Can ChatGPT Pass High School Exams on English Language Comprehension?

de Winter, J.C.F.

DOI
10.1007/s40593-023-00372-z

Publication date
2023

Document Version
Final published version

Published in
International Journal of Artificial Intelligence in Education

International Journal of Artificial Intelligence in Education
https://doi.org/10.1007/s40593-023-00372-z

ARTICLE

Can ChatGPT Pass High School Exams on English Language Comprehension?

Joost C. F. de Winter[1]

Accepted: 4 September 2023

Abstract

Launched in late November 2022, ChatGPT, a large language model chatbot, has garnered considerable attention. However, ongoing questions remain regarding its capabilities. In this study, ChatGPT was used to complete national high school exams in the Netherlands on the topic of English reading comprehension. In late December 2022, we submitted the exam questions through the ChatGPT web interface (GPT-3.5). According to official norms, ChatGPT achieved a mean grade of 7.3 on the Dutch scale of 1 to 10—comparable to the mean grade of all students who took the exam in the Netherlands, 6.99. However, ChatGPT occasionally required re-prompting to arrive at an explicit answer; without these nudges, the overall grade was 6.5. In March 2023, API access was made available, and a new version of ChatGPT, GPT-4, was released. We submitted the same exams to the API, and GPT-4 achieved a score of 8.3 without a need for re-prompting. Additionally, employing a bootstrapping method that incorporated randomness through ChatGPT's 'temperature' parameter proved effective in self-identifying potentially incorrect answers. Finally, a re-assessment conducted with the GPT-4 model updated as of June 2023 showed no substantial change in the overall score. The present findings highlight significant opportunities but also raise concerns about the impact of ChatGPT and similar large language models on educational assessment.

Keywords GPT-3.5 · GPT-4 · Large language model · Educational assessment · Reading comprehension

✉ Joost C. F. de Winter
j.c.f.dewinter@tudelft.nl

[1] Cognitive Robotics Department, Delft University of Technology, Delft, Netherlands

Published online: 13 September 2023

Introduction

ChatGPT, developed by OpenAI, is a chatbot designed to engage in conversation with users and generate responses that are human-like and meaningful. Launched on November 30, 2022, it gained widespread attention due to its ability to generate coherent output across a wide range of topics, including creative writing (King, 2023; Kirmani, 2023; Zhai, 2022), computer coding and bug fixing (Davis et al., 2023; Sobania et al., 2023), teaching innovation (Rudolph et al., 2023; Šlapeta, 2023), sentiment analysis (Tabone & De Winter, 2023; Zhong et al.,2023a), and various annotation tasks, such as genre identification (Kuzman et al., 2023) and the identification of language explicitness (Huang et al., 2023; Rospocher & Eksir, 2023). However, it has also been argued that ChatGPT output tends to contain incorrect statements, cliché-like writing, and faulty references to scientific sources (Han et al., 2023; Kim et al., 2022; Lovin, 2022; Vincent, 2022; Whitford, 2022).

Performing benchmark tasks on large language models like ChatGPT is crucial to understand their strengths and weaknesses, as well as to identify potential areas for improvement. A recent study by Gilson et al. (2022) revealed that ChatGPT demonstrates a reasonable performance on medical licensing exams, with results comparable to third-year medical students in one of the four exams assessed. Kung et al. (2023) corroborated this finding by demonstrating that the model achieved scores near the passing threshold for the United States Medical Licensing Exam (USMLE). However, ChatGPT's performance was found to be below average in physics (Kortemeyer, 2023) and in various mathematical domains, such as Olympiad problem-solving, functional analysis, and symbolic integration (Frieder et al., 2023). Similarly, Newton and Xiromeriti (2023) reported that ChatGPT lagged behind the average student in multiple-choice tests across diverse fields, including ophthalmology, law, economics, and physics. Therefore, further research is necessary to evaluate ChatGPT's performance across different types of exams.

The aforementioned evidence pertains to GPT-3.5. On March 14, 2023, a subsequent version, GPT-4, was introduced, accompanied by API (Application Programming Interface) access. OpenAI (2023) claimed that GPT-4 exhibits superior performance compared to GPT-3.5, particularly in the realm of intricate and nuanced communication. Bordt and Von Luxburg (2023) found that, in an undergraduate computer science exam, GPT-4 obtained 17% more points than GPT-3.5, approaching average student performance. Savelka et al. (2023) investigated the proficiency of GPT-4 in postsecondary Python programming assessments and reported a substantial improvement over previous generations, where GPT-4 can now pass courses independently. However, limitations remained in handling complex coding exercises. Bommarito and Katz (2022) found that GPT-3.5 achieved an average score of 45.1% on the Uniform Bar Examination, encompassing subjects such as civil procedure, constitutional law, and contracts, while in their most recent update, GPT-4 attained a remarkable score of 75.7% (Katz et al., 2023). A similar analysis conducted by OpenAI (2023), comparing the performance of both models on the Uniform Bar Examination, revealed scores of 53% and 75% for GPT-3.5 and GPT-4, respectively. These results correspond to the performance of the lowest 10% of human test-takers for GPT-3.5 and the highest 10% for GPT-4. OpenAI (2023) conducted several other

benchmark evaluations to compare GPT-4 with GPT-3.5 and demonstrated improved performance for GPT-4 in nearly all instances. Of note, while GPT-4 excelled in the GRE verbal section (99th percentile), it scored modestly in GRE writing (~54th percentile).

When we conducted initial explorations of GPT-3.5's and GPT-4's performance on various online tests, we observed that ChatGPT was proficient at verbal analogies commonly found in IQ tests but faced difficulties with solving logical puzzles that require variable tracking, such as determining which guest was staying in a particular hotel room (available at https://www.123test.com/verbal-reasoning-test). This preliminary observation is in line with recent more comprehensive analyses (Arora & Singh, 2023; Bubeck et al., 2023; Zhong et al., 2023b), which explicate that ChatGPT exhibits certain limitations, specifically its inability to globally plan, engage in an 'inner dialogue', or self-correct its errors. While ChatGPT demonstrates remarkable emergent properties such as sophisticated language production and basic arithmetic (Kosinski, 2023; Wei et al., 2022), it essentially operates as an autoregressive model, which means that it predicts each next word by conditioning on the words that have already been produced, thus generating text one word at a time in a sequential manner. In light of these characteristics, it becomes compelling to investigate whether ChatGPT can successfully execute a high school reading comprehension exam—a task that, upon initial observation, one might expect necessitates intricate reasoning.

In this study, we employed ChatGPT versions 3.5 and 4 to analyze their performance on Dutch national exams, focusing on English reading comprehension. In their technical report, OpenAI (2023) recognized the potential risk of contamination in their benchmark evaluations, where the model might inadvertently have access to the test questions and corresponding answers. OpenAI reported only a marginal difference between the performance of 'contaminated' and 'non-contaminated' questions. However, their approach to screening contaminated data could be susceptible to misses and false positives. To mitigate the contamination issue, the current study used exams from 2022 (i.e., after the ChatGPT knowledge cut-off date), ensuring that ChatGPT's training datasets did not contain the exams or answer models being examined. A recent study by Chen et al. (2023) highlighted a marked decline in ChatGPT's (GPT-4) performance on selected subtasks between March and June 2023. Consequently, the criticality of evaluating the capabilities of ChatGPT becomes increasingly apparent. Therefore, in the present study, we assessed two specific versions of GPT-4, those of March and June, to better understand possible shifts in performance.

Methods

We applied ChatGPT (version: December 15, 2022) to national exams of the VWO program (Preparatory Scientific Education) in the Netherlands that tested English reading comprehension. These high-stakes exams are administered by the Dutch organization CITO (CITO, 2023), and are mandatory for all VWO students. The VWO program is considered the most academically rigorous high-school education in the Netherlands, designed for students who intend to pursue university-level studies. In the VWO English exam, students are tasked with reading a variety of passages,

such as newspaper items, and answering associated questions. Although the specific texts and questions change with each exam, the nature of the questions, such as identifying the main idea, making inferences, and interpreting vocabulary within context, remains largely consistent. In this study, only exams from 2022 were used because the database of ChatGPT has a knowledge cut-off in 2021.

The three available exams incorporated a total of 43, 40, and 41 questions, respectively. Each exam comprised an accompanying textbook with 11 textual passages. Of the three exams, there were 31, 32, and 29 multiple-choice questions, respectively. These questions typically had four response options [A, B, C, and D]. However, some questions offered three (A–C: 13 questions), five (A–E: 7 questions), or six (A–F: 1 question) response options. Every multiple-choice question carried the value of one point. Further, the exams incorporated a number of open questions also worth one point each. These questions either required succinct answers, called for a sequence of statements to be arranged in a specific order, or asked for a response to a maximum of two sub-questions in the form of a 'yes' or 'no'. Examples of such questions were as follows (translated from Dutch to English):

- "8. 'why should he lose his perk?' (paragraph 4). What was this 'perk'? Please respond in Dutch." (Exam 1).
- "20. The text divides into critical and non-critical segments. In which paragraph does the critical part commence? Please indicate the paragraph number." (Exam 1).
- "2. 'European ruling' (title). Please determine whether Mike Short identifies each of the following points as an issue with the introduction. Write "yes" or "no" alongside each number on the answer sheet.

1. The implementation is impeded by the current web infrastructure.
2. The interpretation of the 'right to be forgotten' is contingent upon one's cultural background." (Exam 3).

The remaining questions (3 to 6 per exam), worth either 2 or 4 points, were multi-part items. More specifically, these involved scenarios where 4 to 8 statements needed a 'yes' or 'no' response, or items where multiple themes or locations in the text had to be identified.

The text (e.g., a news item or another text fragment that was part of the exam) and the corresponding questions were manually submitted one by one to the ChatGPT web interface (GPT-3.5), as part of the same chat session for each exam. After the text and before each question, a prompt was included, e.g., "*Based on 'Text 5' above, please choose the correct response option between A, B, C, and D for the question below (note that the number in front of each paragraph indicates the number of the paragraph)*".

In 15% of cases, ChatGPT appeared to misconstrue the question, leading to invalid responses. These included not selecting any options in a multiple-choice question, generating an entirely new question, or asserting that the correct answer could not be determined due to insufficient information. When such a scenario occurred, the researcher would either reiterate the question or provide further clarification. This

could involve prompting ChatGPT to choose from the response alternatives or explaining that there was a missing word in the text that required completion. This method occasionally facilitated a response. For example, Question 9 of Exam 3 asked "*Based on 'Text 5' above, please choose the correct response option between A, B, C, D, and E for the question below (note that the number in front of each paragraph indicates the number of the paragraph). Which of the following fits the gap (indicated with "…") in paragraph 1?"*, together with five response alternatives (A–E). Note that the phrase "(indicated with "…")" was our addition to aid ChatGPT. The initial response of ChatGPT was an elaborate general reflection: "*According to data collected by social scientists, there is little evidence that the typical terrorist is poor or poorly educated. … Instead, other factors such as political and religious ideology may be more influential in determining who becomes a terrorist."* This is why the prompt was repeated, after which ChatGPT offered a response: "*B: little to lose*". Out of a total of 124 questions posed, 18 required re-prompting once or twice, resulting in an eventual response in 15 of these instances. The model's input and output during this process were documented and can be found in the supplementary material.

The three available 'VWO English 2022' exams were completed by ChatGPT. ChatGPT's answers were assessed by an independent experimenter and checked by the author of this work. The assessment was done using official scoring booklets containing the correct answers (College voor Toetsen en Examens, 2022). These booklets also contained clear rules for assessment. One of these rules is: "*If more than one example, reason, elaboration, quote or other type of answer is asked for, only the first given answers are assessed, up to the maximum number requested*", and another rule is "*The correct answer to a multiple choice question is the capital letter that belongs to the correct choice. If the answer is given in a different way, but it is unequivocally established that it is correct, then this answer should also be counted as correct.*". It is noted that although the experimenter knew this was an assessment of ChatGPT, the exams were designed in such a way that the answers left very little room for doubt since the correct answers were mostly numbers, keywords, letters (e.g., A, B, C, D), or yes/no statements. The assessor and author had no disagreements about whether particular answers should be scored as correct or incorrect. Finally, the number of points obtained on the exams were converted to a grade mark on a scale from 1.0 to 10.0 using formal conversion tables (College voor Toetsen en Examens, 2022).

On March 17, 2023, the above analysis was repeated using GPT-4 (model GPT-4-0314). Rather than manually inputting the comprehension texts and questions to the ChatGPT web interface, the API was used, and the procedure was fully automated in MATLAB (version R2021b). Specifically, a script was written (see supplementary material) that read the PDF files of the Text booklet and the Question booklet per exam, grouped the questions by text based on their headers, and fed the questions to the OpenAI API. When reading the text and questions, line breaks were removed, but no additional processing was done. For example, page numbers, points per question and page headers and footers were not removed.

The following prompt was used in MATLAB to feed the questions to the OpenAI API: "*Answer all questions, including multiple-choice questions. Only provide the question number and the answer, nothing else*:". The prompt subsequently included the word 'TEXT:' in a new line. This was followed by the entire text and the word

'QUESTIONS:', which preceded all the questions for that text. This prompt design was used for all questions. It was adopted to ensure that the API only outputs the answers without additional clarifying text, which speeds up the response by ChatGPT. In all instances, GPT-4 provided concise answers, typically a letter, number, or keyword as requested. This enabled a straightforward manual evaluation of the examinations.

An example of a complete prompt—in this case, Text 10 from Exam 2 and three corresponding questions (Questions 36–38)—is presented in Fig. 1. It can be seen that the text has not been preprocessed; the prompt represents how the PDF files of the text booklet and the exam were automatically read in, which is accompanied by unnecessary spaces, headers and footers, and information for the candidate about how many points the question is worth. Attempts were also made to submit cleaned-up text to ChatGPT instead of texts in their raw form. However, this did not appear to lead to improvement in the accuracy of the answers provided by GPT-4. We submitted raw text data, as depicted in Fig. 1, to provide an evaluation of ChatGPT free from human intervention.

More recently, in June 2023, an update of GPT-4 was released. An evaluation by Chen et al. (2023) reported that this version performed highly differently on certain tasks, such as answering sensitive questions or generating code, compared to the March version, potentially having far-reaching consequences for users and applications. However, apart from the work of Chen et al. (2023), there is currently limited information in the literature regarding any disparities in output between the March and June versions. Hence, we conducted a re-analysis using the June version (GPT-4-0613).

While the web interface of ChatGPT has an element of stochasticity in its responses, the API offers the ability to modulate this randomness via a 'temperature' parameter that is adjustable within a continuum from 0 to 2. With a temperature setting at 0, the output is highly reproducible. In contrast, a temperature setting at 2 introduces a significant degree of randomness or creativity in the output. For our study, the tem-

Answer all questions, including multiple-choice questions. Only provide the question number and the answer, nothing else:
TEXT:
Tekst 10 Ama-San review: a deep dive into Japan's fisherwomen culture adapted from an article by Leslie Felperin 1 The ama are Japan's fisherwomen, free divers who retrieve abalone, sea snails and other ocean products (they're best known for their pearl fishing) out of the shallows without using oxygen tanks. Portuguese documentarian Cláudia Varejão immerses herself in the daily rhythms and rituals of one group, filming them at home and at work as they go about raising kids, singing karaoke and swimming to the bottom of the sea. 2 Varejão favours an austere approach t hat relies on long, unblinking takes, uses no music that doesn't occur within the action itself and no subtitles that clarify who's who. 37 , there are no explanat ions at all, leaving the viewer to work out why, for instance, the women wear both modern diving suits and traditional linen headscarves over their waterproof balaclavas. Much screen time is devoted to watching the subjects wrapping, folding and tucking thebits of white cloth, a kind of origami that's seemingly both symbolic and se practical, like the tying up of the lacethey favour dress that similarly mixes on a ballet slipper. Out of the water, s modern and traditional, with regular trousers and blouses below the ne ck and white bonnets with deep brims on their heads. 3 This kind of unfiltered anthropologica l study can be mesmerising and there are some lovely sequences here, not ju st of the women diving in the olive- green depths but also moments where they're just hanging out with their families, playing with fireflies or ma king supper. But some viewers may find it frustrating that we never hear them discuss their lives or even learn their names properly, as if we are just ghos ts, weaving among them while they go about their business. It's a style of fil m-making that's as traditional and in its way mannered as the head wraps and di ving techniques that are being observed. theguardian.com, 2019 VW-1002-a-22-2-b lees verder ►►► 17 / 19 Lees bij de volgende tekst steeds eerst de vraag voordat je de tekst zelf raadpleegt.
QUESTIONS:
Tekst 10 Ama-san review 36 What becomes clear about t he ama in paragraphs 1 and 2? A They adhere to economically unviable methods despite needing the extra income badly. B They choose entertainment that comp lies with the strict cultural codes that govern their existence. C They integrate long-established cu stoms and fashions with up-to-date elements. D They tend to be quite secretive about the mysterious rituals employed by their kind. 1p 1p 37 Which of the following fits the gap in paragraph 2? A Besides B Indeed C Instead D Still 38 What is the writer's final verdic t about the documentary discussed in this review? A It cleverly uses innovative methods to capture a way of life that has hardly changed through the years. B It contains some captivating images but never becomes truly involved on a more personal level. C Its promising start is not enough to compensate for the inferior quality of much of the footage. D Its subject matter is fascinating but this mediocre portrayal does not do it justice. 1p 1p Let op: de laatste vragen van dit examen staan op de volgende pagina. VW-1002-a-22-2-o 9 / 10 lees verder ►►► Lees bij de volgende opgaven steeds eerst de vraag voordat je de bijbehorende tekst raadpleegt.

Fig. 1 One of the prompts submitted to GPT-4 (Questions 36–38 of Exam 2)

perature parameter was set to 0. The entire analysis, from automatic reading of the texts and questions to letting GPT-4 produce the responses, took about 40 s per exam.

Results

Table 1 presents the results of the exams completed by ChatGPT. In the Netherlands, a mean grade of 5.50 or higher across all courses would imply a pass of the high school diploma. It can be seen that GPT-3.5 would pass each of the English exams, with an overall mean grade across the exams of 7.3, whereas GPT-4-0314 and GPT-4-0613 had overall mean grades of 8.3 and 8.1, respectively.

Upon categorizing the test items into (1) multiple-choice questions, (2) open one-point questions, and (3) open questions valued at more than one point, it became evident that GPT-3.5 faced some challenges with the second category (79%, 68%, and 77% of the points earned). Compared to GPT-3.5, GPT-4 showed improvement in the multiple-choice questions (GPT-4-0314: 92%, 84%, 80%; GPT-4-0613: 92%, 63%, 83%).

It should be acknowledged that GPT-3.5 had an inherent advantage, as we occasionally provided a repeated prompt through the web interface to procure an explicit response (see Methods section for details). Upon limiting our evaluation to only the first instance of its (non-)response, GPT-3.5's mean score for the three exams was 6.5, as shown in Table 1. In contrast, the interaction with GPT-4 was fully automated via the API, without any re-prompting. We also made an attempt to use GPT-3.5 via the API (model GPT-3.5-turbo-0301), but the performance was deemed unsatisfactory. The grades for Exams 1, 2, and 3 were only 5.7, 6.7, and 7.0 respectively. The score for the first exam was particularly low, as GPT-3.5 did not answer 7 of the 43 questions. Considering these outcomes, we chose not to further investigate this approach.

Bootstrapping Self-Consistency Analysis

As shown above, GPT-3.5 performed comparably to the average Dutch high school student in their final year of Preparatory Scientific Education, while GPT-4 outperformed the average student. However, GPT-4 was not flawless and made several mistakes on each exam. To further investigate, we explored if GPT-4 could self-identify the questions it failed. Initial attempts using specific prompts, such as '*Please rate the difficulty of the question*' did not yield meaningful insights.

Previous research showed that implementing a self-consistency strategy may prove beneficial in directing large language models towards accurate outputs (Wang et al., 2023; Zheng et al., 2023). This entails generating a number of candidate outputs, with the most frequent or internally consistent output being selected. In the present study, we attempted to further this idea by incorporating an element of stochasticity into the output, allowing us to gauge the level of confidence exhibited by GPT-4 with respect to its own outputs. Specifically, we discovered that employing the 'temperature' parameter in conjunction with multiple repetitions yielded valuable insights.

In our self-consistency analysis, we used a temperature parameter of 1. This decision was informed by presenting a single prompt a large number of times under dif-

Table 1 Performance of ChatGPT on Dutch national exams on the topic of 'English'

	Number of points				Corresponding grade (1 to 10)		
	GPT-3.5 (without re-prompting)	GPT-4-0314	GPT-4-0613	Maximum number of points attainable	GPT-3.5 (without re-prompting)	GPT-4-0314	GPT-4-0613
Exam 1 (May 2022, Period 1)	41 (36)	46	44	49	7.9 (7.0)	8.9	8.5
Exam 2 (June 2022, Period 2)	35 (30)	40	40	46	7.0 (6.1)	8.0	8.0
Exam 3 (July 2022, Period 3)	33 (30)	39	38	46	7.0 (6.4)	8.1	7.9
Total	109 (96)	125	122	141			
Average					7.3 (6.5)	8.3	8.1

Note. The prompting of GPT-3.5 was done through the web interface, while the prompting of GPT-4 was done using the API with a temperature setting equal to 0

ferent temperature settings. The particular prompt used (see Fig. 1) required ChatGPT to answer three questions, with the correct answers being C, B, and B, respectively. Figure 2 shows how the correctness of answers varied with different temperature settings, from the minimum value of 0 to the maximum possible value of 2.0, with increments of 0.1. It can be seen that with the temperature set at 0, Questions 36 and 38 were answered correctly, while Question 37 was answered incorrectly. As the temperature increased, Questions 36 and 38 consistently received correct answers, suggesting that ChatGPT had confidence in its answers. Nonetheless, at very high temperature settings, ChatGPT produced incorrect responses. Upon further investigation, we noted that these faulty answers were not inaccurate responses to the multiple-choice questions, but rather 'hallucinations' from ChatGPT, where it generated nonsensical text instead of an answer to the questions. In contrast, Question 37 elicited a level of uncertainty from ChatGPT, where, as the temperature rose, the correct answer surfaced approximately 20% of the time. Based on these observations, we made the decision to conduct a bootstrapping analysis for assessing self-consistency using a temperature setting of 1.0. This value ensures a degree of output variation, yet it restricts excessive variation that might cause ChatGPT to generate arbitrary texts.

After deciding upon the temperature setting of 1.0, we submitted each of the three exams to GPT-4, 50 times each. The number 50 is a trade-off, in which too few repetitions carry the risk that GPT-4 coincidently produces the same output multiple times in a row, appearing consistent when it actually is not. On the other hand, too many repetitions involve unnecessary use of computational resources and can give the false suggestion that GPT-4 is not consistent if it only very rarely produces an alternative output. The 50 repetitions were accomplished by setting the 'n' parameter in the API to 50.

The performance of GPT-4 for each exam question was then manually scored as above, and classified into three categories:

- Questions for which GPT-4 answered correctly in all 50 attempts, indicating high consistency.
- Questions for which GPT-4 provided the same incorrect response in all 50 attempts, indicating high consistency but a wrong answer. In the case of a question worth more than one point, not achieving the full points was also considered an incorrect response for the respective question.
- Questions for which GPT-4 provided at least two different responses over the 50 attempts, indicating inconsistency.

Our analysis showed that out of 124 questions across the three exams combined, GPT-4-0314 displayed inconsistency in 23 cases (19%), while GPT-4-0613 displayed inconsistency in 25 cases (20%) (see Table 2). Among these inconsistent responses, a significant portion (10 or 43% for GPT-4-0314; 14 or 56% for GPT-4-0613) were indeed incorrect. In contrast, of the 101 and 98 consistent responses for GPT-4-0314 and GPT-4-0613, only 4 and 4 were incorrect ("consistently incorrect").

The final grades on a scale from 1 to 10 were calculated by averaging the results over 50 repetitions, and then further averaging across three exams. GPT-4-0314 scored an average of 8.11 (Exams 1–3: 8.63, 7.80, 7.89), while GPT-4-0613 obtained an average score of 8.21 (Exams 1–3: 8.64, 7.97, 8.01). These results are comparable

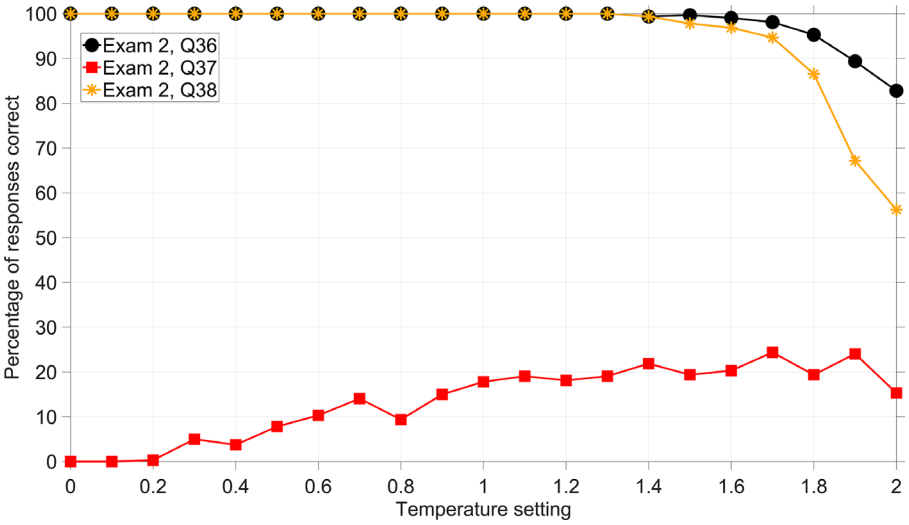

Fig. 2 Percentage of correct responses by submitting the prompt shown in Fig. 1 for a total of 320 times (model: GPT-4-0314). The procedure was repeated for 21 different temperature settings, from 0 to its maximum value of 2.0

Table 2 Number of exam questions per category based on the consistency of 50 GPT-4 attempts (GPT-4-0314 / GPT-4-0613)

	Exam 1	Exam 2	Exam 3	Total
Consistently correct (50 times correct)	36 / 33	31 / 31	30 / 30	97 / 94
Consistently incorrect (50 times the same incorrect response)	1 / 2	3 / 1	0 / 2	4 / 5
Inconsistent (1–49 times the correct response)	6 / 8	6 / 8	11 / 9	23 / 25
of which GPT-4 answered incorrectly at temperature = 0	2 / 4	2 / 5	6 / 5	10 / 14
of which GPT-4 answered correctly at temperature = 0	4 / 4	4 / 3	5 / 4	13 / 11
Total	**43**	**40**	**41**	**124**

to those presented in Table 1, where the mean score across the three exams was 8.3 and 8.1 for GPT-4-0314 and GPT-4-0613, respectively. In summary, the exam grades based on bootstrapping align with the original analysis from Table 1, where a temperature setting of 0 was used.

Figure 3 illustrates the performance of the two ChatGPT versions on Exam 1 relative to all students who completed this exam. The students' mean number of points was 35.88 ($SD = 6.48$) out of a maximum of 49, and the mean of their grades was 6.99. This analysis was conducted only for one of the three exams, as the students' results for the other two exams were not publicly available.

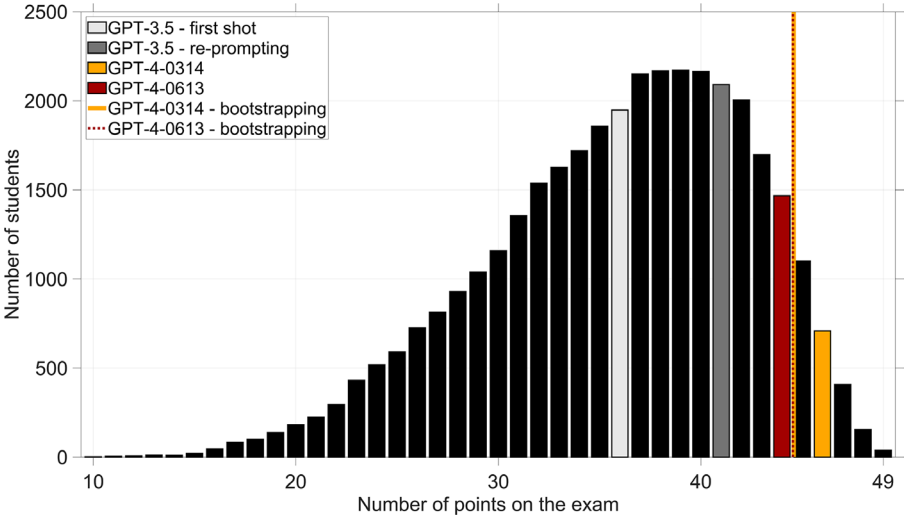

Fig. 3 Distribution of student scores on Exam 1 (n=35,698). The performance of GPT-3.5 (corresponding to the 46th and 76th percentiles), GPT-4 (97th and 91st percentiles), and bootstrapped GPT-4 (mean score of repetitions: 44.58 and 44.54) is depicted

Discussion

The present study's results indicate that GPT-3.5 performs comparably to, while GPT-4 significantly outperforms, the average Dutch student in the domain of English language comprehension. Although students are prohibited from using computers during conventional in-person examinations, our findings suggest that ChatGPT could compromise the integrity of computer-based exams, which have gained popularity in the wake of the COVID-19 pandemic (Kerrigan et al., 2022; Pettit et al., 2021). Educators may presume that online exams with minimal supervision are secure in subjects such as comprehension, where answers are unlikely to be readily accessible online. However, this assumption may no longer hold, given that our study demonstrates the generation of valid answers within minutes. Concurrently, there are concerns that ChatGPT could be exploited for cheating on assessments (Cotton et al., 2023; Mitchell, 2022), necessitating a reevaluation of current methods for assessing student knowledge. Potential solutions include increased proctoring, reduced reliance on essay-based work, and the utilization of alternative assignment formats, such as videos or presentations (Geerling et al., 2023; Graham, 2022; Rudolph et al., 2023; Susnjak, 2022).

On a positive note, ChatGPT holds the potential to foster innovation in the realm of education. Possible applications encompass aiding the development of writing skills, facilitating comprehension through step-by-step explanations, speeding up information delivery via summarization of texts, and enhancing engagement through personalized feedback (Kasneci et al., 2023; Rudolph et al., 2023; Šlapeta, 2023). It is worth considering whether the focus of student assessment should transition towards the effective utilization of ChatGPT and similar language models. For instance, it may be advisable to instruct students on identifying inaccuracies in content generated by

ChatGPT or on integrating ChatGPT to establish a synergistic combination of human and computational capabilities.

A noteworthy observation emerged when a bootstrapping method, consisting of 50 repetitions, was used to determine if GPT-4 exhibited uncertainty in its outputs. This approach has demonstrated its efficacy as a tool for self-assessment, where we found that about half of the responses labeled as 'inconsistent' were incorrect while only about 5% of the responses deemed 'consistent' were incorrect. The exploitation of randomness and bootstrapping has the potential to be a tool for future research. Here it is noteworthy that, although the self-consistency method in our case provided a way for ChatGPT to make statements about the certainty of the answer it delivered, this method did not prove to be useful for arriving at a more accurate answer, contrary to findings by Wang et al. (2023). A possible explanation is that the English exam questions did not involve chain-of-thought reasoning as in Wang et al. (2023), where diverse reasoning paths could lead to the same correct answer. In our case, ChatGPT had to directly converge on the correct answer (such as the letter 'A', 'B', 'C', or 'D'). This can explain why the answer that ChatGPT found most likely (i.e., at the temperature setting of 0) also had the highest probability of being correct, and adding variation by choosing a higher temperature provided no additional value in terms of the accuracy of the output.

Interestingly, the principle of self-consistency may also provide advantages in the construction of tests and exams. For Exam 1, psychometric properties were available online (CITO, 2022). An initial analysis revealed that, for the 'inconsistent' or incorrectly answered questions of Exam 1 ($n=7$ for GPT-4-0314, $n=9$ for GPT-4-0613, as shown in Table 2), students achieved an average score of 63.0% ($SD=10.5\%$) and 67.7% ($SD=12.0\%$), compared to 75.8% ($SD=12.7\%$) and 75.3% ($SD=13.2\%$) for the remaining questions ($n=36$, $n=34$). Thus, questions that ChatGPT answered inconsistently or incorrectly appeared to pose greater challenges to human examinees. Based on this insight, organizations that design exams could apply the principle of self-consistency to verify the level of difficulty of their exams, aiming to maintain consistency year on year. For instance, by having ChatGPT repeatedly attempt newly developed exams (with a temperature setting of 1) and assessing the mean consistency level across all items, one could obtain an initial measure of exam difficulty. However, the feasibility and efficacy of this approach would require further validation through additional research.

Chen et al. (2023) observed a substantial deterioration in the performance of the latest version of GPT-4 compared to its predecessor from a few months earlier. Contrary to this, our analysis found no notable difference between the two versions. Our results revealed an average exam grade of 8.3 for the March version and 8.1 for the June version when using a temperature setting of 0, while upon using a bootstrapping method involving 50 repetitions with a temperature setting of 1.0, the resultant grades were 8.11 and 8.21, respectively. A plausible explanation for the absence of noticeable disparity lies in the fact that the base model, GPT-4, remained the same, with updates, as far as is currently known, focusing exclusively on the fine-tuning layers. Chen et al. (2023) showed a degradation in the June version of GPT-4's capability to determine whether a given number is prime. It is important to note here that ChatGPT, in its essence, is incapable of running algorithms required to determine if a number is prime; nevertheless, through

a smart chain-of-prompt strategy, accurate responses can still be obtained. In a parallel finding, Chen et al. (2023) reported that the June version of GPT-4 exhibited a decreased willingness to respond to sensitive content questions compared to the March version. We also found that there were differences between the GPT-4-0314 and GPT-4-0613 outputs, even though the overall average was comparable. With a temperature setting of 0, GPT-4-0314 answered a total of 14 questions incorrectly (worth 16 points), while GPT-4-0613 incorrectly answered 18 questions (worth 19 points). However, only 9 of these questions were answered incorrectly by both models (see supplementary material for the complete overviews). There was even one question where the bootstrapping analysis showed that GPT-4-0314, in 49 out of 50 instances, gave the answer 'C', while GPT-4-0613, in all 50 instances, correctly gave the answer 'B'. We have no logical explanation for this variation, except that we have noticed that very minor changes in the prompt can have a large impact (see also Reiss, 2023), which could be traced back to a potential lack of robustness in the autoregressive modeling principle (LeCun, 2023). In summary, it seems that GPT-4's performance in English language comprehension has not changed over time, even though it is possible that specific subtasks are performed differently by each version.

One limitation of our study was that ChatGPT was applied to an English exam for Dutch-speaking students, which means that the level of the exam is not equivalent to a native language exam. The estimated level of earlier VWO English exams was reported to be C1, with a score of 64% as the minimum required to achieve this level (College voor Toetsen en Examens, 2020). Furthermore, even though our study was conducted on a well-constructed high school exam, it does not guarantee that ChatGPT will also perform well in other types of verbal tests.

With the introduction of GPT-4, large language models have reached a point where they can outperform average humans in certain areas. There is a high probability that ChatGPT or similar models will evolve to become more intelligent over time. Given that current models can already do well on high school exams today, it raises significant questions about what lies ahead in education and beyond. A wealth of opportunities are at reach for new applications and integrations, such as incorporation into, for example, Microsoft 365 Copilot (Office Microsoft Blog, 2023). As large language models continue to evolve, they have the potential to redefine the boundaries of human-computer interaction.

Acknowledgements Dr. Dimitra Dodou's role in scoring the output of ChatGPT according to the correction instruction is acknowledged.

Declarations The author has no relevant financial or non-financial interests to disclose.

References

Arora, D., & Singh, H. G. (2023). *Have LLMs advanced enough? A challenging problem solving benchmark for large Language Models.* arXiv. https://doi.org/10.48550/arXiv.2305.15074.

Bommarito, M. J., II, & Katz, D. M. (2022). *GPT takes the Bar Exam* arXiv. https://arxiv.org/abs/2212.14402.

Bordt, S., & Von Luxburg, U. (2023). *ChatGPT participates in a computer science exam* arXiv. https://arxiv.org/abs/2303.09461.

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S., Nori, H., Palangi, H., Ribeiro, M. T., & Zhang, Y. (2023). *Sparks of artificial general intelligence: Early experiments with GPT-4* arXiv. https://arxiv.org/abs/2303.12712.

Chen, L., Zaharia, M., & Zou, J. (2023). *How is ChatGPT's behavior changing over time?* arXiv. https://doi.org/10.48550/arXiv.2307.09009.

CITO (2023). CITO: toetsen, examens, volgsystemen, certificeringen en trainingen [CITO: tests, exams, tracking systems, certifications, and trainings]. https://cito.nl.

CITO (2022). Toets en item analyse VWO Engels 2022 tijdvak 1 [Test and item analysis VWO English 2022 period 1]. https://www2.cito.nl/vo/ex2022/VW-1002-a-22-1-TIA.docx.

College voor Toetsen en Examens (2020). Syllabus centraal examen 2022 Arabisch, Duits, Engels, Frans, Russisch, Spaans, Turks [Syllabus central exams 2022 Arabic, German, English, French, Russian, Spanish, Turkish]. https://havovwo.nl/pics/vmvtsyl22.pdf.

College voor Toetsen en Examens. (2022). Engels VWO 2022. https://www.examenblad.nl/examen/engels-vwo-2/2022.

Cotton, D. R. E., Cotton, P. A., & Shipway, J. R. (2023). Chatting and cheating: Ensuring academic integrity in the era of ChatGPT. *Innovations in Education and Teaching International*. https://doi.org/10.1080/14703297.2023.2190148.

Davis, J. C., Lu, Y. H., & Thiruvathukal, G. K. (2023). *Conversations with ChatGPT about C programming: An ongoing study*. Figshare. https://figshare.com/articles/preprint/Conversations_with_ChatGPT_about_C_Programming_An_Ongoing_Study/22257274.

Frieder, S., Pinchetti, L., Griffiths, R. R., Salvatori, T., Lukasiewicz, T., Petersen, P. C., Chevalier, A., & Berner, J. (2023). *Mathematical capabilities of ChatGPT.* arXiv. https://doi.org/10.48550/arXiv.2301.13867.

Geerling, W., Mateer, G. D., Wooten, J., & Damodaran, N. (2023). *ChatGPT has mastered the principles of economics: Now what?* SSRN. https://doi.org/10.2139/ssrn.4356034.

Gilson, A., Safranek, C., Huang, T., Socrates, V., Chi, L., Taylor, R. A., & Chartash, D. (2022). *How well does ChatGPT do when taking the medical licensing exams? The implications of large language models for medical education and knowledge assessment.* medRxiv. https://doi.org/10.1101/2022.12.23.22283901.

Graham, F. (2022). Daily briefing: Will ChatGPT kill the essay assignment? *Nature*. https://doi.org/10.1038/d41586-022-04437-2.

Han, Z., Battaglia, F., Udaiyar, A., Fooks, A., & Terlecky, S. R. (2023). *An explorative assessment of ChatGPT as an aid in medical education: Use it with caution.* medRxiv. https://doi.org/10.1101/2023.02.13.23285879.

Huang, F., Kwak, H., & An, J. (2023). Is ChatGPT better than human annotators? Potential and limitations of ChatGPT in explaining implicit hate speech. *Companion Proceedings of the ACM Web Conference*, Austin, TX, 294–297. https://doi.org/10.1145/3543873.3587368.

Kasneci, E., Seßler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., & Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences*, *103*, 102274. https://doi.org/10.1016/j.lindif.2023.102274.

Katz, D. M., Bommarito, M. J., Gao, S., & Arredondo, P. (2023). *GPT-4 passes the bar exam.* SSRN. https://doi.org/10.2139/ssrn.4389233.

Kerrigan, J., Cochran, G., Tabanli, S., Charnley, M., & Mulvey, S. (2022). Post-COVID changes to assessment practices: A case study of undergraduate STEM recitations. *Journal of Educational Technology Systems*, *51*, 192–201. https://doi.org/10.1177/00472395221118392.

Kim, N., Htut, P. M., Bowman, S. R., & Petty, J. (2022). *$(QA)^2$: Question answering with questionable assumptions.* ArXiv. https://arxiv.org/abs/2212.10003.

King, M. R. (2023). The future of AI in medicine: A perspective from a chatbot. *Annals of Biomedical Engineering*, *51*, 291–295. https://doi.org/10.1007/s10439-022-03121-w.

Kirmani, A. R. (2023). Artificial Intelligence-enabled science poetry. *ACS Energy Letters*, *8*, 574–576. https://doi.org/10.1021/acsenergylett.2c02758.

Kortemeyer, G. (2023). Could an artificial-intelligence agent pass an introductory physics course? *Physical Review Physics Education Research*, *19*, 010132. https://doi.org/10.1103/PhysRevPhysEducRes.19.010132.

Kosinski, M. (2023). *Theory of mind may have spontaneously emerged in large language models.* arXiv. https://doi.org/10.48550/arXiv.2302.02083.

Kung, T. H., Cheatham, M., Medenilla, A., Sillos, C., De Leon, L., Elepaño, C., Madriaga, M., Aggabao, R., Diaz-Candido, G., Maningo, J., & Tseng, V. (2023). Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. *PLOS Digital Health*, *2*, e0000198. https://doi.org/10.1371/journal.pdig.0000198.

Kuzman, T., Ljubešić, N., & Mozetič, I. (2023). *ChatGPT: Beginning of an end of manual annotation? Use case of automatic genre identification.* arXiv. https://arxiv.org/abs/2303.03953.

LeCun, Y. (2023). Do large language models need sensory grounding for meaning and understanding? Spoiler: YES! [Presentation]. https://drive.google.com/file/d/1BU5bV3X5w65DwSMapKcsr0ZvrMRU_Nbi/view.

Lovin, B. (2022, December 3). ChatGPT produces made-up nonexistent references. https://brianlovin.com/hn/33841672.

Mitchell, A. (2022, December 26). Professor catches student cheating with ChatGPT: 'I feel abject terror'. https://nypost.com/2022/12/26/students-using-chatgpt-to-cheat-professor-warns.

Newton, P. M., & Xiromeriti, M. (2023). *ChatGPT performance on MCQ-based exams*. EdArXiv. https://doi.org/10.35542/osf.io/sytu3.

Office Microsoft Blog (2023). Introducing Microsoft 365 Copilot – your copilot for work. https://blogs.microsoft.com/blog/2023/03/16/introducing-microsoft-365-copilot-your-copilot-for-work.

OpenAI (2023). GPT-4 technical report. https://cdn.openai.com/papers/gpt-4.pdf.

Pettit, M., Shukla, S., Zhang, J., Sunil Kumar, K. H., & Khanduja, V. (2021). Virtual exams: Has COVID-19 provided the impetus to change assessment methods in medicine? *Bone & Joint Open*, *2*, 111–118. https://doi.org/10.1302/2633-1462.22.BJO-2020-0142.R1.

Reiss, M. V. (2023). *Testing the reliability of ChatGPT for text annotation and classification: A cautionary remark.* arXiv. https://doi.org/10.48550/arXiv.2304.11085.

Rospocher, M., & Eksir, S. (2023). Assessing fine-grained explicitness of song lyrics. *Information*, *14*, 159. https://doi.org/10.3390/info14030159.

Rudolph, J., Tan, S., & Tan, S. (2023). ChatGPT: Bullshit spewer or the end of traditional assessments in higher education? *Journal of Applied Learning and Teaching*, *6*. https://doi.org/10.37074/jalt.2023.6.1.9.

Savelka, J., Agarwal, A., An, M., Bogart, C., & Sakr, M. (2023). *Thrilled by your progress! Large Language Models (GPT-4) no longer struggle to pass assessments in higher education programming courses*. arXiv. https://doi.org/10.48550/arXiv.2306.10073.

Šlapeta, J. (2023). Are ChatGPT and other pretrained language models good parasitologists? *Trends in Parasitology*. https://doi.org/10.1016/j.pt.2023.02.006.

Sobania, D., Briesch, M., Hanna, C., & Petke, J. (2023). *An analysis of the automatic bug fixing performance of ChatGPT.* arXiv. https://doi.org/10.48550/arXiv.2301.08653.

Susnjak, T. (2022). *ChatGPT: The end of online exam integrity?* arXiv. https://arxiv.org/abs/2212.09292.

Tabone, W., & De Winter, J. (2023). Using ChatGPT for human–computer interaction research: A primer. *Royal Society Open Science*, *10*, 231053. https://doi.org/10.1098/rsos.231053

Vincent, J. (2022, December 5). AI-generated answers temporarily banned on coding Q&A site Stack Overflow. https://www.theverge.com/2022/12/5/23493932/chatgpt-ai-generated-answers-temporarily-banned-stack-overflow-llms-dangers.

Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., Chowdhery, A., & Zhou, D. (2023). *Self-consistency improves chain of thought reasoning in language models. Proceedings of the International Conference on Learning Representations*, Kigali, Rwanda. https://doi.org/10.48550/arXiv.2203.11171.

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., Chi, E. H., Hashimoto, T., Vinyals, O., Liang, P., Dean, J., & Fedus, W. (2022). *Emergent abilities of large language models.* arXiv. https://doi.org/10.48550/arXiv.2206.07682.

Whitford, E. (2022, December 9). A computer can now write your college essay — Maybe better than you can. https://www.forbes.com/sites/emmawhitford/2022/12/09/a-computer-can-now-write-your-college-essay---maybe-better-than-you-can/?sh=35deca9ddd39.

Zhai, X. (2022). *ChatGPT user experience: Implications for education*. ResearchGate. https://www.researchgate.net/publication/366463233_ChatGPT_User_Experience_Implications_for_Education.

Zheng, C., Liu, Z., Xie, E., Li, Z., & Li, Y. (2023). *Progressive-hint prompting improves reasoning in large language models.* arXiv. https://doi.org/10.48550/arXiv.2304.09797.

Zhong, Q., Ding, L., Liu, J., Du, B., & Tao, D. (2023a). *Can ChatGPT understand too? A comparative study on ChatGPT and fine-tuned BERT.* arXiv. https://doi.org/10.48550/arXiv.2302.10198.

Zhong, W., Cui, R., Guo, Y., Liang, Y., Lu, S., Wang, Y., Saied, A., Chen, W., & Duan, N. (2023b). *AGIEval: A human-centric benchmark for evaluating foundation models.* arXiv. https://doi.org/10.48550/arXiv.2304.06364.

Supplementary Information All inputs (prompts) and outputs of ChatGPT, as well as the MATLAB scripts used to access the API and create the figures can be found here: https://doi.org/10.4121/545f8ead-235a-4eb6-8f32-aebb030dbbad.