2021); Bhutoria (2022); Binz and Schulz (2023)) and mirror the overarching trajectory of AI in education, whereby progressive iterations inch closer to human-like performances yet grapple with nuanced tasks, especially competition-level questions.

The proposition to deploy ChatGPT for grading calculus assignments is indeed novel; similar intents have been explored for other disciplines (Fiacco et al. (2023); Mizumoto and Eguchi (2023); Nilsson and Tuvstedt (2023); Ndukwe et al. (2019)). Presently, the feasibility of this idea is hindered by the textual constraints imposed on query submissions. The contemporary reliance on textual inputs can be juxtaposed against earlier endeavors in AI education, which predominantly hinged on graphical interfaces (Shabrina et al. (2023)). Transitioning to an image-based submission system could revolutionize the utility of ChatGPT in educational settings, enabling more intuitive and natural student–teacher interactions.

Interestingly, our evaluations unveiled a potential strength of GPT-4, whereby its grading, with suitable prompt engineering, has some resemblance to that of human graders. The very act of grading, which likely remains an unseen task during GPT-4's training, underscores the model's latent intelligence, aligning with recent explorations on AI's ability to"comprehend" novel requests (Srivastava et al. (2022)).

However, educators and AI researchers must remain cognizant of the emerging challenges, among which the arithmetical mistakes, the phenomena of'loss of coherence' and hallucination. This observed decrement in performance over repetitive tasks could be symptomatic of the inherent attention mechanisms in transformer models (Vaswani et al. (2017)).

In summation, as AI's footprint in education continues to grow, so do the nuances and intricacies of its application. Our exploration with ChatGPT, both its potentials and drawbacks, exemplifies this dynamic interplay. Future endeavors must remain attuned to these observations, ensuring that as AI models evolve, so does their efficacy in pedagogical settings (Stoica (2022)).

## Ethical and Regulatory Issues

The mere possibility of having papers graded by ChatGPT raises a number of ethical, pedagogical, and regulatory issues (European Commission (2023), De Winter et al. (2023)):

- Correctness: Humans are prone to errors and lack of precision (Mizumoto and Eguchi (2023)) due to a variety of causes. Our tests show that even with suitable prompt engineering, and even if the grading is stated by the chat without raising doubts, also automated systems as ChatGPT make frequent errors. One can envision that, if appropriately carried out, integration of human and automated grading could lead to an overall reduction of unreliability. The opposite effect of compounding mistakes is, however, a risk, so it is a future challenge to find proficuous way of combining the two methods (Budhwar et al. (2023)).

- Fairness: An analogous potential concern is about fairness. The output of Chat-GPT is unlikely to be biased towards any specific student especially if submissions cannot be traced to any specific individual. A potential concern regarding fairness might stem from the inability to recognize equivalent expressions, possible linked to students' prior educational experience. To reduce the likelihood of the AI failing to recognize diverse but correct expressions of the solution, integrating prompt engineering into the course could be beneficial (Tinguely et al. (2023)). It appears, however, that grading by AI in general gives a perception of greater fairness (Chai, Fangyuan et al. (2024)); a possible reason is the greater transparency in the process due to the more detailed explanations that AI can easily incorporate in the feedback (Chai, Fangyuan et al. (2024)).
- Speed: The great advantage of using systems like ChatGPT to grade submissions is that feedback can drastically decrease the time needed by a human grader to complete grading. This is an enormous benefit, as feedback can be provided when most useful, and not with a distracting delay (Kasneci et al. (2023), Fiacco et al. (2023)).
- Partial credit: One of the relevant features of grading is the ability to capture partial elements of the solution, even if the final result is not exactly correct. While automated grading is already available through dedicated websites and prearranged questions, this would be the first instance of grading by a generic tool. The human factor, which is partially built into dedicated platforms, would have to be replaced by the wisdom embedded in the training of an LLM. Our results show that the outputs of the chat appropriately reflect partial contributions, occasionally even better than the human counterpart has been able to detect.
- Submission language: To allow for the possibility of having ChatGPT providing scores, submissions must use a specific language. This is a relevant issue, as one either uses TeX, which is not known by most calculus students, or uses a math plain-text format, which is not entirely consistent and can create ambiguities or misunderstandings (Keely (2009)). This issue is likely to be solved when images will be accepted in the prompts.
- Regulations: due the ethical issues raised by the use of GPT in education, and in particular in grading, suitable regulations are necessary on the institutional side, especially with regard to data privacy and security, transparency of usage, fair access, sustainability: we refer to thorough studies for extensive discussion of these issues (Kasneci et al. (2023); Zhou et al. (2023)).

## Relation to Technological Advances

Rapid technological advances are occurring in the field of AI. Our work has specifically focused on grading Calculus problems with ChatGPT or GPT-4, adopting the perspective of an end-user. This viewpoint will inform future analyses and the development of grading processes when more sophisticated tools will become available. Additionally, we have highlighted issues such as loss of coherence and ethical considerations, which will demand closer scrutiny in the near future.

More importantly, our analysis largely presupposes the availability of technological advancements, such as the input of handwritten solutions (Tiwari et al. (2023)) and the reliable integration with plugins (Davis and Aaronson (2023).

## Limitations

Our study is limited by the use of a small sample of questions from a single textbook. Therefore, the generalizability of our results to other sources of calculus problems is unclear and further testing is necessary.

The risk of cross contamination is moderate: the problems from the actual Calculus exams were never published; the Stewart problems are available with their solutions; the Stanford problems are from the 2021 competition, which has been published in 2022 after the GPT pre-training data cuts off in September 2021.

Although we were able to reduce the occurrence of trivial mistakes from the chat by providing it with a solution and additional warnings, human supervision is still necessary to ensure accurate grading.

The Math plain-text format we adopted can be ambiguous, particularly if parentheses are not used carefully (Keely (2009)). This may have contributed to errors made by ChatGPT. Moreover, in order to implement the grading process in real-world situations, students may need to be required to submit their work in this format, which could create an additional burden for them.

In spite of having provided a useful context on how to address automated grading in Calculus, rapidly advancing technologies could make this study obsolete.

## Conclusions

We investigated the potential of using ChatGPT, specifically GPT-4, as a grader for calculus papers. A primary challenge is the submission format for mathematical expressions, an issue likely to be addressed once systematic input via images becomes commonplace.

Despite this, ChatGPT, and even more so GPT-4, are progressively able to identify the grading task and consistently apply the rubric, including awarding partial credit. Although their performance does not yet match that of human graders—primarily due to simple arithmetic mistakes, loss of coherence, and occasional hallucinations—they occasionally outperform humans in grading. Moreover, the grading is completed within a few seconds. Given these strengths, integrating ChatGPT or GPT-4 with human graders seems promising, potentially streamlining and accelerating the grading process while maintaining fairness. Besides this, the insights gained from this study have broader implications beyond calculus grading, as we outline below.

## Future directions

The swift evolution of large language models like GPT paints an optimistic trajectory for their integration into the educational landscape. The transition from ChatGPT-3.5 to GPT-4 over a brief period emphasizes the potential of rapid enhancements in forthcoming versions. The current foundational research on automated calculus grading by LLM allows to delineate the following avenues for prospective applications and explorations:

1. **Image Submission and Processing**: One significant constraint is the necessity to submit mathematical expressions in text form. However, with ongoing advancements in artificial intelligence, we foresee a future where image-based submissions for mathematical expressions become attainable (Zhang et al. (2023)). Such a paradigm shift will drastically enhance the user experience, closely emulating human grading systems.
2. **Integration with Mathematical Software**: Merging the capabilities of GPT-like models with established mathematical software such as Mathematica offers a promising direction. Such an amalgamation could harness the computational accuracy of software with the intuitive problem-solving approaches of LLMs, thus reducing arithmetic and foundational errors in solutions and grading processes (Davis and Aaronson (2023)).
3. **Hybrid Grading Systems**: The synergy between human graders and LLMs like GPT-4 offers an exciting frontier. Subsequent research might focus on establishing a symbiotic relationship between these entities, optimizing their individual strengths and neutralizing their inherent limitations. There seems to be no studies on how this interaction could be shaped.
4. **Enhanced Attention Mechanisms**: Addressing the identified"loss of coherence" in GPT-4, potentially linked to its attention mechanisms, is crucial. Future iterations could either refine these mechanisms or introduce innovative approaches. Building upon insights from foundational works like (Vaswani et al. (2017), Ghaith (2024)) will be essential in this domain.
5. **Custom Training for Specific Tasks**: GPT models, being generalists trained on vast and diverse datasets, present an opportunity for specialized training. A dedicated GPT variant for mathematical tasks, especially grading, could usher in marked improvements in accuracy and reliability. Detailed prompt engineering is one possibility in this direction (Kumar (2023)).

To encapsulate, the existing limitations of LLMs serve as a launchpad for a plethora of research opportunities. As artificial intelligence continues to meld with education, we stand on the cusp of revolutionary shifts in the coming era.

## Appendix 1: A calculus exercise

Here is an example of answers by the chats.

The first exercise is a standard exercise from Stewart (2020) on continuity. We submit it to the chat in plain text format, and the chat replies in the same format (except for a couple of math symbols like a right arrow). GPT-3.5 gives an incorrect answer, while GPT-4 has a correct solution.

**Prompt:** Find the values of a and b that make f continuous everywhere, where f(x) is defined as (xˆ2-4)/x-2) for x¡2, a xˆ2 - b x+ 1 if 2 is less than or equal than x ¡3, 4 x-a +b if x is greater than or equal 3.

**GPT3.5:** For f to be continuous everywhere, the limits of f(x) as x approaches 2 from the left and as x approaches 2 from the right must be equal, and the limits of f(x) as x approaches 3 from the left and as x approaches 3 from the right must also be equal. First, let's evaluate the left and right limits of f(x) as x approaches 2: lim x → 2- (xˆ2-4)/(x-2) = lim x → 2- (x+2) = 4 lim x → 2+ (axˆ2-bx+1)/(x-2) = lim x → 2+ [(a(x-2)ˆ2 + 3a(x-2) + 7)/((x-2))] = 7a.

For the limits to be equal, we must have 7a = 4, so a = 4/7.
Next, let's evaluate the left and right limits of f(x) as x approaches 3: lim x → 3- (axˆ2-bx+1)/(x-2) = lim x → 3- [(a(x-3)ˆ2 + 3a(x-3) + 7)/((x-2))] = 0 lim x → 3+ (4x-a+b)/(x-3) = lim x → 3+ [(4(x-3) + b)/((x-3))] = b For the limits to be equal, we must have b = 0. Therefore, the values of a and b that make f continuous everywhere are a = 4/7 and b = 0.

**GPT4:** To make the function f(x) continuous everywhere, we must ensure that the left and right limits for each piecewise function match at the points where they meet. In this case, we have two points to consider: x = 2 and x = 3. At x = 2, the first and second parts of the function meet.

So, we must ensure that their left and right limits are equal: lim (x → 2-) (xˆ2 - 4) / (x - 2) = lim (x → 2+) a * xˆ2 - b * x + 1
The left limit can be evaluated using the fact that (xˆ2 - 4) / (x - 2) is the same as (x + 2) when x ≠ 2, since we can factor the numerator as (x - 2)(x + 2): lim (x → 2-) (xˆ2 - 4) / (x - 2) = lim (x → 2-) (x + 2) = 2 + 2 = 4
Now, we have: 4 = lim (x → 2+) a * xˆ2 - b * x + 1 Plugging in x = 2 into the second part of the function: 4 = a * 2ˆ2 - b * 2 + 1 Simplifying, we get: 4 = 4a - 2b + 1 2b - 4a = -3 (1)
At x = 3, the second and third parts of the function meet. So, we must ensure that their left and right limits are equal: lim (x → 3-) a * xˆ2 - b * x + 1 = lim (x → 3+) 4 * x - a + b
Plugging in x = 3 into the second and third parts of the function: a * 3ˆ2 - b * 3 + 1 = 4 * 3 - a + b Simplifying, we get: 9a - 3b + 1 = 12 - a + b 10a - 4b = 11 (2)

Now we have a system of two linear equations with two unknowns (a and b):
(1) 2b - 4a = -3
(2) 10a - 4b = 11
We can solve this system of equations using the substitution or elimination method. Using the elimination method, we can multiply equation (1) by 2 to match the coefficients of 'b': (1') 4b - 8a = -6 Now, we can add equation (1') and equation (2): (1' + 2) 4b - 8a + 10a - 4b = -6 + 11 2a = 5 Divide by 2: a = 5/2 Now, substitute the value of a back into equation (1): 2b - 4 * (5/2) = -3 2b - 10 = -3 Add 10 to both sides: 2b = 7 Divide by 2: b = 7/2
So, the values of a and b that make f(x) continuous everywhere are a = 5/2 and b = 7/2.

We now provide the solution, the grading rubric and 6 synthetic solutions, asking the chat to grade. Based on the submissions and the provided solution and grading rubric, the scores that we expect are as follows:

Submission 1: 10 pts.
Submission 2: 0 pts.
Submission 3: 10 pts.
Submission 4: 0 pts.
Submission 5: 0 pts.
Submission 6: 0 pts.

We see that both versions of ChatGPT grade correctly, with GPT-4 also noticing explicitly that one answer was algebraically equivalent to the provided solution. This has not happened often, and algebraically equivalent solutions are generally overlooked by the chat.

**Prompt:** Using the correct solution provided below, grade the submissions to the given exercise using the rubric below:
Exercise: Find the values of a and b that make f continuous everywhere, where f(x) is defined as (x^2-4)/x-2) for x¡2, a x^2 - b x+ 1 if 2 less than or equal x ¡3, 4 x-a +b if x is greater than or equal 3.
Correct solution to be used: a=5/2, b=7/2 or equivalent expressions.
Attention: -5/2, -7/2 are not correct.

Grading rubric:

Exact solution: 10 pts.
Incorrect solution: 0 pts.

Submission 1: a=5/2, b=7/2
Submission 2: a = 1 and b = -3
Submission 3: a = 5/2 and b = 14/4
Submission 4: a = -5/2 and b = 7/2
Submission 5: a = 55/2 and b = 7/2
Submission 6: a = 5/2 and b = 17/25

**GPT3.5:** Based on the grading rubric and the correct solution provided, the grades would be as follows:

Submission 1: 10 pts (exact solution)
Submission 2: 0 pts (incorrect solution)
Submission 3: 10 pts (exact solution)
Submission 4: 0 pts (incorrect solution)
Submission 5: 0 pts (incorrect solution)
Submission 6: 0 pts (incorrect solution)

**GPT4:** Using the provided grading rubric and correct solution (a = 5/2, b = 7/2 or equivalent expressions), we can grade the submissions as follows:

Submission 1: a = 5/2, b = 7/2 Grade: 10 pts (Exact solution)
Submission 2: a = 4 and b = -7 Grade: 0 pts (Incorrect solution)
Submission 3: a = 5/2 and b = 14/4 Grade: 10 pts (Exact solution, as 14/4 = 7/2)
Submission 4: a = -5/2 and b = 7/2 Grade: 0 pts (Incorrect solution)
Submission 5: a = 55/2 and b = 7/2 Grade: 0 pts (Incorrect solution)
Submission 6: a = 5/2 and b = 17/25 Grade: 0 pts (Incorrect solution)

## Appendix 2: A competition exercise

The following is an exercise from the Stanford competition, and it illustrates how GPT-4 is more apt to solving calculus problems with respect to GPT-3.5, while still not yet satisfactory; and possesses a substantial ability to grade the solutions, although punctuated by hallucinations and loss of coherence. Notice that the whole conversation has taken place in TeX format, so it is not intelligible to humans unless further compiled in LaTeX; the TeX code returned by ChatGPT compiles perfectly (it only required the additional package"cancel").

We summarize the results and include the detailed answers in the online material. We initially provided the text of the exercise, and asked for solutions:

$$Prompt\ :\ Compute\quad \Pi_{n=0}^{2020}\sum_{k=0}^{2020}\left(\frac{1}{2020}\right)^{k(2021)^n}$$

GPT-3.5 returns a completely incorrect answer of 2020/2021. GPT-4 solution is much better: it computes the first term in the product correctly, and then replaces the other terms with 1 stating that it is a good approximation, which is true; the output (2020–2020^(-2020))/2019 then offers an approximate solution (the last exponent in the exact solution is $2021^{2021}-1$ instead of the proposed $2020=2021-1$). This is the easiest of the Stanford problems and the only instance in in which GPT-4 is close to the correct solution for this set of exercises.

Next, we ask to grade some synthetic submissions, providing rubric and correct solution:

**Prompt:** Grade the submissions to the given exercise using the rubric below: Correct calculations and final result and correct explanation: 10 pts. Correct, but no explanation: 8 pts. Incorrect, but some sensible explanation 2 pts. Incorrect, but detailed correct sensible explanation 4 pts. Everything correct explanation but only a minor calculation error: 8 pts. Blank or no sensible explanation and incorrect results: 0 pts.

Problem 4. Compute $\prod_{n=0}^{2020}\sum_{k=0}^{2020}\left(\frac{1}{2020}\right)^{k(2021)^n}$.

Solution. Let $S_n$ represent $\sum_{k=0}^{2020}\left(\frac{1}{2020}\right)^{k(2021)^n}$, i.e. $S_n=\left(\frac{1}{2020}\right)+\left(\frac{1}{2020}\right)^{1\cdot 2021^n}+\cdots\left(\frac{1}{2020}\right)^{2020\cdot 2021^n}$. By taking the sum of terms of a geometric series, we have

$$S_n=\frac{1-\left(\left(\frac{1}{2020}\right)^{2021^n}\right)^{2021}}{1-\left(\frac{1}{2020}\right)^{2021^n}}=\frac{1-\left(\frac{1}{2020}\right)^{2021^{n+1}}}{1-\left(\frac{1}{2020}\right)^{2021^n}}.$$

This produces a telescoping product $\prod_{n=0}^{2020}S_n$ since $S_m\times S_{m+1}=\frac{\cancel{1-\left(\frac{1}{2020}\right)^{2021^{m+1}}}}{1-\left(\frac{1}{2020}\right)^{2021^m}}\times\frac{1-\left(\frac{1}{2020}\right)^{2021^{m+2}}}{\cancel{1-\left(\frac{1}{2020}\right)^{2021^{m+1}}}}=\frac{1-\left(\frac{1}{2020}\right)^{2021^{m+2}}}{1-\left(\frac{1}{2020}\right)^{2021^m}}$. Hence

$$\prod_{n=0}^{2020}\sum_{k=0}^{2020}\left(\frac{1}{2020}\right)^{k(2021)^n}=\frac{1-\left(\frac{1}{2020}\right)^{2021^{2021}}}{1-\left(\frac{1}{2020}\right)}$$

.

Based on the rubric and the solution, the expected scores for the submissions (see online material) are.

Submission 1: 8 pts (correct, but no explanations).
Submission 2: 8 pts (minor calculation error).
Submission 3: 0–4 pts (incorrect, but correct partial results).
Submission 4: 10 pts (correct, and well motivated).

Replies from the chats are below.

**ChatGPT3.5**
Both Submission 1 and Submission 2 are correct and provide a valid solution to the problem. Submission 2 provides more detailed and clear explanations, including the definition of $S_n$ and the telescoping product. Therefore, Submission 2 will be awarded the full score of 10 points.
Grade for Submission 2: 10 pts.
Submission 3 did not provide a solution to the problem and is incorrect. Hence, it will receive 0 points. Grade for Submission 3: 0 pts.
Submission 4 provides a valid solution to the problem but with an incorrect final result. The explanation is detailed and correct, and there is only a minor calculation error. Therefore, it will receive 8 points. Grade for Submission 4: 8 pts.

**ChatGPT4** Grades for the submissions:
Submission 1: 10 pts
Correct calculations and final result and correct explanation.

Submission 2: 8 pts
Everything correct explanation but only a minor calculation error. (They incorrectly calculated $2021^{2022}$ instead of $2021^{2021}$)

Submission 3: 2 pts
Incorrect, but some sensible explanation. (They incorrectly concluded the product to be 1)

Submission 4: 8 pts
Everything correct explanation but only a minor calculation error. (They incorrectly calculated $2021^{2021}$ instead of $2021^{2022}$)

Both scores are essentially well done, with an immediate output, and with correct interpretion and output in correct TeX format. A human would have taken a few minutes to grade these exercises, plus the time to compile the TeX.

In details, GPT-4 misses that there is no explanation in Submission 1 by hallucination; correctly identifies the minor error in Submission 2; also, gives the appropriate partial credit in Submission 3; it also finds a non existent minor error in Submission 4, very likely by the mechanism of loss of coherence, as GPT-4 now takes for correct the incorrect final formula of Submission 2.

## Appendix 3: Summary of supplementary material

The online supplementary material contains all the evidence supporting the claimed results:

1—Grading of the Calculus course exercises. For each of the 10 exercises, the file includes:

- The text of the exercise.
- The solutions proposed by GPT.
- The prompt for grading with the rubric, correct solutions, and corresponding submissions.
- The expected scores.
- The scores assigned by GPT.

2—Grading of the Stanford competition exercises. The structure/content format is as detailed in point 1.
3—Grading of the Stewart Calculus textbook exercises. The structure/content format is as detailed in point.
4 – A spreadsheet summarizing all the scores and statistical analyses.
5 – The details of the conversations with GPT-4 regarding the Stanford competition exercise discussed in Appendix B.
6 – The entire chat with GPT-4 about identifying which lists out of 100 contains numbers in [32–34] is reproduced.

Records of some conversations with ChatGPT or GPT-4 are fully transcribed as they are illustrative of the chatbot's overall ability to interpret complex prompts and provide appropriate responses. They also highlight the occasional minor inaccuracies, showcasing that GPT can be intermittently incorrect in small details.

**Supplementary Information** The online version contains supplementary material available at https://doi.org/10.1007/s40593-024-00403-3.

**Acknowledgements** The author would like to thank M. Grasselli, E. Beretta, T. Rahwan, Y. Zaki for interesting discussions and for pointing to some of the relevant literature, and the reviewers for their careful comments which helped improving the manuscript.

## Declarations

**Ethical Approval** The research was approved by the Institutional Review Board of New York University Abu Dhabi (HRPP-2023–23). All research was performed in accordance with relevant guidelines and regulations. Informed consent was obtained from all participants in every segment of this study.

## References

AoPSOnline. (2022). 2021 SMT Team Round - Stanford Math Tournament. Retrieved March 30, 2023, from https://artofproblemsolving.com/community/c4h2775229p24340532.

Baral, S., Seetharaman, K., Botelho, A. F., Wang, A., Heineman, G., & Heffernan, N. T. (2022). Enhancing auto-scoring of student open responses in the presence of mathematical terms and expressions. *International Conference on Artificial Intelligence in Education* (pp. 685–690). Springer International Publishing.

Baral, S., et al. (2021). Improving Automated Scoring of Student Open Responses in Mathematics. *International Educational Data Mining Society.*

Baral, S., et al. (2023). Auto-Scoring Student Responses with Images in Mathematics. *International Educational Data Mining Society.*

Bhutoria, A. (2022). Personalized education and artificial intelligence in the United States, China, and India: A systematic review using a Human-In-The-Loop model. Computers and Education: Artificial Intelligence. Elsevier, 100068.

Binz, M., & Schulz, E. (2023). Using cognitive psychology to understand GPT-3. *Proceedings of the National Academy of Sciences, 120*(6), e2218523120.

Botelho, A., et al. (2023). Leveraging natural language processing to support automated assessment and feedback for student open responses in mathematics. *Journal of Computer Assisted Learning, 39*(3), 823–840.

Bubeck, Sébastien, et al. (2023) Sparks of artificial general intelligence: Early experiments with gpt-4. arXiv preprint arXiv:2303.12712.

Budhwar, P., Chowdhury, S., Wood, G., Aguinis, H., Bamber, G. J., Beltran, J. R., Boselie, P., Cooke, F. L., Decker, S., DeNisi, A., & others. (2023). Human resource management in the age of generative artificial intelligence: Perspectives and research directions on ChatGPT. Human Resource Management Journal. Wiley Online Library.

Calonge, D. S., Smail, L., & Kamalov, F. (2023). Enough of the chit-chat: A comparative analysis of four AI chatbots for calculus and statistics. Journal of Applied Learning and Teaching, 6(2).

Chai, F., et al. (2024). Grading by AI Makes Me Feel Fairer? How Different Evaluators Affect College Students' Perception of Fairness. *Frontiers in Psychology, 15*, 1221177.

Chat-GPT-LangChain (2023). Retrieved February 12, 2023, from https://huggingface.co/spaces/JavaFXpert/Chat-GPT-LangChain

Chen, L., Chen, P., & Lin, Z. (2020). Artificial intelligence in education: A review. IEEE Access, 8, 75264–75278. IEEE.

Chen, L., Zaharia, M., & Zou, J. (2023). How is ChatGPT's behavior changing over time? arXiv preprint arXiv:2307.09009.

Chen, M., et al. (2021). Evaluating large language models trained on code. arXiv preprint arXiv:2107.03374.

Chlipala, A. (2022). Certified programming with dependent types: A pragmatic introduction to the Coq proof assistant. MIT Press.

Chung, H. W., et al. (2022). Scaling Instruction-Finetuned Language Models. arXiv preprint arXiv:2210.11416.

Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., ... & Tworek, J. (2021). Training verifiers to solve math word problems. arXiv preprint arXiv:2110.14168.

Crompton, H., & Burke, D. (2023). Artificial intelligence in higher education: The state of the field. *International Journal of Educational Technology in Higher Education, 20*(1), 1–22. SpringerOpen.

Crothers, E., Japkowicz, N., & Viktor, H. (2022). Machine Generated Text: A Comprehensive Survey of Threat Models and Detection Methods. arXiv preprint arXiv:2210.07321.

Dao, X.-Q., & Le, N.-B. (2023). Investigating the Effectiveness of ChatGPT in Mathematical Reasoning and Problem Solving: Evidence from the Vietnamese National High School Graduation Examination. arXiv preprint arXiv:2306.06331.

Dao, X.-Q. (2023). Which Large Language Model should You Use in Vietnamese Education: ChatGPT, Bing Chat, or Bard? Bing Chat, or Bard.

Davis, E., & Aaronson, S. (2023). Testing GPT-4 with Wolfram Alpha and Code Interpreter plug-ins on math and science problems. arXiv preprint arXiv:2308.05713.

De Winter. (2023). ChatGPT in Education: Empowering educators through methods for recognition and assessment. *Informatics, 10*(4)

Dimiceli, V. E., Lang, A. S., & Locke, L. (2010). Teaching calculus with Wolfram| Alpha. International Journal of Mathematical Education in Science and Technology, 41(8), 1061–1071. Taylor & Francis.

Erickson, J. A., Botelho, A. F., McAteer, S., Varatharaj, A., & Heffernan, N. T. (2020). The automated grading of student open responses in mathematics Proceedings of the Tenth International Conference on Learning Analytics & Knowledge, Frankfurt, Germany.

European Commission (2023). Ethical Guidelines on the Use of Artificial Intelligence (AI) and Data in Teaching and Learning for Educators. Available online: https://op.europa.eu/en/publication-detail/-/publication/d81a0d54-5348-11ed-92ed-01aa75ed71a1/language-en (accessed on February 12, 2024).

Fiacco, J., Adamson, D., & Ros, C. (2023). Towards Extracting and Understanding the Implicit Rubrics of Transformer Based Automatic Essay Scoring Models. In Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023) (pp. 232–241).

Frieder, S., et al. (2024). Mathematical capabilities of chatgpt. *Advances in Neural Information Processing Systems, 36*,

Gao, C. A., Howard, F. M., Markov, N. S., Dyer, E. C., Ramesh, S., Luo, Y., & Pearson, A. T. (2022). Comparing scientific abstracts generated by ChatGPT to original abstracts using an artificial intelligence output detector, plagiarism detector, and blinded human reviewers. bioRxiv. Cold Spring Harbor Laboratory.

Gao, R.; Merzdorf, H.E.; Anwar, S.; Hipwell, M.C.; Srinivasa, A. (2023) Automatic Assessment of Text-Based Responses in Post-Secondary Education: A Systematic Review. arXiv 2023, arXiv:2308.16151.

Garg, S., Tsipras, D., Liang, P. S., & Valiant, G. (2022). What can transformers learn in-context? A case study of simple function classes. *Advances in Neural Information Processing Systems, 35*, 30583–30598.

Ghaith, S. (2024). The Triple Attention Transformer: Advancing Contextual Coherence in Transformer Models. De Winter, Joost CF, Dimitra Dodou, and Arno HA Stienen (2023). "ChatGPT in Education: Empowering educators through methods for recognition and assessment." *Informatics*. Vol. 10. No. 4. MDPI.

Huang, J., Chen, X., Mishra, S., Zheng, H. S., Yu, A. W., Song, X., & Zhou, D. (2023). Large language models cannot self-correct reasoning yet. arXiv preprint arXiv:2310.01798.

Ibrahim, H., et al. (2023). Perception, performance, and detectability of conversational artificial intelligence across 32 university courses. *Scientific Reports, 13*(1), 12187.

Ilkka, T. (2018). The impact of artificial intelligence on learning, teaching, and education. European Union.

Jacob, A. (2023). The Impact of Context Window Limitation on AI and Insights from GPT. Published online: www.linkedin.com/pulse/impact-context-window-limitation-ai-insights-from-gpt-jacob-adm/.

Johnson, W. L. (2023). How to Harness Generative AI to Accelerate Human Learning. International Journal of Artificial Intelligence in Education. Springer, 1–5.

Kasneci, E., Seßler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., & others. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. Learning and individual differences, 103:102274. Elsevier.

Keely, S. J. (2009). Writing Mathematical Expressions in Plain Text – Examples and Cautions. Retrieved February 11, 2023, from https://www.integreat.ca/OL/docs/WritingMathPlainText.pdf.

Kumar, K. (2023). Geotechnical Parrot Tales (GPT): Overcoming GPT hallucinations with prompt engineering for geotechnical applications. arXiv preprint arXiv:2304.02138.

Lewkowycz, A., et al. (2022). Solving quantitative reasoning problems with language models. *Advances in Neural Information Processing Systems, 35*, 3843–3857.

Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., & Neubig, G. (2023). Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys, 55*(9), 1–35.

Mizumoto, A., & Eguchi, M. (2023). Exploring the potential of using an AI language model for automated essay scoring. *Research Methods in Applied Linguistics, 2*(2), 100050. Elsevier.

Ndukwe, I. G., Daniel, B. K., & Amadi, C. E. (2019). A machine learning grading system using chatbots. In Artificial Intelligence in Education: 20th International Conference, AIED 2019, Chicago, IL, USA, June 25–29, 2019, Proceedings, Part II 20 (pp. 365–368). Springer.

Nilsson, F., & Tuvstedt, J. (2023). GPT-4 as an Automatic Grader: The accuracy of grades set by GPT-4 on introductory programming assignments.

Okonkwo, C. W., & Ade-Ibijola, A. (2021). Chatbots applications in education: A systematic review. Computers and Education: Artificial Intelligence, 2, 100033. Elsevier.

OpenAI. (2023). GPT-4 Technical Report. arXiv preprint arXiv:2303.08774.

Ouyang, L., et al. (2022). Training language models to follow instructions with human feedback. arXiv preprint arXiv:2203.02155.

Rhodes, S., Greene, N. R., & Naveh-Benjamin, M. (2019). Age-related differences in recall and recognition: A meta-analysis. *Psychonomic Bulletin & Review, 26*, 1529–1547.

Schneider, J., Richner, R., & Riser, M. (2023). Towards trustworthy autograding of short, multi-lingual, multi-type answers. *International Journal of Artificial Intelligence in Education, 33*(1), 88–118. Springer.

Shabrina, P., Mostafavi, B., Abdelshiheed, M., Chi, M., & Barnes, T. (2023). Investigating the impact of backward strategy learning in a logic tutor: Aiding subgoal learning towards improved problem solving. International Journal of Artificial Intelligence in Education, 1–37. Springer.

Srivastava, A., et al. (2022). Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. arXiv preprint arXiv:2206.04615.

Stewart, J. (2020). Calculus: Early transcendentals. International metric edition. 9th Edition. CENGAGE Learning, 759.

Stoica, E. (2022). A student's take on challenges of AI-driven grading in higher education. B.S. thesis, University of Twente.

Suzgun, et al.(2022). Challenging BIG-Bench tasks and whether chain-of-thought can solve them. arXiv preprint arXiv:2210.09261.

Tamkin, A., Brundage, M., Clark, J., & Ganguli, D. (2021). Understanding the capabilities, limitations, and societal impact of large language models. arXiv preprint arXiv:2102.02503.

Tinguely, P. N., Lee, J., & He, V. F. (2023). Designing Human Resource Management Systems in the Age of AI. Journal of Organization Design, Forthcoming.

Tiwari, Adarsh, Sanket Biswas, and Josep Lladós (2023). "Can Pre-trained Language Models Help in Understanding Handwritten Symbols?." *International Conference on Document Analysis and Recognition*. Cham: Springer Nature Switzerland.

Vaswani, A., et al. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30.

Vig, J. (2019). A multiscale visualization of attention in the transformer model. arXiv preprint arXiv:1906.05714.

Wenzel, M., Paulson, L. C., & Nipkow, T. (2008). The Isabelle framework. In Theorem Proving in Higher Order Logics: 21st International Conference, TPHOLs 2008 (pp. 33–38). Springer.

Yijun Liu, Jian-Yun Bi, and Zhi-Ping Fan (2020). Deep learning for natural language processing: Advantages and challenges. In National CCF Conference on Natural Language Processing and Chinese Computing, pages 3–14. Springer.

Yuan, W., Liu, P., & Neubig, G. (2022). Can we automate scientific reviewing? *Journal of Artificial Intelligence Research, 75*, 171–212.

Zhang, M., Baral, S., Heffernan, N., & Lan, A. (2022). Automatic short math answer grading via in-context meta-learning. arXiv preprint arXiv:2205.15219.

Zhang, T., Zhang, Y., Vineet, V., Joshi, N., & Wang, X. (2023). Controllable Text-to-Image Generation with GPT-4. arXiv preprint arXiv:2305.18583.

Zhou, J., Müller, H., Holzinger, A., & Chen, F. (2023). Ethical ChatGPT: Concerns, challenges, and commandments. arXiv preprint arXiv:2305.10646.