Computers and Education: Artificial Intelligence 7 (2024) 100248

Contents lists available at ScienceDirect

Computers and Education: Artificial Intelligence

journal homepage: www.sciencedirect.com/journal/computers-and-education-artificial-intelligence

Assessing the quality of automatic-generated short answers using GPT-4

Luiz Rodrigues [a,*], Filipe Dwan Pereira [b], Luciano Cabral [c], Dragan Gašević [d], Geber Ramalho [d,e], Rafael Ferreira Mello [f,**]

[a] Computing Institute, Federal University of Alagoas, Macei'o, Alagoas, Brazil
[b] Department of Computer Science, Federal University of Roraima, Boa Vista, Brazil
[c] Federal Institute of Pernambuco, Jaboat–ao Dos Guararapes, Brazil
[d] Faculty of Information Technology, Monash University, Melbourne, Australia
[e] Federal University of Pernambuco, Recife, Pernambuco, Brazil
[f] CESAR School, Centro de Estudos e Sistemas Avancados Do Recife, Recife, Brazil

ARTICLE INFO

Keywords:
Automatic answer generation
Question-answering
Large language models
GPT-4
Natural language processing

ABSTRACT

Open-ended assessments play a pivotal role in enabling instructors to evaluate student knowledge acquisition and provide constructive feedback. Integrating large language models (LLMs) such as GPT-4 in educational settings presents a transformative opportunity for assessment methodologies. However, existing literature on LLMs addressing open-ended questions lacks breadth, relying on limited data or overlooking question difficulty levels. This study evaluates GPT-4's proficiency in responding to open-ended questions spanning diverse topics and cognitive complexities in comparison to human responses. To facilitate this assessment, we generated a dataset of 738 open-ended questions across Biology, Earth Sciences, and Physics and systematically categorized it based on Bloom's Taxonomy. Each question included eight human-generated responses and two from GPT-4. The outcomes indicate GPT-4's superior performance over humans, encompassing both native and non-native speakers, irrespective of gender. Nevertheless, this advantage was not sustained in 'remembering' or 'creating' questions aligned with Bloom's Taxonomy. These results highlight GPT-4's potential for underpinning advanced question-answering systems, its promising role in supporting non-native speakers, and its capacity to augment teacher assistance in assessments. However, limitations in nuanced argumentation and creativity underscore areas necessitating refinement in these models, guiding future research toward bolstering pedagogical support.

1. Introduction

Automatic Answer Generation (AAG) offers significant value in Educational Technology by enhancing learning experiences and facilitating personalized feedback. Through AAG, educational platforms can generate accurate responses to (any) student queries, fostering self-directed learning and immediate clarification of concepts (Wang & Demszky, 2023; Yan et al., 2023). This automation supports the process of providing tailored assessments, reducing teachers' workload while ensuring timely feedback for students. Moreover, AAG enables adaptive learning environments, where algorithms analyze student responses to tailor subsequent questions, addressing individual learning gaps effectively (Divya et al., 2023; Wang & Demszky, 2023). Relevant has explored AAG systems, however, the literature has highlighted the need for advancing educational technology by equipping them with innovative solutions such as state-of-the-art large language models (LLMs) (Yan et al., 2023).

The recent advancements in LLMs, such as GPT and Palm-2, have provoked the research community to study their implications within educational settings (Yan et al., 2023; Zirar, 2023). These models can be used in a wide range of applications, such as powering question-answering systems and interactive chatbots. Previous studies have evaluated the efficacy of LLMs, predominantly those based on the GPT models, in automating response generation (Divya et al., 2023; Wang & Demszky, 2023; Yan et al., 2023). These investigations have primarily concentrated on the models' capabilities in addressing multiple-choice (Huang et al., 2019; Rosoł et al., 2023) and open-ended questions, the latter often accompanied by contextual reading material

* Corresponding author.
** Corresponding author.
E-mail addresses: luiz.rodrigues@nees.ufal.br (L. Rodrigues), filipedwan@gmail.com (F. Dwan Pereira), rflm@cesar.org.br (R. Ferreira Mello).

https://doi.org/10.1016/j.caeai.2024.100248
Received 13 March 2024; Received in revised form 27 May 2024; Accepted 3 June 2024
Available online 29 June 2024

(de Winter 2023; Li et al., 2023; Liu et al., 2019).

On the other hand, the existing literature on the application of LLMs to answer open-ended questions without supporting texts is limited, whereas this task is crucial for AAG-based systems. For instance, this capability is vital for generative meaningful, personalized feedback to students even when they present solutions/question unexpected by the system's designer, which is important to promote critical thinking and problem-solving skills by supporting students' own responses (Wang & Demszky, 2023; Yan et al., 2023). Additionally, it enables the system to adapt its teaching strategies dynamically, catering to the individual learning pace and style of each student (Divya et al., 2023; Wang & Demszky, 2023). Therefore, exploring and advancing the application of LLMs in this context can significantly enhance the effectiveness of AAG systems in fostering comprehensive and tailored educational experiences.

However, past research concerning LLMs and AAG without supporting texts often proposes studies using small data samples, focusing on narrow domains, or not considering the posed questions' varying difficulty levels (Jayaraman & Black, 2022; Nguyen et al., 2023; Parsons & Curry, 2023). The relevance from broadening the context of those studies might be addressed from two main perspectives. First, if one is to rely on LLMs' answers, it is important to understand how they compare to a reliable source. In this perspective, we often rely on humans' answers (e.g., peers and teachers), and one's answers might vary depending on their language understanding. Consequently, it is important to understand how LLMs answers compare to humans, and whether being a native speaker or not plays a role, given that language understanding is a key feature enabling question answering OpenAI (2023). Additionally, as AI research has faced bias issues, ensuring LLMs abilities do not reflect societal inequalities related to gender is prominent so that this technology does not perpetuate these biases (Memarian & Doleck, 2023; Woolf, 2022). Second, it is important to not only understand LLMs' AAG overall capabilities, but how they vary depending on the questions' features, such as the topic they address and its complexity. This is important so that educators and students are aware of situations in which LLMs are more or less likely to reliably answer their queries, hence, providing an understanding of how LLMs support learners with different needs Pedro et al. (2019); Vinuesa et al. (2020).

Therefore, this study aims to assess the capability of GPT-4 to respond to open-ended questions applicable to high-school subject matter. To address the challenges previously outlined, our research involved the compilation of a comprehensive dataset comprising 7380 responses to 738 open-ended questions across three distinct disciplines: Biology, Earth Sciences, and Physics. Particularly, those disciplines were selected due to their representativeness in early school years. Hence, as they concern questions likely to be asked and answered by students and teachers, they provide valuable external validity for our dataset. This dataset uniquely includes eight human-generated answers alongside two responses generated by GPT-4 for each question. Furthermore, the questions are systematically categorized according to the levels of cognitive demand, as defined by Bloom's Taxonomy (Anderson & Sosniak, 1994).

The inclusion of Bloom's Taxonomy as a factor serves as a robust framework for evaluating GPT-4's performance. While there are many models and theories about cognitive complexities, given Bloom's widespread acceptance and usage in teacher education, relying on its hierarchical structure provides a clear and standardized way to categorize the cognitive demands of the open-ended questions Rawat et al. (2023). For instance, the revision of Bloom's taxonomy presents six complexity levels that include remembering (e.g., recalling historical dates), understanding (e.g., explaining the causes of an economic recession), applying (e.g., applying mathematical formulas to solve real-world problems), analyzing (e.g., analyzing a literary work's themes and character development), evaluating (e.g., evaluating the effectiveness of a scientific experiment), and creating (e.g., writing an original short story) Anderson and Sosniak (1994).

These levels offer a nuanced understanding of cognitive processes, allowing for a comprehensive assessment of GPT-4's capabilities across various intellectual tasks Rawat et al. (2023). The hierarchy allows researchers to understand where models excel/require further developments, facilitating targeted improvements in performance and usability to attend varied learning needs Pedro et al. (2019); Vinuesa et al. (2020). Therefore, this not only enhances the precision of our evaluation but allows for a comparative analysis across different levels of question complexity. Thereby, by aligning our dataset with Bloom's taxonomy, we can systematically assess how well GPT-4 responds to varying cognitive challenges, making our study methodologically sound and insightful for educational applications.

Our initial results show that GPT-4 outperformed humans in our dataset. Overall, GPT-4's scores were higher than those of both native and non-native speakers. Interestingly, GPT-4's advantage compared to the former was higher than that of the latter, while people's gender did not affect this difference. Nevertheless, when we controlled for the question's complexity, based on the levels of Bloom's taxonomy, we found that GPT-4 did not outperform humans in either *remembering* or *creating* questions. Therefore, our findings suggest that, compared to humans, GPT-4 is able to provide more accurate answers to open-ended questions of all complexity levels, except factual and creative-based ones.

2. Background

2.1. LLM models

Large language models (LLM) have experienced unprecedented and rapid advancement in development and application. These advancements have created more powerful models, enabling them to perform a broader range of NLP tasks with accuracy and fluency (Yenduri et al., 2023). Among these models, the *encoder-only* architecture, used by BERT, DEBERTa, and RoBERTa, reaches good performance in capturing contextual information for tasks like language understanding. *Decoder-only* models, typified by GPT-2 and PALM, focus on generating coherent and contextually rich text, showcasing their proficiency in creative language generation. Additionally, *sequence-to-sequence* models like T5 contribute to tasks requiring input-to-output mapping, showcasing their versatility in handling diverse applications (Devlin et al., 2019; He et al., 2020; Lewis et al., 2020; Radford et al., 2019; Raffel et al., 2023).

LLMs are built upon the transformer architecture (Vaswani et al., 2017), utilizing self-attention mechanisms to capture contextual dependencies within input sequences. The scale of these models can be exemplified by GPT-3, a model with approximately 175 billion parameters (Brown et al., 2020). The pre-training process involves exposure to massive corpora, allowing the model to learn nuanced contextual language representations.

Initially, these models were limited to executing a singular task effectively only when fine-tuned with many examples, compromising their generalization capabilities as language models (Brown et al., 2020). However, recent studies unveiled the potential of language models to seamlessly adapt to various tasks without the need to train new models (Raffel et al., 2023). Therefore enabling knowledge transfer between diverse activities (Alzubaidi et al., 2023; Brown et al., 2020). For instance, GPT-3 has demonstrated interesting few-shot learning capabilities (Brown et al., 2020). This implies the model's ability to perform tasks with minimal task-specific examples.

Despite this versatility, the computational expense of training for each new task and the frequent unavailability of appropriately annotated datasets may still present limitations and challenges.

2.2. LLM for answering open-ended questions

The analysis, understanding, and response to open-ended questions

present a relevant challenge for the scientific community, demanding diverse models capable of addressing these tasks contextually. In this context, LLM could be a relevant support in this domain, as it addresses both comprehension and text generation activities (Vaswani et al., 2017).

The efficacy of language models in open-ended question analysis and answering is prominently demonstrated. Recent empirical studies have systematically assessed the efficacy of LLMs, specifically those built on GPT architectures, in the realm of automated response generation (Divya et al., 2023; Wang & Demszky, 2023; Yan et al., 2023). For instance, Divya et al. (2023) comprehensively compare seven pre-trained embedding models. This analysis is centered on assessing the degree of similarity between these models and student responses. To this end, the researchers employed regression models with the objective of forecasting scores for short-answer questions. This predictive modeling was conducted within the framework of the Mohler dataset. The evaluation methodology adopted in this study used RMSE and Pearson correlation coefficients for each of the embedding models evaluated.

The study by Wang and Demszky (2023), they assessed whether generative AI could effectively complement expert feedback as an automated teacher coach. Three coaching tasks were proposed: (A) scoring transcript segments, (B) identifying instructional highlights and missed opportunities, and (C) providing actionable suggestions for student reasoning. Expert math teachers evaluated ChatGPT's zero-shot performance on these tasks using elementary math transcripts. The results showed that ChatGPT's insights were relevant but lacked novelty, with 82% of suggestions aligning with existing teacher actions in the transcript.

In Yan et al. (2023), a systematic scoping review of 118 peer-reviewed papers (published since 2017) explored the current state of research on employing Language Models (LLMs) for automating educational tasks. The findings unveiled 53 use cases across nine categories: profiling/labelling, detection, grading, teaching support, prediction, knowledge representation, feedback, content generation, and recommendation. Practical and ethical challenges, such as low technological readiness, replicability issues, and privacy concerns, were identified. Three key recommendations for future studies emerged: updating innovations with state-of-the-art models (e.g., GPT-3/4), advocating for open-sourcing models/systems, and prioritizing a human-centered approach in the developmental process.

Similarly, the study conducted by Rosoł et al. (2023) offers an in-depth assessment of the efficacy of two LLMs, ChatGPT (based on GPT-3.5) and GPT-4. This evaluation focused on applying these models in the Polish Medical Final Examination (MFE) context, utilizing different temperature settings for each model. The study encompassed three distinct editions of the MFE: Spring 2022, Autumn 2022; Spring 2023, with the examinations presented in both English and Polish. This comparative analysis examined both models' accuracies and the relationship between the correctness of the answers and various answer metrics. The main finding of this research was the consistent superiority of GPT-4 over GPT-3.5 in all three examined iterations of the MFE, irrespective of the language of the examination.

The study by Liu et al. (2019) is key in K-12 education, as it pioneers developing a hybrid automatic question-answering system. This innovative system is a confluence of Knowledge-Based Question Answering (KB-QA) and Information Retrieval-based Question Answering (IR-QA) methodologies. It leverages resources from Chinese textbooks and employs a comprehensive K-12 knowledge graph, accessible at edukg.org. The system's extensive scope covers nine diverse subjects, including mathematics, Chinese language, geography, and history. A critical aspect of this research is its empirical evaluation, which involved a rigorous testing process using over 9000 questions. The results of this evaluation are remarkable, with the system achieving an average accuracy rate exceeding 70%. This high level of accuracy demonstrates the hybrid approach's effectiveness in processing and answering educational questions and underscores its potential utility in enhancing learning experiences and outcomes in the K-12 educational sector.

In de Winter (2023), they explored the capabilities of GPT-3.5 in completing Dutch national high school exams, explicitly focusing on English reading comprehension, conducted in late December 2022. The research revealed that ChatGPT attained an average score of 7.3, closely mirroring the national student average in the Netherlands, which stood at 6.99. Notably, the study observed that while re-prompting occasionally enhanced response clarity, ChatGPT's initial average score was 6.5 without such interventions. The performance of the GPT-4 was also assessed. It achieved a higher average score of 8.3, which was accomplished without needing re-prompting. The study employed a novel bootstrapping method, leveraging ChatGPT's 'temperature' parameter to identify potential inaccuracies in the model's responses. A subsequent reassessment in June 2023, utilizing the updated version of GPT-4, indicated no significant deviation from the previously recorded overall score.

Li et al. (2023) performed a study that identified nine distinct prompting strategies tailored explicitly for ChatGPT. This approach was aimed at eliciting a wide range of reflective responses. These generated responses and reflections penned by students were subjected to a thorough evaluation process conducted by experienced educators. The evaluative framework employed was a theory-aligned assessment rubric meticulously designed for application in university-level pharmacy courses. This study explores the efficacy of Deep Learning classification techniques, particularly using BERT, in the automatic differentiation between student-authored and ChatGPT-produced reflective responses. The findings of this investigation are significant: ChatGPT was shown to possess the capacity to generate reflective responses of superior quality, applicable to various aspects of pharmacy education. The reflections generated by ChatGPT outperformed those written by students across all six criteria of the assessment rubric. Moreover, the study highlights the ability of a domain-specific BERT-based classifier, registering up to a 38% enhancement in four key accuracy metrics, surpassing both the evaluations conducted by teaching staff and those performed using a general-domain classifier.

Although previous works have evaluated multiple dimensions of this problem, studies in this domain either rely on small datasets, concentrate on specific domains, or overlook variations in question difficulty (Jayaraman & Black, 2022; Nguyen et al., 2023; Parsons & Curry, 2023). To the best of our knowledge, no previous study has evaluated the performance of GPT-4 in a dataset encompassing all these features.

2.3. Research questions

Based on the literature, we aim to address the lack of studies on answer generation using LLMs, addressing the problems of dataset size, content-focus approaches and measurement of the algorithm performance for different difficult levels. To this end, we evaluated the performance of GPT-4 in a newly created dataset, including diverse features. Thus, the first research question in the current study was:

Research Question 1 (RQ1):

To what extent is GPT-4 capable of answering open-ended questions?

The second goal of our study was to understand how individual features affect the capability of GPT-4 to answer the questions correctly. To this end, we utilized the Hierarchical Linear Modeling (Hox et al., 2010) to measure the performance of GPT-4 under different features. More formally, we asked the following research questions to guide our investigation at this stage:

Research Question 2 (RQ2):

Does the difference between GPT-4 and humans in answering open short questions change depending on the demographic information?

Research Question 3 (RQ3):

Does the difference between GPT-4 and humans in answering open short questions change depending on the characteristics of the questions?

3. Method

3.1. Dataset

As mentioned before, the limitations in prior research are primarily related to the dataset's limitations. These datasets are typically characterized by their small size, narrow domain focus, and an omission of the questions' difficulty level in the analytical process (Jayaraman & Black, 2022; Nguyen et al., 2023; Parsons & Curry, 2023). To address these shortcomings, we have created a comprehensive dataset comprising 7380 responses to 738 open-ended questions, spanning three distinct academic fields: Biology, Earth Sciences, and Physics. This dataset was created in English in order to improve the adoption for future studies.

The construction of our dataset was facilitated through a crowdsourcing platform, employing an approach similar to that used in earlier studies (Basu et al., 2013; Horbach et al., 2018). Our process was delineated into three distinct phases. The first phase required crowd workers to formulate a question ("Generated Question") that corresponded to the different Bloom's Taxonomy levels and was tailored to specific subjects ("topic"). These Generated Questions contained between 5 and 20 words each. In sum, 738 questions were created and evenly distributed across each Bloom's Taxonomy level, resulting in 246 questions per level. This approach ensured a balanced representation of cognitive complexity and topic specificity in our dataset.

During the second phase of our methodology, we produced the answers for each question. This step involved a dual approach. Firstly, we enrolled crowd workers to create eight responses per question. These responses were stratified based on two criteria: language proficiency (native and non-native speakers) and gender (male and female). Secondly, we utilized the GPT-4 model to produce two additional answers for each question automatically. Thus, for every question in our dataset, we compiled a total of ten answers, two each from the following categories: native female speakers, non-native female speakers, native male speakers, non-native male speakers, and answers generated by GPT-4.

In the final phase of our process, crowd workers were tasked with evaluating the answers using a grading scale ranging from 0 to 5, guided by specific criteria: a score of 5 signified an 'excellent' answer, 4 indicated a 'very good' answer, 3 was awarded to a 'good' answer, 2 denoted an 'acceptable but somewhat simplistic and lacking in detail' answer, 1 was given to a 'slightly unclear' answer, and 0 was reserved for answers that were 'incorrect or do not match the question'. During this assessment, the evaluators were instructed to consider three key aspects: the completeness of the content, the stylistic presentation, and the quality of argumentation. This structured evaluation process was designed to ensure a comprehensive assessment of each response, thereby enhancing the reliability and utility of our dataset.

Table 1 presents descriptive statistics of our dataset. In short, it has 7380 rows distributed among three topics: Biology, Earth Science and Physics. Similarly, the dataset is evenly distributed among the six levels of Bloom's taxonomy (1230 for each). For gender and speaker, the dataset is evenly distributed between male and female and native and non-native, respectively, with 40% of it for each one, while the remaining 20% concerns GPT-4 answers. When grouped together (i.e., gp_sp), there are 20% of the dataset for each combination. Furthermore, Fig. 1 demonstrates the average length of the answers in our dataset, given that it was a key component of our prompt to GPT-4 (see Section 3.2).

To further exemplify our dataset, Table 2 presents a sample of its answers along with the respective grades. Particularly, this table presents questions from three complexity levels (i.e., Remembering, Analyzing, and Creating) with answers from non-native, native speakers, and GPT4 for each of them, along with their respective score. Hence, we expect it to give an initial idea of our dataset. Furthermore, we highlight that this dataset will be openly available following this manuscript's acceptance for widespread usage.

Table 1
Our dataset's descriptive statistics.

Variable	Count (%)
Topic	
Biology	2460 (33%)
Earth Science	2460 (33%)
Physics	2460 (33%)
Bloom's level	
Remembering	1230 (16.7%)
Understanding	1230 (16.7%)
Applying	1230 (16.7%)
Analyzing	1230 (16.7%)
Evaluating	1230 (16.7%)
Creating	1230 (16.7%)
Gender	
GPT-4	1476 (20%)
Male	2952 (40%)
Female	2952 (40%)
Speaker	
GPT-4	1476 (20%)
Male	2952 (40%)
Female	2952 (40%)
GD_SP	
GPT-4	1476 (20%)
Male-native	1476 (20%)
Male-non-native	1476 (20%)
Female-native	1476 (20%)
Female-non-native	1476 (20%)

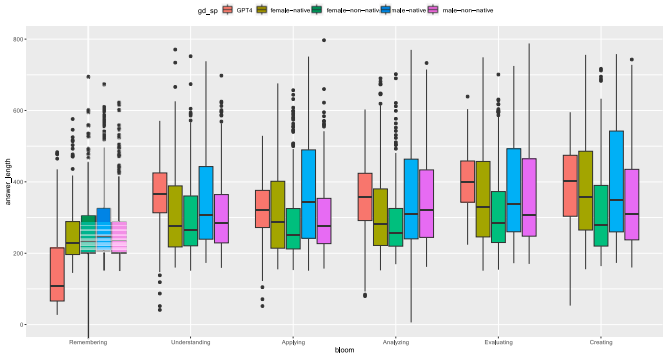

Fig. 1. Average length of the answers in our dataset.

3.2. GPT model and prompt

The GPT models have become prominent in academic discourse for addressing multiple challenges (Kasneci et al., 2023; Ziyu et al., 2023). Language models such as GPT-3.5 have demonstrated their capability to learn new tasks across various prompt paradigms, including zero-shot learning (without any examples), few-shot learning (using a limited number of examples), and in-context learning (utilizing contextual cues up to the model's input boundary) (Brown et al., 2020). For instance, GPT-3.5 outperformed the then-leading T5 model in specific tasks, such as question-answering, as demonstrated using the TriviaQA dataset (Brown et al., 2020).

In 2023, OpenAI introduced GPT-4, a more advanced language model, which reached better results than GPT-3.5 in many natural language processing (NLP) tasks (OpenAI, 2023). For example, GPT-4 has achieved top-tier performance, ranking in the upper 10% in professional examinations, including the Uniform Bar Exam in the United States and various academic tests in physics and psychology (OpenAI, 2023). However, it is essential to highlight that GPT-4 continues to have some

Table 2
Dataset sample.

Question	Answer	Speaker	Score
Remembering sample			
What are the characteristics of a mammal?	Mammals have seven specific traits: They are warm-blooded, have backbones and four limbs. Another specific is their fur and they produce milk to feed their children. They can live in different types of environments.	non-native	3
What are the characteristics of a mammal?	Mammals are unique beings. They possess the ability to dispense milk to their offspring. Mammals are covered in hair or fur for protection from others and the weather. Their heart organ is unique to other animals because it has four chambers. Unlike other animals, they are warm blooded. They also possess sweat glands. Breathing is easier for mammals because they have lungs.	native	3
What are the characteristics of a mammal?	Mammals are characterized by their ability to produce milk through mammary glands, have hair or fur, give live birth (with few exceptions), maintain a constant body temperature, and possess a neocortex region in their brain. They also have three middle ear bones and a lower jaw made up of a single bone.	GPT4	4
Analyzing sample			
Explain the genetic contribution to behavioral or mental health disorders such as bipolar disorder and schizophrenia.	Genetic factors play an important role in the development of behavioral and psychiatric disorders, such as bipolar disorder and schizophrenia. Multiple genes are involved, each with a small effect. Mutations in these genes can impair brain development and function, making individuals more susceptible to these disorders. However, genetic factors alone are not sufficient; environmental factors also play an important role in the development of these disorders.	non-native	4
Explain the genetic contribution to behavioral or mental health disorders such as bipolar disorder and schizophrenia.	Mental disorders are the result of genetic and environmental factors. Genetic disorders that can contribute to mental disorders are Espigenetic Regulation, how a person reacts to environmental factors. Genetic Polymorphisms, changes in our DNA and environmental factor combinations. Single gene changes, which is rare.	native	5
Explain the genetic contribution to behavioral or mental health disorders such as bipolar disorder and schizophrenia.	Genetic contribution to mental health disorders like bipolar disorder and schizophrenia is significant. Research shows these disorders often run in families, suggesting a genetic link. Specific genes associated with these conditions have been identified. However, having these genes doesn't guarantee development of these disorders; environmental factors also play a key role. Thus, it's an interplay of genetic predisposition and environmental triggers.	GPT4	2
Creating sample			
How might humans go about unlocking the full potential of their brains?	I totally recommend not even thinking about that. Just look at what a few percent of our use causes to the planet and other human beings. Any more usage will reveal more of the same, may it be 3 or bad.	non-native	0
How might humans go about unlocking the full potential of their brains?	Humans can tap into their brains' untapped potential through different things such exercising which is believed to help the brain work better. Learning new things and challenging your mind is another way as well as taking care or our overall health by eating well, sleeping enough and reducing stress. In addition to this brain training games or puzzles are also effective in keeping your mind sharp.	native	5
How might humans go about unlocking the full potential of their brains?	Humans can unlock the full potential of their brains through continuous learning, maintaining a healthy lifestyle, and practicing mindfulness. Brain exercises like puzzles, reading, or learning a new language can stimulate brain activity. Regular physical exercise, a balanced diet, enough sleep, and avoiding harmful substances contribute to overall brain health. Mindfulness practices, like meditation, can increase focus and cognitive flexibility.	GPT4	5

limitations, such as the tendency for hallucinations, context window constraints, and limitations in learning from extended historical contexts (OpenAI, 2023). Given the capability of using the GPT model for diverse tasks, we have incorporated GPT-4 in our experimental framework.

A critical aspect to consider in utilizing large language models is the development of well-structured prompts. This process significantly influences the model's ability to generate relevant and accurate outcomes (White et al., 2023). In our context, we developed the prophet based on the instructions used during the second phase of the dataset creation; it included aspects related to the criteria for evaluating the answer and the output format. Moreover, we also included the 'think step by step' command as it improves the outcome of GPT-4 (Kojima et al., 2022). Table 3 presents the final prompt for this task.

Table 3
Prompt proposed.

Element	Text
Instruction	Think step by step to answer the question.
Criteria	A good output should be coherent, include the main concepts related to the question, and present a clear argumentation in the response.
Output format	The answer should have up to 100 words in length, and should be in English only.
Data input:	Question: [question] Answer:

3.3. Evaluation methodology

As mentioned in section 3.1, the proposed dataset presents a hierarchical structure with answers nested within questions, questions within Bloom's taxonomy levels, and Bloom's taxonomy levels within Topics. This nesting violates the assumption of independence among observations, a fundamental assumption in traditional regression analysis (Gelman & Hill, 2006). Therefore, to address this hierarchical nature and the non-independence of data points, we employed a Hierarchical Linear Modeling (HLM), also known as Multilevel Regression Analysis (Hox et al., 2010). By using HLM, we addressed this hierarchical nature and the non-independence of data points, ensuring our data analysis' validity.

HLM is particularly suited for analyzing nested data structures as it allows for incorporating random effects at different hierarchy levels (Hox et al., 2010). By accounting for the variability within and between the nested levels, HLM enables us to model the complex relationships among variables while appropriately handling the hierarchical dependencies within the dataset (Gelman & Hill, 2006). In our study, HLM serves as a robust statistical technique to examine the effects of predictors at various levels of the hierarchy while controlling for the nested structure of the data. This approach provides insights into how specific questions influence answers, how questions relate to different levels of Bloom's taxonomy, and how Bloom's taxonomy levels manifest within various topics. Thus, HLM helps in ensuring the validity and reliability of our data analysis by accounting for the variability within and between the nested levels, thereby modeling the complex relationships among variables while appropriately handling the hierarchical dependencies within the dataset (Gelman & Hill, 2006).

In that context, understanding fixed and random effects is crucial for interpreting the model's parameters and capturing the variability within hierarchical data structures. Fixed effects concern the overall relationship between predictors and the outcome variable, providing information about the average impact across the entire dataset. Meanwhile, random effects account for the variability in this relationship at different levels of the hierarchy, acknowledging that the effects of predictors may vary between or within different groupings in the dataset (Hox et al., 2010). By simultaneously considering both fixed and random effects, HLM offers a robust framework to understand the complex relationships within hierarchical data and investigate the effects of predictors while accounting for the inherent dependencies in the data structure (Gelman & Hill, 2006). Thereby, according to literature recommendations, our data analysis procedure followed two steps: data preparation and model development.

Our data analysis procedure followed two steps: data preparation and model development. In data preparation, we took several steps to ensure the validity and reliability of our analysis. First, we transformed the score column from 0 to 4 instead of 1–5, as this approach helps with model interpretability once the intercept becomes meaningful. Second, it was defined *GPT-4* as the reference level for both *speaker* and *gender columns* to ensure it could be compared against other levels based on their coefficient (e.g., if GPT-4 is the reference level and *male* has a significant coefficient, it means there is a significant difference between them). Third, we created a new column named *gd_sp*, which merged columns *speaker* and *gender* to allow us to compare both factors to GPT-4 at the same time. This was necessary because both speaker and gender had the GPT-4 level, so adding both to the same model would lead to a deficient ranking. All of these steps were based on established best practices in HLM to increase our data analysis' validity and reliability (Hox et al., 2010).

Model development was based on a systematic bottom-up approach (Hox et al., 2010). First, we created a baseline model with no predictors, which only accounted for the different questions in its random structure, to cope with the repeated answers for the same question. Second, we defined the fixed effects structure, in which we tested gender, speaker, and gd_sp, to understand how they affect the outcome variable compared to GPT-4, aligned to **RQ1** and **RQ2**. Third, we introduced random intercepts to account for intercept variability across different hierarchy levels. Lastly, we incorporated random slopes to examine the variability in the relationships between predictors and the outcome variable across different levels of the hierarchy. The last two substeps align with **RQ3**. For each of those substeps, each inclusion was assessed using likelihood ratio tests (LRT) to compare the nested models (Gelman & Hill, 2006). By systematically comparing nested models at each inclusion substep, LRTs allow us to assess the significance of model improvements and additions. This rigorous statistical method helps mitigate the risk of overfitting and ensures that each enhancement to the model is statistically justified, aligning with best practices in HLM.

Moreover, because we have a somewhat large dataset, it might increase the chances of Type I error. Therefore, we set the significance level for LRTs tests at a 99% confidence level and adjusted p-values using the Bonferroni approach to account for multiple comparisons (Cairns, 2019). Setting a stringent significance level for LRTs enhances the reliability of our model evaluations, reducing the likelihood of false-positives and providing a robust framework for interpreting the relationships within our hierarchical data structure. Overall, the strategic application of LRTs strengthens the validity of our data analysis by ensuring that our model accurately captures the underlying complexities of the data while maintaining statistical reliability.

Lastly, note that all analyses were conducted using R,[1] R studio,[2] and the *lme4* package (Bates et al., 2014). Overall, this packages uses maximum likelihood for model estimation, similar to prior research on HML Hox et al. (2010). Notably, *lme4* builds on previous efforts to facilitate the practical usage of HML, offering benefits such as improved performance for large problems, efficiency for fitting models with crossed random effects, implementation of profile likelihood confidence intervals on random-effects parameters, among others (Bates et al., 2014). Hence, while R has been extensively used for data analysis, lme4 is a package that has been extensively used and particularly designed to optimize HML. Therefore, we believe it provides reliable and valid background for our data analysis and refer the interested reader to (Bates et al., 2014) for further details on its workings.

4. Results

4.1. RQ1: quality of GPT-4 answers

Overall, our dataset shows that the quality of GPT4's answers (Mean, M = 3.69; Standard Deviation, SD = 1.33) were higher than those of humans (M = 3.19; SD = 1.38). Furthermore, Fig. 2 demonstrates descriptive statistics regarding answers' evaluations according to its subgroups. The figure suggests GPT-4's evaluations were better than those of humans, regardless of gender and speaker, for most levels of Bloom's taxonomy (e.g., understanding, applying, and analyzing). However, the figure suggests some deviations from that pattern in specific cases (e.g., GPT-4 compared to male-native for understanding). Moreover, the figure also offers variations depending on the topic (e.g.,

[1] https://www.r-project.org/.
[2] https://rstudio.com/.

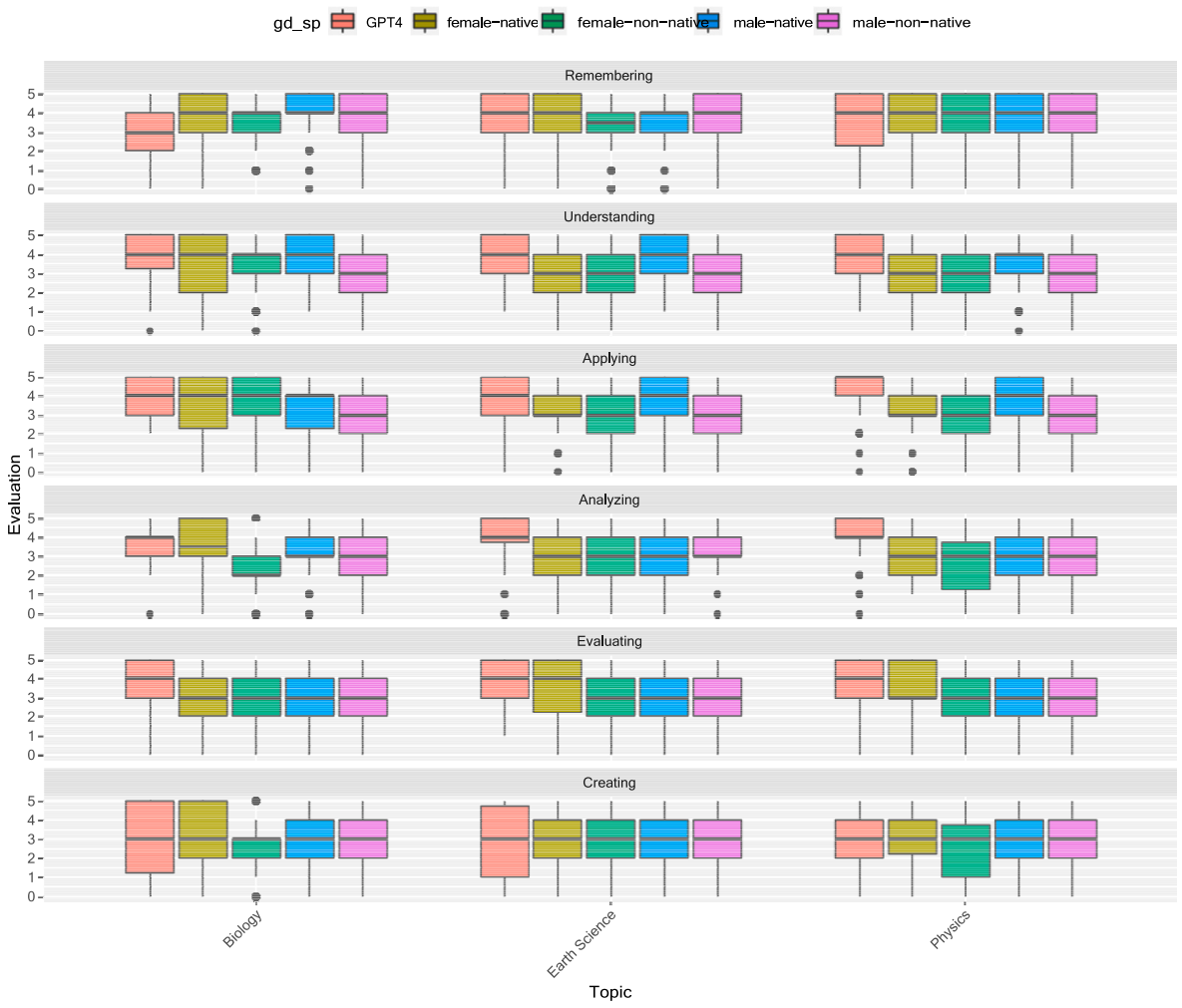

Fig. 2. Descriptive statistics on the quality of GPT-4 answers compared to those of humans in terms of gender, speaker (gp_sp), level of Bloom's taxonomy, and topic.

GPT-4 compared to female-native for Biology, Earth Science, and Physics). Accordingly, we proceed with the HLM to disentangle those complex relationships and provide more details for the proposed analysis.

4.2. RQ2: comparison of GPT-4 and humans considering gender and proficiency in English

To answer RQ2, we tested how the quality of GPT-4 and human answers compare depending on demographic information: a) gender and b) language proficiency. For this, we followed the HLM procedure as described in Section 3.3. Mainly, we focused on analyzing fixed effects because people's gender and being a native speaker concern answer-level characteristics.

Initially, we created a baseline (i.e., intercept-only) model, M0, to ensure that adding some variables improves model fit, as suggested in the literature (Hox et al., 2010). Following, we fitted models featuring only gender (M1) and only speaker (M2) as the predictor to answer RQ2a and RQ2b, respectively. Next, we did a model featuring the combination of both variables (M3) to respond to RQ2 in light of their combination. Table 4 summarizes these models. Then, we conducted LRT tests to understand which of those models yield a better fit to our data and how they compare to each other (see Table 5).

Overall, Table 4 demonstrates that the intercept (i.e., GPT-4 answers) is positive for all models and that they do not include zero. While none of the remaining coefficients have zero either, they are all negative, indicating that the quality of their answers is worse than those of GPT-4.

Table 4
Fixed effects-based models comparing the quality of GPT answers (reference) to males, females, native speakers, and non-native speakers. Data are shown as Coefficients (Lower Confidence Interval - Upper Confidence Interval).

Predictors	M0	M1	M2	M3
(Intercept)	3.29 (3.25–3.33)	3.69 (3.62–3.76)	3.69 (3.62–3.76)	3.69 (3.62–3.76)
gender [f]		−0.53 (−0.61 to −0.45)		
gender [m]		−0.46 (−0.55 to −0.38)		
spk [n]			−0.35 (−0.43 to −0.26)	
spk [nn]			−0.65 (−0.73 to −0.56)	
gd_sp[f-n]				−0.35 (−0.45 to −0.25)
gd_sp[f-nn]				−0.71 (−0.81 to −0.61)
gd_sp[m-n]				−0.35 (−0.44 to −0.25)
gd_sp[m-nn]				−0.58 (−0.68 to −0.48)

f = female; m = male; spk = speaker; n = native; nn = non-native; gd_sp = gender_speaker.

Furthermore, Table 4 demonstrates significant improvements in each model comparison but between M2 and M3. Therefore, suggesting the model with the best fit was M2, which features the speaker as a

Table 5
Comparison statistics (likelihood ratio tests - LRT) for the fixed-effect models (M) used to answer RQ2.

M	Df	AIC	BIC	logLik	Deviance	Chisq	Chi Df	p-value	Adj p
M0	3	25735	25756	−12865	25729	–	–	–	NA
M1	5	25577	25611	−12783	25567	162.5955	2	<0.001	0
M2	5	25508	25542	−12749	25498	68.9025	0	<0.001	0
M3	7	25505	25553	−12746	25491	6.8291	2	0.033	0.099

predictor, as summarized in Fig. 3. Thus, we considered M2 as our new baseline in the subsequent analyses.

4.3. RQ3: comparison of GPT-4 and humans considering different difficult levels and topics for the questions

To answer RQ3, we analyzed how the quality of GPT-4 and human answers for different questions features: a) complexity and b) the topic, following the procedure described in Section 3.3. Specifically, this analysis focused on random intercepts and slopes, given that question complexity (based on Bloom's taxonomy) and topic concern second and third-level characteristics of our dataset (Hox et al., 2010) (see 3.1 for details).

Building upon M2, we first investigated whether allowing the intercepts of Bloom's levels (M4) and topic (M5) to vary would improve M2 toward answering RQ3a and RQ3b, respectively. For this, we fitted two new models, in which we summarize their random effects in Table 6 (please refer to Table 4 for the fixed effect structure). Notably, the variance for the topic is zero, in contrast to that of Bloom. Accordingly, after conducting LRT tests, the results show (see Table 7) that M4 improves M2, but M5 yields no improvement compared to M4. This finding indicates that how the quality of GPT-4 and human answers compare does not depend on the topic (RQ3b). Therefore, we proceed with our analysis using M4, which features the speaker as a predictor and bloom as the grouping level.

Next, we investigated whether the addition of random slopes would improve our model fit compared to M4. Particularly, we investigated adding random slopes based on Bloom's levels (M6) and the question id (M7) to increase the reliability of our answer to RQ3a. As our previous find revealed topic did not play a role in modeling our data, we did not consider it in this analysis, while we kept question id as it enables the regression model to account for repeated-measures (i.e., multiple answers to the same question). Table 8 summarizes the random effects' variances for this investigation, in which we found that M6 yielded an improvement compared to M4, as well as M7 did so compared to M6 (see Table 9). Therefore, our final model for RQ3a is M7, which features random slopes for Bloom's levels and question id, besides speaker as predictor, as summarized in Fig. 4. Thus, those results suggest the quality of GPT-4 answers differed depending on the level of Bloom's

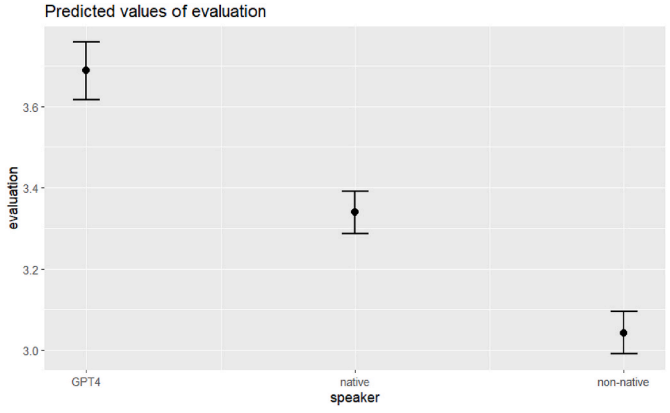

Fig. 3. Coefficients and confidence intervals for M2, the model featuring speaker as fixed-effect, which yielded the best fit for answering RQ2.

Table 6
Summary of random coefficients' variance in random-intercept models (M) fit to answer RQ3.

Model	M2	M4	M5
Groups			
question_id			
(Intercept)	0.072 (0.269)	0.014 (0.120)	0.014 (0.120)
bloom	–	0.058 (0.241)	0.058 (0.241)
topic	–	–	0.000 (0.000)
Name			
(Intercept)	–	–	–
Residual	1.792 (1.339)	1.792 (1.339)	1.792 (1.339)

taxonomy, overcoming those of humans in understanding, applying, analyzing, and evaluating questions, but not in remembering and creating ones (RQ3a).

5. Discussion

Overall, our results suggested an advantage for GPT-4 compared to humans. Our main results were: (i) GPT-4 shows an overall advantage over humans in answering short open questions, as indicated by negative coefficients in fixed effect analysis (RQ1); Gender differences among humans do not affect performance, but there's a notable contrast between native and non-native speakers, where GPT-4 outperforms both groups but has a minor advantage over native speakers (RQ2); Contextual characteristics like the topic do not significantly influence GPT-4's advantage over humans, but levels of Bloom's taxonomy in questions impact GPT-4's performance, wherein its advantage encompass understanding, applying, analyzing, and evaluating levels (RQ3). Next, we elaborate on how our findings answer our RQs in light of the relevant literature and discuss this article's contributions and implications.

5.1. Research questions and findings

This paper addressed three RQs. RQ1 concerned the extent to which GPT-4 is capable of answering open-ended questions. Our findings demonstrate that, overall, the scores of GPT-4's answers are slightly higher than those of humans' (see Section 4.1 and Fig. 2). Hence, providing evidence that GPT-4 is capable of answering open-ended questions slightly better than humans. The literature has advocated towards integrating AI innovative solutions into educational products (Yan et al., 2023), and GPT-4 offers a means to empower several AAG-based features, such as real-time generation of personalized feedback and assessment recommendation (Divya et al., 2023; Wang & Demszky, 2023). Therefore, our findings background such claims by expanding it with empirical evidence supporting GPT-4 capabilities to contribute to AAG.

RQ2 concerned possible differences between GPT-4 and humans in answering open-ended questions depending on the demographic information. Our findings demonstrate that one's native language significantly affected GPT-4 and humans difference, in contrast to gender (see Section 4.2, Fig. 3, and Tables 5 and 4). Hence, providing evidence that GPT-4 and humans capabilities differ depending on whether one is a native speaker or not and that gender was not a significant factor when controlling for being a native. While AI research has been concerned with potential gender-related bias that could perpetuate social

Table 7
Comparison statistics (likelihood ratio tests - LRT) for the fixed-effect models (M) used to answer RQ3.

M	Df	AIC	BIC	logLik	Deviance	Chisq	Chi Df	p-value	Adj p
M2	5	25508	25542	−12749	25498	–	–	–	
M4	6	25338	25379	−12663	25326	172.11	1	<0.01	<0.01
M5	7	25340	25388	−12663	25326	0.00	1	1	1

Table 8
Summary of random coefficients' variance in random-slope models (M) fit to answer RQ3.

Parameter	M4	M6	M7
Groups			
question_id (Intercept)	0.014 (0.120)	0.017 (0.129)	0.000 (0.005)
bloom (Intercept)	0.058 (0.241)	0.001 (0.023)	0.153 (0.391)
bloom.1	–	–	0.000 (0.000)
speakernative	–	0.136 (0.369)	0.008 (0.090)
speakernon-native	–	0.126 (0.356)	0.104 (0.322)
topic (Intercept)	–	–	0.000 (0.000)
Name			
(Intercept)	–	–	–
Residual	1.792 (1.339)	1.768 (1.330)	1.739 (1.319)

inequalities (Memarian & Doleck, 2023; Woolf, 2022), our results demonstrate people's answers and how they compare to AI-generated ones were not affected by gender. Differently, our results revealed that language skills, from the perspective of being a native speaker or not, play a significant role on how GPT-4 answers compare to human ones. Thus, our findings advance the understanding of how GPT-4's capabilities to answer open-ended questions compare to humans and demographic factors that (do not) affect this difference.

Lastly, RQ3 concerned possible differences between GPT-4 and humans in answering open-ended questions depending on characteristics of the to-be-answered questions. Our findings demonstrate that questions Bloom level significantly affected GPT-4's capabilities, whereas the question itself did not (see Section 4.3, Fig. 4, and Tables 6–9). Hence, providing evidence that GPT-4 and humans capabilities differ depending on the question's complexity, as defined through Bloom's taxonomy, rather than the question itself. Learning is a complex process that involves multiple cognitive processes, ranging from recalling information to creation, and Bloom's taxonomy provides a reliable framework to categorize these processes Anderson and Sosniak (1994). Thereby, regardless of the AAG task, learners might need different levels of support for an incluse AI-powered system Pedro et al. (2019); Vinuesa et al. (2020). Accordingly, our findings reveal evidence that one might expect GPT-4's capabilities to vary depending on the question's complexity level.

5.2. Contributions and implications

From a technical perspective, it is important to highlight three principal dimensions. First, the prompt influences the performance of the LLM models (White et al., 2023). This dependency suggests a prompt engineering process can optimize output quality (Karmaker Santu & Feng, 2023). This study focused on evaluating a single prompt to ensure parity in the experimental conditions, mirroring the approach where students were similarly instructed with a single directive. Such a methodological choice is crucial for maintaining the integrity and fairness of the comparative analysis between the LLM's performance and the students' responses (Patil & Adhiya, 2022).

Second, it is noteworthy that GPT-4 demonstrated inferior performance compared to native-speaking human participants in questions related to the remembering bloom taxonomy category. This outcome was unexpected, particularly since these questions predominantly pertain to content recall (Herrmann Werner et al., 2023). Nonetheless, it is crucial to acknowledge that the original evaluative criteria encompassed both the content and the argumentative quality of the responses, see section 3 for more details. More detailed observations revealed that GPT-4's responses were predominantly content-centric and typically concise in nature. In contrast, human responses exhibited a more comprehensive approach, integrating both content and argumentative elements. This distinction could highlight a critical aspect of GPT-4's current limitations in mimicking the depth and breadth of human cognitive processing in response generation, particularly in tasks requiring nuanced argumentation and elaboration (Li et al., 2023).

Third, it is pertinent to highlight that GPT-4 also does not outperform native speakers in the 'creating' category of Bloom's Taxonomy. This category is generally associated with the ability to extrapolate new concepts and apply pre-acquired knowledge in novel contexts. In this domain, GPT-4 demonstrated a marginally diminished capacity to generate innovative scenarios for the given activity. This finding is significant as it underscores the current limitations of GPT-4 in engaging with tasks that require original thought and the creative application of

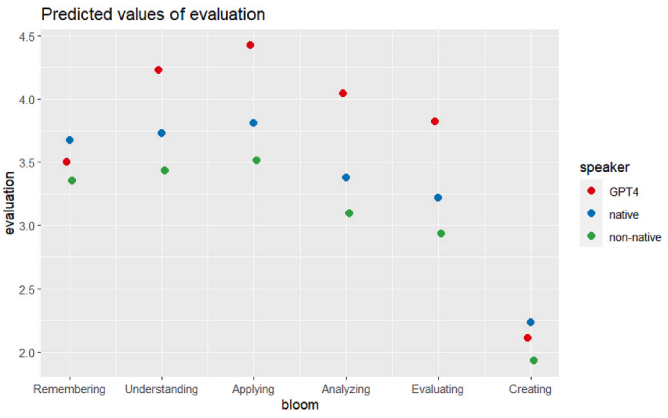

Fig. 4. Coefficients for M7, the model featuring speaker as fixed effect and random slopes for both questions and question complexity (measured through the level of Bloom's taxonomy), which yielded the best fit for answering RQ3.

Table 9
Model (M) comparison statistics.

M	Df	AIC	BIC	logLik	Deviance	Chisq	Chi Df	p-value	Adj p
M4	6	25338	25379	−12663	25326	–	–	–	–
M6	12	25279	25362	−12628	25255	70.818	6	<0.01	<0.01
M7	18	25275	25399	−12619	25239	16.373	6	0.01189	0.024

existing information, a key aspect of higher-order cognitive processing (Chang, 2023).

Our investigation identified three significant practical implications:

1. Our analysis utilizing the GPT-4 model has demonstrated its capability to respond to diverse questions regarding complexities and topics accurately. This finding underscores the potential of GPT-4 as a basis for developing advanced question-answering systems within educational settings.
2. The observation that GPT-4 consistently surpassed non-native-speaking human participants in performance suggests a promising direction for leveraging this model in supporting non-native speakers more extensively.
3. The quality of GPT-4's responses and its potential for further detailed elaboration on activities posits it as a potential basis for a more comprehensive system to enhance teacher support mechanisms in student assessment processes.

Furthermore, the dataset created is this study is available.[3] This empowers researchers to expand this research line by facilitating replications and future research. Thereby, sharing our dataset expands this paper's contribution.

6. Limitations

The validity of the model employed in our study is critical to our findings, warranting rigorous evaluation. As with most statistical methodologies, HLM operates under specific assumptions, such as normality of residuals and homogeneity of variances (Hox et al., 2010). While statistical tests designed to verify these assumptions exist, it's essential to acknowledge that these tests are predicated on their own assumptions. This recursive nature of assumption verification introduces new layers of complexity and potential validity threats to the overall analytical framework (Cairns, 2019). However, it's pertinent to consider the robustness of HLM in the face of assumption violations. The literature suggests that with an adequately sized sample, HLM can maintain high accuracy in results, even when certain foundational assumptions are not fully met. This strength is highlighted by previous work (Darandari, 2004; Montgomery et al., 2012; Schielzeth et al., 2020). In light of these considerations, we posit that the potential threats to the validity of our findings stemming from the inherent complexities of assumption testing in HLM are mitigated. The robust nature of HLM, particularly in the context of a sufficiently large sample size, lends credence to the reliability and accuracy of our study's outcomes. Consequently, we assert that the issues related to assumption verification do not pose significant challenges to the validity of our findings.

Our dataset's validity is another important consideration. While relying on crowd workers for question formulation introduces variability in question quality, potentially impacting dataset integrity, this approach allows for diverse perspectives and contributions, enriching the dataset with a wide range of questions that align with people's different perspectives. Additionally, whereas we understand a question might fit into more than one level of Bloom's taxonomy and one might face challenges in categorizing questions Anderson and Sosniak (1994), similar research shows promising evidence on crowd workers' ability to perform such a categorization Moore et al. (2023). Nevertheless, to mitigate this issue, we provided descriptions for each level of the taxonomy and requested crowd workers to assign questions into the categorize it fitted the most to ensure that all questions would be attributed to the level it has a better fit.

Furthermore, the involvement of crowd workers in multiple areas of the study (e.g., response generation and evaluation), raises concerns about consistency and expertise, which could influence the dataset's reliability. However, crowd workers' diverse backgrounds and experiences also contribute to a holistic evaluation process, capturing a broader spectrum of responses and perspectives that enhance the dataset's comprehensiveness. Another important issue is that we were not able to track crowd workers' interactions or background (e.g., educational level), but past research shows encouraging evidence that the overall contribution from these workers' are reliable Soprano et al. (2021); Pavlidou et al. (2020); Auer et al. (2021). Thus, despite these dataset-related limitations, the structured evaluation process and diverse contributions from crowd workers contributes to achieving a robust dataset that reflects the complexities of learning and assessment across various domains as well as mitigates individual-level issues based on the sample size.

7. Final consideration

Recent research has called for equipping existing educational technology with innovative solutions such as state-of-the-art LLMs. While previous studies have explored the development of question-answering systems, research investigating LLMs in that context is limited. In this context, LLMs' (e.g., GPT-4) capabilities and contextual understanding pave the way for innovative approaches in assessing and guiding student learning, such as deployment of automated question-answering systems, possibly leading to more efficient, personalized, and adaptive educational approaches. However, current research lacks comprehensive evaluations of LLMs, particularly in terms of their performance in responding to diverse open-ended questions. Existing studies often focus on limited domains or neglect varying question complexities, hindering a holistic understanding of these models' capabilities compare to humans. Bridging this gap is crucial for unlocking the true potential of LLMs in education, necessitating robust assessments across diverse question types, demographics, and cognitive dimensions to fully leverage their benefits in enhancing learning outcomes.

To address that gap, this paper presented an empirical analysis comparing GPT-4 to humans' ability to answer short open-ended questions. Tackling prior research's limitations, this study generated a new dataset composed of 7380 answers to 738 open-ended questions of three topics (i.e., Biology, Earth Sciences, and Physics) and all levels of Bloom's taxonomy, featuring answers from GPT-4 and humans, including males and females, native and non-native speakers. By analyzing our dataset, we mainly found that GPT-4 has a consistent advantage over humans in answering short open-ended questions and that there are noticeable distinctions between native and non-native speakers' performances, although GPT-4 surpasses both. On the other hand, while contextual factors like question topics did not significantly impact GPT-4's advantage, differing levels of Bloom's taxonomy in questions does. GPT-4 outperforms humans in understanding, applying, analyzing, and evaluating cognitive levels, but yielded comparable performances for remembering and creating ones.

These results underscore critical dimensions within GPT-4's performance. Despite the overall advantage, GPT-4 exhibited inferiority to native-speaking humans in 'remembering' and 'creating' categories of Bloom's Taxonomy, revealing limitations in mimicking nuanced argumentation and innovative idea generation, respectively. These insights illuminate GPT-4's current constraints in emulating human cognitive depth and creativity. Thus, the practical implications encompass GPT-4's potential for diverse and accurate question-answering systems, its supportive role for non-native speakers, and its prospective application in refining teacher support mechanisms for student assessments. Furthermore, we share our dataset to facilitate replicating and expanding upon our findings, extending this article's contribution. In summary, our findings advocate for strategic LLM usage, considering their strengths and limitations, within educational frameworks to bolster learning and assessment paradigms effectively.

[3] https://tinyurl.com/2ppv6mdp.

8. Statements

The study was not submitted to an ethical committee for review because it used crowdsourcing platforms for data collection and, at the time of writing, Brazilian regulation do not require reviewing such projects. Informed consent was obtained from all participants, and their privacy rights were strictly observed. The participants were protected by hiding their personal information during the research process. They knew that the participation was voluntary and they could withdraw from the study at any time. There is no potential conflict of interest in this study. The data can be obtained by sending request e-mails to the corresponding author. The paper was partially funded by Open AI research grant.

CRediT authorship contribution statement

Luiz Rodrigues: Writing – original draft, Methodology, Formal analysis, Conceptualization. **Filipe Dwan Pereira:** Methodology, Formal analysis. **Luciano Cabral:** Formal analysis. **Dragan Gašević:** Writing – review & editing, Methodology. **Geber Ramalho:** Writing – review & editing, Methodology. **Rafael Ferreira Mello:** Writing – review & editing, Resources, Methodology, Data curation, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Alzubaidi, L., Bai, J., Al-Sabaawi, A., Santamaría, J., Albahri, A. S., Al-dabbagh, B. S. N., et al. (2023). A survey on deep learning tools dealing with data scarcity: Definitions, challenges, solutions, tips, and applications. *J Big Data, 10 46*. https://doi.org/10.1186/s40537-023-00727-2

Anderson, L. W., & Sosniak, L. A. (1994). *Bloom's taxonomy*. Chicago, IL, USA: Univ. Chicago Press.

Auer, E. M., Behrend, T. S., Collmus, A. B., Landers, R. N., & Miles, A. F. (2021). Pay for performance, satisfaction and retention in longitudinal crowdsourced research. *PLoS One, 16*, Article e0245460.

Basu, S., Jacobs, C., & Vanderwende, L. (2013). Powergrading: A clustering approach to amplify human effort for short answer grading. *Transactions of the Association for Computational Linguistics, 1*, 391–402.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). Fitting linear mixed-effects models using lme4. *arXiv preprint arXiv:1406.5823*.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., et al. (2020). Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, & H. Lin (Eds.), *Advances in neural information processing systems, 33* pp. 1877–1901). Curran Associates, Inc.).

Cairns, P. (2019). *Doing better statistics in human-computer interaction.* Cambridge University Press.

Chang, E. Y. (2023). Examining gpt-4: Capabilities, implications and future directions. In *The 10th international conference on computational science and computational intelligence*.

Darandari, E. Z. (2004). *Robustness of hierarchical linear model parameter estimates under violations of second-level residual homoskedasticity and independence assumptions*. The Florida State University. Ph.D. thesis.

de Winter, J. C. F. (2023). Can chatgpt pass high school exams on English language comprehension? *International Journal of Artificial Intelligence in Education*. https://doi.org/10.1007/s40593-023-00372-z

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). *Bert: Pre-training of deep bidirectional transformers for language understanding*.

Divya, A., Haridas, V., & Narayanan, J. (2023). Automation of short answer grading techniques: Comparative study using deep learning techniques. In *2023 fifth international conference on electrical, computer and communication technologies (ICECCT)* (pp. 1–7). IEEE).

Gelman, A., & Hill, J. (2006). *Data analysis using regression and multilevel/hierarchical models*. Cambridge university press.

He, P., Liu, X., Gao, J., & Chen, W. (2020). Deberta: Decoding-enhanced bert with disentangled attention. *CoRR abs/2006.03654*.

Herrmann Werner, A., Festl-Wietek, T., Holderried, F., Herschbach, L., Griewatz, J., Masters, K., et al. (2023). Assessing chatgpt's mastery of bloom's taxonomy using psychosomatic medicine exam questions. *medRxiv*, 2023–08.

Horbach, A., Stennmanns, S., & Zesch, T. (2018). Cross-lingual content scoring. In *Proceedings of the thirteenth workshop on innovative use of NLP for building educational applications* (pp. 410–419).

Hox, J. J., Moerbeek, M., & Van de Schoot, R. (2010). *Multilevel analysis: Techniques and applications.* Routledge.

Huang, Z., Shen, Y., Li, X., Wei, Y., Cheng, G., Zhou, L., et al. (2019). GeoSQA: A benchmark for scenario-based question answering in the geography domain at high school level. In K. Inui, J. Jiang, V. Ng, & X. Wan (Eds.), *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)* (pp. 5866–5871). Hong Kong, China: Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1597.

Jayaraman, J., & Black, J. (2022). Effectiveness of an intelligent question answering system for teaching financial literacy: A pilot study. In *Innovations in learning and technology for the workplace and higher education: Proceedings of 'the learning ideas Conference'2021* (pp. 133–140). Springer.

Karmaker Santu, S. K., & Feng, D. (2023). TELeR: A general taxonomy of LLM prompts for benchmarking complex tasks. In H. Bouamor, J. Pino, & K. Bali (Eds.), *Findings of the association for computational linguistics: Emnlp 2023* (pp. 14197–14203). Singapore: Association for Computational Linguistics). https://doi.org/10.18653/v1/2023. findings-emnlp.946.

Kasneci, E., Seßler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., et al. (2023). Chatgpt for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences, 103*, Article 102274.

Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., & Iwasawa, Y. (2022). Large language models are zero-shot reasoners. *Advances in Neural Information Processing Systems, 35*, 22199–22213.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., et al. (2020). Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In D. Jurafsky, J. Chai, N. Schluter, & J. Tetreault (Eds.), *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 7871–7880). Online: Association for Computational Linguistics). https://doi.org/10.18653/v1/2020.acl-main.703.

Li, Y., Sha, L., Yan, L., Lin, J., Raković, M., Galbraith, K., et al. (2023). Can large language models write reflectively. *Computers in Education: Artificial Intelligence, 4*, Article 100140.

Liu, Y., Xu, B., Yang, Y., Chung, T., & Zhang, P. (2019). Constructing a hybrid automatic q&a system integrating knowledge graph and information retrieval technologies. In *Foundations and trends in smart learning: Proceedings of 2019 international conference on smart learning environments* (pp. 67–76). Springer.

Memarian, B., & Doleck, T. (2023). Fairness, accountability, transparency, and ethics (fate) in artificial intelligence (ai), and higher education: A systematic review. *Computers in Education: Artificial Intelligence*, Article 100152.

Montgomery, D. C., Peck, E. A., & Vining, G. G. (2012). *Introduction to linear regression analysis, 821*. John Wiley & Sons.

Moore, S., Fang, E., Nguyen, H. A., & Stamper, J. (2023). Crowdsourcing the evaluation of multiple-choice questions using item-writing flaws and bloom's taxonomy. In *Proceedings of the tenth ACM conference on learning@ scale* (pp. 25–34).

Nguyen, H. A., Stec, H., Hou, X., Di, S., & McLaren, B. M. (2023). Evaluating chatgpt's decimal skills and feedback generation in a digital learning game. In *European conference on technology enhanced learning* (pp. 278–293). Springer.

OpenAI. (2023). *Gpt-4 technical report*.

Parsons, B., & Curry, J. H. (2023). Can chatgpt pass graduate-level instructional design assignments? Potential implications of artificial intelligence in education and a call to action. *TechTrends*, 1–12.

Patil, S., & Adhiya, K. P. (2022). Automated evaluation of short answers: A systematic review. *Intelligent Data Communication Technologies and Internet of Things: Proceedings of ICICI, 2021*, 953–963.

Pavlidou, I., Papagiannidis, S., & Tsui, E. (2020). Crowdsourcing: A systematic review of the literature using text mining. *Industrial Management & Data Systems, 120*, 2041–2065.

Pedro, F., Subosa, M., Rivas, A., & Valverde, P. (2019). *Artificial intelligence in education: Challenges and opportunities for sustainable development*.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. (2019). Language models are unsupervised multitask learners. *OpenAI blog, 1*, 9.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., et al. (2023). *Exploring the limits of transfer learning with a unified text-to-text transformer*.

Rawat, A., Kumar, S., & Samant, S. S. (2023). A systematic review of question classification techniques based on bloom's taxonomy. In *2023 14th international conference on computing communication and networking technologies (ICCCNT)* (pp. 1–7). IEEE).

Rosoł, M., Gasior, J. S., Łaba, J., Korzeniewski, K., & Młyńczak, M. (2023). Evaluation of the performance of gpt-3.5 and gpt-4 on the polish medical final examination. *Scientific Reports, 13*, Article 20512.

Schielzeth, H., Dingemanse, N. J., Nakagawa, S., Westneat, D. F., Allegue, H., Teplitsky, C., et al. (2020). Robustness of linear mixed-effects models to violations of distributional assumptions. *Methods in Ecology and Evolution, 11*, 1141–1152.

Soprano, M., Roitero, K., La Barbera, D., Ceolin, D., Spina, D., Mizzaro, S., et al. (2021). The many dimensions of truthfulness: Crowdsourcing misinformation assessments on a multidimensional scale. *Information Processing & Management, 58*, Article 102710.

Vaswani, A., Shazeer, N. M., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). Attention is all you need. In *Neural information processing systems*.

Vinuesa, R., Azizpour, H., Leite, I., Balaam, M., Dignum, V., Domisch, S., et al. (2020). The role of artificial intelligence in achieving the sustainable development goals. *Nature Communications, 11*, 1–10.

Wang, R., & Demszky, D. (2023). Is ChatGPT a good teacher coach? Measuring zero-shot performance for scoring and providing actionable insights on classroom instruction. In E. Kochmar, J. Burstein, A. Horbach, R. Laarmann-Quante, N. Madnani, A. Tack, V. Yaneva, Z. Yuan, & T. Zesch (Eds.), *Proceedings of the 18th workshop on innovative use of NLP for building educational applications (BEA 2023)* (pp. 626–667). Toronto, Canada: Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.bea-1.53.

White, J., Fu, Q., Hays, S., Sandborn, M., Olea, C., Gilbert, H., et al. (2023). *A prompt pattern catalog to enhance prompt engineering with chatgpt. arXiv preprint arXiv: 2302.11382*.

Woolf, B. (2022). Introduction to ijaied special issue, fate in aied. *International Journal of Artificial Intelligence in Education, 32*, 501–503.

Yan, L., Sha, L., Zhao, L., Li, Y., Martinez-Maldonado, R., Chen, G., et al. (2023). Practical and ethical challenges of large language models in education: A systematic scoping review. *British Journal of Educational Technology*. https://doi.org/10.1111/bjet.13370. n/a.

Yenduri, G., M, R., G, C. S., Y, S., Srivastava, G., Maddikunta, P. K. R., et al. (2023). *Gpt (generative pre-trained transformer) – a comprehensive review on enabling technologies, potential applications, emerging challenges, and future directions*.

Zirar, A. (2023). Exploring the impact of language models, such as chatgpt, on student learning and assessment. *The Review of Education, 11*, e3433.

Ziyu, Z., Qiguang, C., Longxuan, M., Mingda, L., Yi, H., Yushan, Q., et al. (2023). Through the lens of core competency: Survey on evaluation of large language models. *Proceedings of the 22nd Chinese National Conference on Computational Linguistics, 2*(Frontier Forum), 88–109.