Fine-tuning ChatGPT for Automatic Scoring

Ehsan Latif[1] and Xiaoming Zhai[1*]

[1*]AI4STEM Education Center, University of Georgia, Athens, GA, USA .

*Corresponding author(s). E-mail(s): xiaoming.zhai@uga.edu;
Contributing authors: ehsan.latif@uga.edu;

Abstract

This study highlights the potential of fine-tuned ChatGPT (GPT-3.5) for automatically scoring student written constructed responses using example assessment tasks in science education. GPT-3.5 has been trained over enormous online language materials such as journals and Wikipedia; however, direct usage of pre-trained GPT-3.5 is insufficient for automatic scoring as students do not utilize the same language as journals or Wikipedia, and contextual information is required for accurate scoring. All of these imply that a fine-tuning of a domain-specific model using data for specific tasks can enhance model performance. In this study, we fine-tuned GPT-3.5 on six assessment tasks with a diverse dataset of middle-school and high-school student responses and expert scoring. The six tasks comprise two multi-label and four multi-class assessment tasks. We compare the performance of fine-tuned GPT-3.5 with the fine-tuned state-of-the-art Google's generated language model, BERT. The results show that in-domain training corpora constructed from science questions and responses for BERT achieved average accuracy = 0.838, SD = 0.069. GPT-3.5 shows a remarkable average increase (9.1%) in automatic scoring accuracy (mean = 9.15, SD = 0.042) for the six tasks, p =0.001 $<$ 0.05. Specifically, for each of the two multi-label tasks (item 1 with 5 labels; item 2 with 10 labels), GPT-3.5 achieved significantly higher scoring accuracy than BERT across all the labels, with the second item achieving a 7.1% increase. The average scoring increase for the four multi-class items for GPT-3.5 was 10.6% compared to BERT. Our study confirmed the effectiveness of fine-tuned GPT-3.5 for automatic scoring of student responses on domain-specific data in education with high accuracy. We have released fine-tuned models for public use and community engagement.

Keywords: Large Language Model (LLM), BERT, GPT-3.5, Finetune, Education, Automatic Scoring

1 Introduction

The importance of education in stimulating creativity, encouraging critical thinking, and establishing the groundwork for individual and societal improvement has long been acknowledged. This claim becomes more relevant today, marked by extraordinary technological advancements, a developing digital ecosystem, and a complex globalized society (Linn et al., 2014). There has never been a greater need for a strong and effective education system that provides students with the knowledge and skills they need to successfully traverse the challenges of the 21st century. Despite the needs, modern educational frameworks sometimes find themselves up against complex problems. One of these is adaptability, as conventional teaching and learning methodologies occasionally find it challenging to keep up with the rapidly changing state of knowledge and technology. The challenge of personalization is particularly challenging– a one-size-fits-all strategy (Teasley, 2017) frequently fails to provide meaningful and unique learning experiences given the variety of student backgrounds and demands (Limna et al., 2022).

Artificial intelligence (AI) has taken center stage in this dynamic environment as a possible game-changer for the education industry (Zhai, 2023a; Latif et al., 2023; Lee et al., 2023). AI promises to address some of the most challenging issues in education thanks to its unmatched data processing powers, predictive analytics, and adaptable algorithms. AI has the potential not just to augment but also completely transform current educational practices, whether it be through the creation of personalized learning pathways, the provision of real-time feedback, or even the facilitation of massive online courses with thousands of students (Namatherdhala et al., 2022; Limna et al., 2022; Zhai et al., 2020). The nexus of AI and education creates opportunities for more individualized, adaptable, and inclusive learning experiences, ushering in a new era of pedagogy (Holmes and Tuomi, 2022; Wu et al., 2023; Zhai and Wiebe, 2023). Although the use of AI in education is growing, not all AI solutions are suited to the particular needs of the educational field. Generic AI models frequently lack the precise contextual awareness required for successful educational interactions (Zhai et al., 2021; Liu et al., 2023; Tahiru, 2021; Ng et al., 2023).

The ChatGPT program, created by OpenAI, is a leading example of natural language processing and the promise of AI in education (Mhlanga, 2023; Zhai, 2023b). While ChatGPT demonstrates outstanding broad conversational skills, there are inherent limitations when using it without adjustments in specialized domains like education (Zhai, 2022). The general trend towards fine-tuning AI technology for certain activities resonates with improving ChatGPT for educational purposes (Liu et al., 2023). A fine-tuned and applying Chain-of-Thought (Lee et al., 2023) using ChatGPT may offer individualized, engaging, and successful learning experiences consistent with the philosophy of lifelong learning by integrating domain-specific knowledge, context awareness, and pedagogical techniques (Zhai, 2022).

Inspired by the capabilities of ChatGPT in predicting sentences even with zero-shot (Wei et al., 2023) and remarkable responses with few-shots (Dai et al., 2023; Liu et al., 2023) for different applications, we propose a fine-tuned ChatGPT for science educational automatic scoring. In this study, we fine-tuned the public version of ChatGPT, i.e., GPT-3.5, with six complex assessment tasks in science education.

Our findings show high accuracy in automatic scoring using fine-tuned ChatGPT compared to the performance of Google's pre-trained language model BERT (Devlin et al., 2018). This study can set grounds for future uses of fine-tuned Chatbots in education and shows high potential for science education assessments.

The study contributes to the field in multiple aspects:

- We propose a GPT-3.5-tubo's fine-tuning scheme that utilizes a simple and explainable principle to select and obtain high accuracy for automatic scoring. This approach strives for both effectiveness and portability in evaluating and curating data subsets.
- We collected empirical evidence for how accurate the fine-tuned GPT-3.5-turbo is in scoring students' written responses to science assessments. Specifically, the study provides evidence for both multi-label and multi-class assessment tasks.
- This study compared the fine-tuned GPT-3.5-turbo model with a state-of-the-art language model (BERT) and showed the outperformance of GPT-3.5-turbo backed by the accuracy results.
- This study shows the potential of fine-tuned GPT-3.5-turbo in addressing unbalanced training data.

2 Automatic Scoring in Education

In the educational field, automatic scoring has become a cutting-edge approach to evaluating student-generated content without manual grading. This strategy is especially useful in large-scale classrooms when manual scoring is impractical (Susanti et al., 2023; Hahn et al., 2021). A critical development in learning technologies leverages the powerful machine learning algorithms (Teasley, 2017; Zhai et al., 2020). These methods were designed to provide timely information for teachers to make informed instructional decisions, as well as provide customized feedback to students to guide self-regulated learning (Gerard and Linn, 2016). Student-centered learning theories also started to be incorporated into the design of automatic scoring (Matcha et al., 2019).

An earlier attempt at automatic scoring concentrated on short responses. For example, the c-rater was introduced in (Sukkarieh and Blackmore, 2009) to score the content of such responses. The algorithm attempted to assess the quality and relevance of the content using sophisticated linguistic techniques. Researchers started looking into more complex scoring algorithms as the number and complexity of student comments increased, particularly in the context of classroom assessment practices. Aspect-level sentiment analysis was used by (Ren et al., 2023) to grade student input automatically. Their strategy allowed for a more detailed evaluation of the feedback by classifying feelings following particular facets of training. There have also been developments in the area of automatic question-generating. The rising research in this field, which has the potential to supplement automatic scoring by creating standardized, AI-generated questions, was noted in a thorough assessment by (Kurdi et al., 2020).

Additionally, recognizing the inherent variety in student responses, (Horbach and Zesch, 2019) looked into how this variability affected the automatic content scoring. Their study emphasized the significance of considering the many methods by which

students express their ideas. In the context of essay exams, a systematic evaluation by (Susanti et al., 2023) demonstrated a variety of approaches, including the use of neural networks and natural language processing, to grade lengthy responses. Meanwhile, (Fu et al., 2020) studied the advantages of AI-enabled scoring applications in promoting students' intent to study continuously and found good correlations.

Although these developments are encouraging, there are still challenges to address. For example, (Zhai et al., 2021) conducted a meta-analysis of automatic scoring and reported varied success in scoring accuracy and identified a number of factors that contribute to machine scoring accuracy. The expectations of the automatic scoring system and the actual material are frequently out of sync due to the huge diversity of student-generated content (Horbach and Zesch, 2019). Additionally, it's still difficult to ensure fairness and eliminate biases, particularly in aspect-level sentiment studies (Ren et al., 2023). Another issue that limits the general applicability of scoring systems is their flexibility across different disciplines and languages (Susanti et al., 2023).

The potential of fine-tuning language models emerges as a possible option in the face of these difficulties. Automatic scoring can be made more precise, adaptive, and context-aware by utilizing the extensive knowledge base and contextual understanding of models like ChatGPT and adding domain-specific expertise (Kasneci et al., 2023). This has the ability to overcome the current drawbacks and pave the way for an enhanced and all-encompassing automatic evaluation system for use in education. This study captures the high volume of student written responses and processed data for better results from fine-tuning ChatGPT; further, we also gather data from different demographics such as gender and race and compile the data in such a way to overcome the possible biases in automatic scoring.

3 Large Language Models for Automatic Scoring

Large language models (LLMs) have drawn increasing attention in automatic scoring because of their extensive knowledge base, context awareness, and flexibility. Different from traditional AI models, LLMs are pre-trained using large datasets. For example, Google pre-trained a Transformer model BERT using Wikipedia data, which includes 12 encoders with 12 bidirectional self-attention heads totaling 110 million parameters in its base variant and can conduct various tasks such as word sense disambiguation, natural language inference, and sentiment classification.

Automatic scoring of student-written responses is a common application of LLMs. It was shown in (Organisciak et al., 2023) that LLMs can efficiently evaluate divergent thinking problems by considering factors other than semantic distance. (Bertolini et al., 2023) used LLMs to assess the emotional content of dream reports, demonstrating the variety of applications in a more specialized application. LLMs have proven to be reliable reviewers for traditional essay scoring, a field where machine evaluation has historically been difficult owing to the range and complexity of human expression. Transformer-based models are used for this purpose in works by (Rodriguez et al., 2019) and (Ormerod et al., 2021), demonstrating LLMs' effectiveness in grading lengthy text-based responses.

Even in specialized disciplines like mathematics and science education, there is an increasing trend towards using LLMs. For instance, (Liu et al., 2023) and (Shen et al., 2021) explore BERT tactics specific to assess math and science learning, respectively. Additionally, (Wu et al., 2022) suggested adopting LLMs for automatic grading in the translation domain, highlighting the models' ability to understand linguistic nuance.

Despite their abilities, LLMs face challenges in educational settings, such as concerns about prejudice and justice (Yan et al., 2023; Zhai and Krajcik, 2024). These models may unintentionally reinforce preexisting biases because they are trained on large datasets. Furthermore, due to their generic character, they frequently miss out on the nuances of a specific topic (Caines et al., 2023). The ability to interpret the predictions or classifications made by these models is another possible challenge. The 'black-box' aspect of some LLMs can be a source of dispute because stakeholders in education frequently require transparency and openness in grading (Kasneci et al., 2023).

Moreover, most of the LLM-based automatic scoring models are based on BERT (Devlin et al., 2018), which has limitations despite being groundbreaking in its bidirectional approach to contextual embeddings. BERT is primarily intended for tasks that call for a fixed-length input, making generating sequences or handling more open-ended tasks more difficult. Because of its architecture, it is necessary to utilize laborious tokens like [SEP] (a separator token) and [CLS] (a classification token), which will be used to predict whether or not second part (A) is a sentence that directly follows first part (B). Additionally, BERT is not naturally generative, even though it can encode context in both directions. However, student-written responses are not fixed-length and sometimes not grammatically correct, making them hard to embed in BERT for automatic scoring.

As an example of LLMs, ChatGPT naturally has the benefits and drawbacks outlined above. However, many of these difficulties can be resolved as ChatGPT is open to be fine-tuned. Prior research (e.g., (Wu et al., 2023)) has presented how fine-tuned models can be made to do particular tasks, such as computerized grading in science education. More precise, equitable, and domain-appropriate automatic scoring can be achieved by aligning ChatGPT with corpora specific to a given domain and fine-tuning based on context. Furthermore, the model's dependability and credibility in educational environments can be further increased by iterative fine-tuning based on feedback loops with educators (Liu et al., 2023). Fine-tuned GPT models are more suited to tasks like text completion, response evaluation, or open-ended queries because of their autoregressive nature, which excels in sequence formation. GPT's already strong generating capacities will be strengthened by fine-tuning it for specific tasks or domains. This will enable it to produce more contextually relevant and accurate outputs, effectively solving some of the limitations of BERT.

A recent advancement to Bard (Manyika and Hsiao, 2023), Gemini Pro, a multimodal, which shows the technical competition for state-of-the-art (SOTA) AI services. Google Bard and Gemini Pro have also been tested and used for reasoning, answering knowledge-based questions, solving math problems, translating between languages, generating code, and acting as instruction-following agents through SOTA benchmarks

(Team, 2023). However, Google only released its APIs[1] on 13th December 2023 to create applications for Gemini Models; it still doesn't provide public APIs to fine-tune the models. Hence, our study has to restrict our findings to GPT3.5 as we focus on providing capabilities of fine-tuned GPT-3.5 Turbo for automatic scoring; hence, the publicly released Bard is not the right fit for the study.

4 Methodology

4.1 Datasets

This study is a secondary analysis of existing datasets, consisting of responses from middle school students with expert scoring for two multi-label assessment tasks from the PASTA project (Harris et al., 2024; PASTA, 2023) and four multi-class assessment tasks from the Mathematical Thinking in Science (MTS) project (Jin et al., 2019; ETS-MTS, 2023) responded to by high-school students. The data division based on each assessment item can be seen in Table. 1.

Table 1

Dataset information for both multi-label and multi-class tasks

Model Type	Item	No.	Labels/Classes	Training size	Testing size
Multi-label	Gas-filled balloons	5	5 binary PEs	1200	240
	Layers in test tube	10	10 binary PEs	1428	285
Multi-class	Falling weights	4	(0-3) level	1150	230
	Gelatin	5	(0-4) level	1148	226
	Bathtub	5	(0-4) level	1145	230
	Sandwater1	4	(0-3) level	1144	230

4.2 Assessment Task

4.2.1 Multi-label Assessment Tasks

The assessment tasks are designed to assess the middle school student's ability to use multi-label knowledge to explain scientific phenomena. Aiming to support students in developing knowledge-in-use across grade levels, NGSS provides performance expectations for K–12 students that integrate disciplinary core ideas (DCIs), crosscutting concepts (CCCs), and science and engineering practices (SEPs). The tasks employed in this study are aligned with the NGSS performance expectation at the middle school level: Students analyze and interpret data to determine whether substances are the same based on characteristic properties (Council et al., 2013). This performance expectation requires students to be able to use the structure and properties of matter and chemical reactions (DCIs) and patterns (CCC) to analyze and interpret data (SEP). 1200 students from grades 6-8 participated in this study. Middle school teachers in the

[1] https://ai.google.dev/tutorials/python_quickstart

Gas filled balloons (ID#: 034.02-c01)

Tap text to listen

Alice did an experiment that caused four balloons to fill with gas, as shown in the figure to the right. Alice tested the flammability of each gas. She also measured the volume and mass of each gas to calculate the density. The tests and measures all occurred under the same conditions. The data is in Table 1.

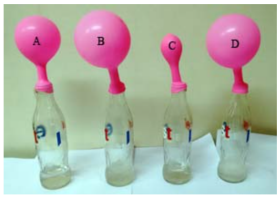


Table 1. Data of four gases in the balloons.

Sample	Flammability	Density	Volume
Gas A	Yes	0.089 g/L	180 cm^3
Gas B	No	1.422 g/L	270 cm^3
Gas C	No	1.981 g/L	35 cm^3
Gas D	Yes	0.089 g/L	269 cm^3

Question #1

Which, if any, of the gases listed in the data table could be the same? Using information from the table, explain your answer.

Please type your answer here.

Fig. 1. Example Multi-label Task: Gas-Filled Balloons

US were offered the opportunity for their students to test open-ended NGSS-aligned science tasks (Zhai et al., 2022). Student responses from this system were randomly selected as a representative participant set for the machine learning algorithms' training, validation, and testing. Student data was masked in this sample to protect student privacy; thus, no demographic information is available to researchers. However, given the geographic distribution of the teachers choosing to access the system, the data set is considered a representative cross-section of US middle school students.

These assessment tasks were selected from the Next Generation Science Assessment (Harris et al., 2024). These questions dealt with applying *chemistry* ideas in real-world situations. These tasks come under the physical sciences category of "Matter and its Characteristics," which assesses students' analytical and interpreting skills to compare different substances based on their distinctive properties. The tasks aim to examine students' multi-perspective learning and give teachers the information they may utilize to help decision-making in the classroom. The automatic reports from the students' rubric scores show what subjects they still need to learn more.

For example, In the first task, students were asked to identify gases in the experiment based on how closely their attributes match those listed in the data table (see Fig. 1). To answer this question, students have to understand the structure and properties of matter and chemical reactions and be able to use the knowledge to plan and carry out investigations and identify patterns in the data.

We developed a scoring rubric containing five response perspectives based on the dimensions of science learning: SEP+DCI, SEP+CCC, SEP+CCC, DCI, and DCI. This scoring rubric reflects multi-perspective thinking (He et al., 2024). Table 2 lists specific components for each dimension. Students would receive multiple scores automatically and simultaneously based on a multi-perspective (MPS) rating if they understood the DCIs, CCCs, or SEPs aligned with the rubric. The research team collaborated with science teachers to further validate the multi-perspective rubrics.

Table 2

Scoring rubric for task: Gas-filled balloons.

ID	Perspective	Description
E1	SEP+DCI	Student states that Gas A and D could be the same substance.
E2	SEP+CCC	Student describes the pattern (comparing data in different columns) in the table flammability data of Gas A and Gas D as the same.
E3	SEP+CCC	Student describes the pattern (comparing data in different columns) in density data of Gas A and Gas D, which is the same in the table.
E4	DCI	Student indicate flammability is one characteristic of identifying substances.
E5	DCI	Student indicate density is one characteristic of identifying substances.

4.2.2 Multi-Class Assessment Tasks

The multi-class assessment tasks ask students to solve a science problem using scientific knowledge and mathematical reasoning (e.g., proportional reasoning) (Jin et al., 2019). The tasks intend to assess the nature of mathematical reasoning in science and how students acquire this ability as they progress through high school science courses. The tasks target two NGSS disciplinary core ideas: PS3 (Energy) and LS2 (Ecosystems: Interactions, Energy, and Dynamics) at the high school level. Approximately 1400 students from grades 9-12 participated in this study, and eight teaching experts graded the responses based on the respective rubric.

To respond to these tasks, students should consider the following concepts:

1. Heat (Q) and temperature (T): Heat is an extensive variable that depends on the amount of substance (mass, M), while temperature is an intensive variable that does not depend on the amount of substance. Lower-level performing students often are confused with heat and temperature. They often think that temperature is the measure of heat (and not the average energy of molecular motion in a substance). They may also use a single undifferentiated variable that conflates the meaning of heat and the meaning of temperature to explain phenomena. In contrast, the

higher-performing students understand that heat and temperature are two different quantities.

2. Specific Heat as a compound concept: Specific heat is the amount of heat required to raise the temperature of a unit of mass of a substance by one degree Celsius. The scientific meaning of specific heat emphasizes that specific heat is a substance's internal property to regulate temperature. The material determines this internal property is the type of material or the material's microscopic/atomic/molecular structure. Specific heat is a compound concept because it results from calculating other variables. Lower-performing students may only recognize mathematical patterns in diagrams of $Q - T$ but do not understand how specific heat is related to the slope $\left(C = \frac{Q}{M\Delta T}\right)$.
3. For the same amount of Q, ΔT depends on both M and C: Heat input/output will cause the temperature of a substance to increase/decrease. For a certain amount of heat, the temperature changes depending on two factors—the amount of the substance (i.e., mass, M) and the type of the substance (i.e., heat capacity, C).

One example of a multi-class task shows a phenomenon in which a falling weight drives a paddle wheel to stir the water. Students are asked to determine whether the water will turn warm and write an explanation. For this example task, students have to identify the heat change in water based on falling weight, which causes the paddle to stir the water(see Fig. 2).

To assess students' knowledge use proficiency in responding to the example item, a four-level rubric was developed. Table 3 lists specific descriptions for each level.

4.3 Machine Algorithms

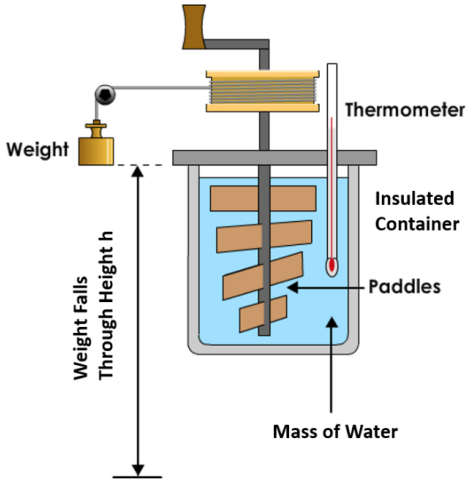
4.3.1 BERT

The cutting-edge language representation model BERT (Bidirectional Encoder Representations from Transformers) was developed by (Devlin et al., 2018) based on the Transformer model (Vaswani et al., 2017). One of its key advances is the Transformer model's self-attention mechanism, which compares the significance of several words in a phrase to a particular word. In contrast to conventional models, which usually concentrate in one direction, this method enables the model to collect context in both the forward and backward directions. Due to its bi-directionality, BERT can better comprehend the precise meaning of each word in a phrase.

BERT's in-depth comprehension of context can be quite useful for automatic scoring. For instance, BERT's capacity to comprehend complex linkages between many response components becomes essential in evaluating student failures in experimentation. The effectiveness of Large Language Models in evaluating student experimentation errors was compared with that of human raters in a study by (Bewersdorff et al., 2023). Their findings suggested that models like BERT, which have a deep awareness of context, could nearly mimic the evaluation abilities of human experts in some science education. Similar research that employed BERT for automatic scoring of student written responses in science (Riordan et al., 2020) reported robust accuracy.

Item ID: CH11_FallingWeight

A group of students conducted an experiment to study energy. In the experiment, a falling weight was attached to a rope. The rope was attached to a paddle wheel that was immersed in the water in an insulated container. When the weight falls, the paddle wheel will turn and stir the water.



1. The falling weight causes the paddle to stir the water. Do you think that will warm the water? Choose one option.
   A. Yes
   B. No
   C. Not enough information is provided.
2. Please explain your answer.

Fig. 2. Example Multi-class Task: Falling weights

Given the outstanding performance of BERT, compared to other machine learning models, in the automatic scoring of written responses in science education, we used BERT as a baseline.

4.3.2 GPT-3.5

Modern language model GPT-3.5, created by OpenAI (Brown et al., 2020) as a variation of the Generative Pre-trained Transformer (GPT) series, is a state-of-the-art LLM. The Transformer that uses the self-attention mechanism (Ghojogh and Ghodsi, 2020) to balance the importance of various words in a sequence forms the foundation of GPT-3.5's internal architecture. This model offers a potent representation of contextual relationships.

Table 3

Scoring rubric for task: Falling weights

Level	Description	Example
3	A Level 3 response suggests that the student understands the measurability of variables. Level 3 responses must choose A. There are two patterns at Level 3:	Example 1: A [Yes.] Because the weight will stir the water, the movement in water particles will cause the temperature to rise.
	3a: [Identification of heat/energy] The response 1) associates movement (e.g., movement of the weight, paddle, and/or water) with energy/heat AND/OR 2) associates particle movement with temperature, heat, or energy.	Example 2: A. [Yes.] The falling weight transfers its energy to the paddles that spin. The spinning paddles transfer their energy to the water. Because water (the system) absorbs energy, the temperature of the water will increase even if it is very small.
	3b: [Identification of heat/energy with errors.] The response 1) associates movement (weight, paddle, and/or water) with energy or heat AND/OR 2) associates particle movement with temperature, heat, or energy. But, the response also confuses energy/heat/temperature with other variables such as forces.	Example 3: A. [Yes.] The weight falling will cause the paddle to start moving, which turns gravity into kinetic energy, and the kinetic energy is then transferred to the water and moves the water's molecules, which would then start heating the water.
2	A Level 2 response:	Example 1: C [Not enough information is provided.] It depends on how fast the paddles are turning. If they are turning fast enough, they could warm the water.
	2a: [Threshold] The response indicates a threshold where the movement must be fast to a certain degree or the energy input must be large enough to cause an increase in temperature.	Example 2: B. [No.] Stirring the water will not increase the temperature of the water because no energy is being added.
	2b: [General understanding] The response indicates that the student generally understands that work/energy will cause the temperature to increase but cannot apply that general understanding to this specific scenario.	Example 3. C. [Not enough information is provided.] there is not enough information to explain more about the temperature; just will it spin, and yes, it will spin pretty fast.
	2c: [Macroscopic causation] The response provides a causal relationship at a macroscopic scale without using energy or heat.	Example 4: B. [No.] We don't know if the water is warm or cold already, room temperature, etc.
	2d: [Irrelevant variables ONLY] The response analyzes the scenario based on irrelevant variables.	Example 5: A. [Yes] Due to the friction between the paddle and the water, the water will get warmer after some time.
1	Level 1 responses have the following patterns: **1a:** [IDK] "I don't know" type of response. **1b:** [No information] The response does not provide information about the student's ideas about the phenomenon or data.	Example 1: A. [Yes.] It's placed in an insulated container, and it has a thermometer.
0	Blank or random letters	Acareba artouth

The potential applications of GPT-3.5 go beyond straightforward text production. It can adapt to various tasks without requiring tedious fine-tuning because of its few-shot learning capabilities. Translation, summarizing, answering questions, and even

programming duties fall under this category (Floridi and Chiriatti, 2020). The analysis by Floridi clarifies the nature of GPT-3.5, highlighting its ability to function as an agency even though it lacks accurate intelligence and emphasizes the creative power of such massive models. (Kung et al., 2023) demonstrated the model's success on the USMLE (United States Medical Licensing Examination) to illustrate its potential in AI-assisted medical education. Researchers can even use GPT to assist in writing academic papers in education (Zhai, 2022). GPT-3.5 is a strong choice for automatic scoring in educational settings because of its few-shot learning capacity. Traditional automatic scoring systems need a large number of labeled domain-specific datasets to fine-tune the algorithm. The requirement for domain-specific fine-tuning, however, is lessened by GPT-3.5. (Zhai, 2023b) found that GPT-3.5 shows great potential for automatically scoring written responses to science equations. The model can align its scoring pattern with human raters by comprehending the scoring requirements and evaluating responses with minimum input (Wang et al., 2022).

Additionally, because of its ability to comprehend and produce human-like language, it can provide feedback on student responses that include both a score and qualitative observations, or even customized learning guidance based on students' performance (Zhai, 2023b). GPT-3.5 and comparable models would be feasible to acquire with the democratization of huge language models espoused by (Candel et al., 2023), making AI-driven automatic scoring a common occurrence in educational institutions.

GPT-3.5's effectiveness in situations requiring automatic scoring can significantly improve with minor adjustments. GPT-3.5 can be modified to capture the nuances and intricacies of the scoring criteria by altering the model's weights on a particular scoring dataset, leading to enhanced alignment with human raters' expectations. The InstructionGPT-4 study by (Wei et al., 2023) shows how big models may be tuned using a paradigm of 200 instructions, resulting in more accurate task-specific performance. Such methods could significantly reduce scoring differences between humans and artificial intelligence (AI), guaranteeing that student responses are evaluated fairly and consistently for the GPT-3.5.

Overall, GPT-3.5 is a robust candidate for the development of autonomous scoring due to its architecture and capabilities. It has great potential to transform the field of automatic assessment thanks to its in-depth context awareness and flexibility. We have used a refined version (GPT-3.5-turbo) from the family of GPT-3.5 for fine-tuning because of its API availability. GPT-3.5-Turbo is an improved version of GPT-3 and GPT-3.5 to balance performance and efficiency. GPT-3.5-Turbo has capabilities comparable to GPT-3 but with fewer settings. This makes its use in a variety of applications affordable.

4.3.3 Domain-Specific Training

Large language models like GPT-3.5-turbo that have undergone domain-specific training can produce even more precise outcomes adapted to certain situations, topics, or scoring criteria. Text data augmentation is one creative way to use GPT-3.5-turbo's capabilities for this objective. A larger and more varied training set for scoring mechanisms can be produced using GPT-3.5 to supplement textual datasets, as demonstrated by (Cochran et al., 2023). By enhancing the model's comprehension and offering a

wider variety of sample responses, such augmentation has been proven to improve the performance of automated evaluation of student text responses.

The PandaLM benchmark was introduced by (Wang et al., 2023) to optimize domain-specific training. This evaluation environment is specifically for adjusting and optimizing LLM instructions. The benchmark enables a thorough understanding of how well the model performs across various tasks, identifying potential improvement areas and fine-tuning strategies to increase the model's effectiveness in domain-specific applications, such as automatic scoring.

To sum up, domain-specific training of GPT-3.5-turbo paves the way for more accurate and consistent automated scoring systems and ensures that the model grows more adept at comprehending the intricate aspects of particular subjects or tasks.

5 Experimental Setup

5.1 Training Scheme

For their outstanding generalization across a broad range of applications, LLMs like GPT-3.5-turbo have shown great potential. However, optimizing these models on domain-specific datasets for specialized domains like education is crucial, particularly in the context of automated scoring of student responses. Here, we outline our training scheme for optimizing GPT-3.5-turbo using academic data.

5.1.1 Data Collection and Preprocessing

A thorough dataset with student replies is necessary to start with. To guarantee a broad training set, these responses should include a range of themes, question kinds, and levels of complexity. Each response should have a human-assigned grade or score to give the model a clear supervisory signal. Following steps are applied for data collection and processing while considering OpenAI data processing guidelines[2]:

- **Data Cleaning:** Remove any irrelevant or personally identifiable information from the dataset to maintain student privacy.
- **Tokenization:** Convert the textual responses into a format suitable for the model using appropriate tokenization strategies.
- **Uploading Files:** After processing the dataset, JSON files are generated and can be uploaded to the OpenAI server using file upload API[3].

5.1.2 Model Initialization

The pre-trained GPT-3.5-turbo model needs to be loaded first. Starting with a previously trained model, such as GPT-3.5-turbo, has the advantage of understanding the language's structure, which speeds up convergence during fine-tuning.

[2]https://platform.openai.com/docs/guides/fine-tuning/preparing-your-dataset
[3]https://platform.openai.com/docs/api-reference/files

5.1.3 Fine-tuning Procedure

1. **Loss Function:** Use a regression-based loss function if the scores are continuous or a classification loss if the scores are categorical.
2. **Learning Rate:** Start with a small learning rate to ensure that the fine-tuning process doesn't diverge from the pre-trained weights too aggressively.
3. **Epochs:** Depending on the dataset's size, several epochs might be required to achieve convergence. Monitor the validation loss to prevent overfitting.
4. **Batch Size:** Choose an appropriate batch size, considering the available computational resources.

5.1.4 Evaluation and Validation

Once the model is fine-tuned, an evaluation set (distinct from the training set) is necessary to assess the model's performance.

- **Metrics:** Depending on the scoring nature, metrics like Mean Absolute Error (MAE) or classification accuracy can be used to gauge the model's scoring precision against human raters.
- **Comparison with Baseline:** Compare the fine-tuned model's performance with the original GPT-3.5-turbo to quantify the benefits of domain-specific fine-tuning.

5.2 Baselines

This study intends to investigate how the training data context affects the performance of pre-trained models (like BERT) and to investigate methods to enhance model performance further. To accomplish this goal, we use a variety of datasets to train and improve the models. First, we use 7T as the downstream task and the original BERT as the pre-trained model, which serves as the standard model. Then, using 7T as the downstream task, we train a BERT model using the dataset from each task. We contrast the two refined models to meet our goals, as the 7T is based on the context of science education.

6 Results

We performed a paired-sample T-test to compare the accuracies of fine-tuned GPT-3.5-turbo and BERT. We also calculated the p-value for the two-tailed test as we are interested in analyzing the accuracy difference from both directions. Across the trained and validated datasets yielded significant gains for GPT-3.5-turbo over BERT (Mean Difference = 0.076, SD = 0.043) conditions; $t(5) = 4.44$, $p = 0.007 < 0.05$ (Kahn, 2010). The results indicate that the average accuracy of fine-tuned GPT-3.5-turbo is significantly higher than that of the fine-tuned BERT, with 9.1% increase. A direct accuracy comparison plot can be seen in Fig. 3.

6.1 Multi-label Tasks

Two categories were evaluated for the multi-label tasks dataset: 'Gas-filled balloons' and 'Layers in test tube'.

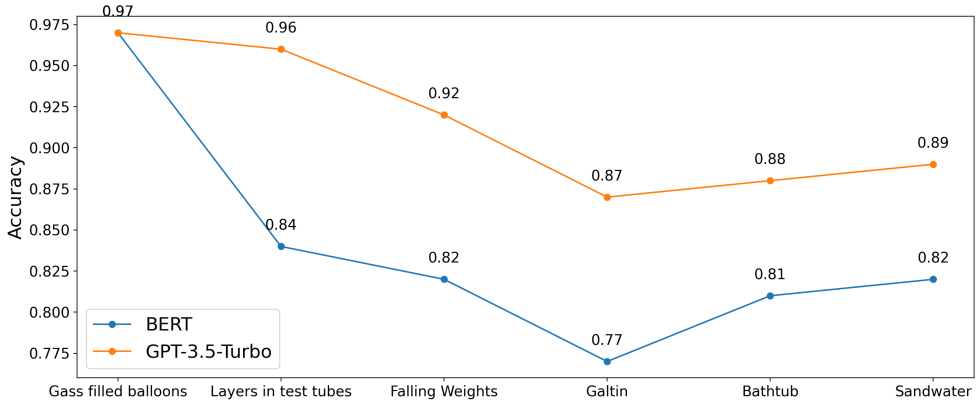


Fig. 3. Accuracy comparison of fine-tuned GPT-4.5-Turbo and BERT for different assessment tasks

- **Gas-Filled Balloons:** Both BERT and GPT-3.5-turbo achieved a remarkable accuracy of 0.97 in this multi-label classification task. No significant difference was observed between BERT and GPT-3.5-turbo, $t(2) = 0$. This suggests that the two models are equally adept at identifying and categorizing elements related to gas-filled balloons.
- **Layers in Test Tubes:** A significant difference of 0.12 was noted, with GPT-3.5-turbo outperforming BERT. The significant performance boost indicates GPT-3.5-turbo's heightened capability in grasping the complexities associated with this category.

As both tasks are multi-labeled, we also performed a paired-sample t-test to compare the individual label's accuracy for fine-tuned GPT-3.5-turbo and BERT. A significant difference was observed for both tasks. For the first task, a Mean Difference = 0.008 (SD = 0.008) of scoring accuracy between GPT-3.5-turb and BERT was observed, $t(4) = 2.14$ and $p = 0.001 < 0.05$. For the second task, which has more unbalanced labels (Mean Difference = 0.065, SD = 0.044) conditions; $t(9) = 4.67$ and $p = 0.001 < 0.05$. Upon examining the multi-label dataset, particularly the 'Layers in Test Tubes' category, it's evident that GPT-3.5-turbo consistently achieved higher accuracy on certain unbalanced labels than BERT. This consistent performance across varying label distributions underscores GPT-3.5-turbo's capability to effectively handle unbalanced labels in multi-label datasets, offering a clear advantage (7.2% increase) over models like BERT in specific scenarios.

6.2 Multi-class Tasks

We also performed a paired-sample t-test to compare the accuracy of fine-tuned GPT-3.5-tubo and BERT for all the multi-class tasks. There was a significant difference in the accuracies between the two models (Mean Difference = 0.085, SD = 0.017)

conditions; $t(4) = 9.82$ and $p = 0.0006 < 0.05$. The results suggest that GPT-3.5-turbo can predict scores for multi-class tasks significantly more accurately, with an average increase of 10.6%. Specifically,

- **Falling Weights:** GPT-3.5-turbo outperformed BERT with an accuracy of 0.92 compared to BERT's 0.82. This indicates GPT-3.5-turbo's superior ability to distinguish between different classes related to falling weights.
- **Gelatin:** Once again, GPT-3.5-turbo took the lead with an accuracy of 0.87, a notable improvement from BERT's 0.77. The results highlight GPT-3.5-turbo's refined understanding of concepts tied to gelatin.
- **Bathtub:** Both models performed commendably in this category, with GPT-3.5-turbo scoring 0.88 and BERT trailing slightly at 0.81. The performance differential suggests the advanced capabilities of GPT-3.5-turbo in discerning nuances related to bathtubs.
- **Sandwater1:** Here too, GPT-3.5-turbo surpassed BERT, scoring 0.89 against BERT's 0.82. This underscores GPT-3.5-turbo's adeptness at handling concepts associated with 'sandwater1'.

It is evident from the findings that GPT-3.5-turbo consistently outperformed BERT, particularly in the multi-class tasks. The architecture of GPT-3.5-turbo, the large amount of training data, or its innate skills to comprehend context more effectively may be responsible for its continual performance improvement. While GPT-3.5-turbo has shown higher proficiency across various educational categories, making it a preferable option for automatic scoring in certain scenarios, BERT has displayed commendable performances, particularly with "Gas-filled balloons."

7 Discussion

Automatic scoring has been long regarded as essential for effective assessment practices in education. Researchers have applied various machine learning algorithms and strategies to achieve high-scoring accuracy (Zhai et al., 2020), though with varying degrees of success (Zhai et al., 2021). The results from comparing accuracies of fine-tuned BERT and GPT-3.5-turbo in automatic scoring provide informative information about the potential of using fine-tuned GPT for automatic scoring in educational contexts. Pre-trained models can be effectively adapted to tasks specifically to a certain domain using the optimization technique known as fine-tuning. According to our results in this study, instruction-based fine-tuning can improve accuracy in tasks like automatic scoring for both multi-label and multi-class scoring tasks. Thanks to this domain-specific adaption, the fine-tuned scoring models can capture the complexities and intricacies unique to the datasets in education.

Our results provide evidence of GPT-3.5-turbo's improved performance in automatic scoring of student written responses, highlighting the technology's potential as a cutting-edge tool for educational applications. Prior research has demonstrated the potential of large-language models in automatic scoring of student written responses (Liu et al., 2023; Riordan et al., 2020), the architecture of GPT-3.5-turbo further shows its capacity to comprehend context more holistically compared to existing LLMs, and

its training strategy tailored for educational data contribute to the consistent and remarkable performance gains. The performance is consistent with expectations for ChatGPT due to its powerful language analysis and generation capacity.

The availability of fine-tuned ChatGPT is a huge advance for specific tasks such as automatic scoring in education. Our findings show the promise of GPT-3.5-turbo in education, where evaluating student responses calls for both accuracy and context knowledge. Its fine-tuned version, created especially for educational data, allows scoring procedures to be automated without sacrificing accuracy. The findings show that the GPT-3.5-turbo can be a beneficial tool in educational technology, opening the door for faster and more reliable evaluations with the proper training and fine-tuning techniques.

While the fine-tuned GPT-3.5-turbo achieved outstanding results compared to BERT, it also shows the promise of minority scoring categories in unbalanced data. Data unbalance has been a significant concern in machine training and finetuning processes (Zhai and Nehm, 2023), it is thus expected that the more advanced algorithms can contribute to this gap. Our findings show that in "Gas-filled balloons," the GPT-3.5-turbo has a clear advantage in minority scoring categories, demonstrating its prowess in understanding various educational content.

Overall, the rise of models like GPT-3.5-turbo and their potential after fine-tuning signifies a transformative shift in the automatic scoring domain. The promising results achieved by GPT-3.5-turbo, especially compared to BERT's performance, reinforce the importance of leveraging advanced models for educational applications. As models evolve, the potential for more refined and domain-specific applications in education is vast, warranting further exploration and research.

8 Conclusions and Limitations

This study highlights the potential of ChatGPT, notably GPT-3.5, for automatic scoring in science education. In our study, we focused on multi-label and multi-class assessment tasks to fine-tune GPT-3.5, utilizing varied datasets that included responses from middle-school and high-school students. The study closes the knowledge gap between the pre-trained model and the contextual demands of the educational sector. Our finding suggests that for automatic scoring accuracy, fine-tuned GPT-3.5 outperformed the refined version of Google's cutting-edge language model, BERT. This finding demonstrates GPT-3.5's effectiveness in automating scoring tasks for students' responses specific to a given domain and presents a scalable and replicable approach to utilizing this capability for various science education tasks that require pronounced accuracy. This study represents a proof-point for the expanding educational sector potential of fine-tuned language models and a lighthouse for future initiatives trying to tap into this potential for various educational applications.

Despite of the significant findings, the study has limitations. First, as a secondary analysis of existing assessment responses, researchers were constrained to the assessment process and specific student and teacher information. Future research should may focus on the implementation of automatic scoring in authentic classroom settings and examine the student learning outcomes using experimental design.

Second, we acknowledge that a boom of LLMs emerged in the market, such as the recently released Google's Gemini Pro (Team, 2023) incorporated in Bard (Manyika and Hsiao, 2023), and our study has only compared two SOTA LLMs (BERT and GPT-3.5) for fine-tuning and automatic scoring. Once the open-source and public API-provided LLMs for fine-tuning and Google APIs[4] is released, we suggest that future research should compare GPT-3.5 turbo with Gemini and other similar models for automatic scoring in education.

Third, integrating AI into educational assessment, specifically fine-tuned ChatGPT models, raises ethical dilemmas. Despite the advancement of automatic scoring (Latif and Zhai, 2023) in providing detailed feedback, poses questions about the fairness and transparency of the scoring algorithms need to be addressed (Latif et al., 2023). Future research should further study these ethical issues to better implement automatic scoring in classroom settings.

Fourth, the availability of automatic scoring has drawn concerns about teachers' roles in the era of AI-assisted education. The use of ChatGPT for providing feedback and scoring might alter the traditional role of educators, shifting their focus from assessment to more personalized guidance and support (Adiguzel et al., 2023). Furthermore, the efficiency and ease of training and uploading tests for AI-based assessment compared to traditional human marking are undeniable. However, this efficiency comes with concerns regarding the potential reduction in critical human engagement in the learning process. The reliance on AI for feedback and scoring might lead to a diminished capacity for critical thinking and personal interaction, which are crucial components of effective education (Mhlanga, 2023). Future research should examine teachers' pedagogical needs to navigate the teacher-AI relationships in educational processes to better serve student learning (Halaweh, 2023).

Last, despite the promise of automatic scoring, the use of ChatGPT in educational settings must be scrutinized for issues related to bias, privacy, and data security (Huallpa et al., 2023). Ensuring that the AI systems are free from inherent biases and that student data is handled responsibly is paramount to maintaining ethical standards in education.

Acknowledgment

This study secondary analyzed data from projects supported by the National Science Foundation (grants numbers #) and the Institute of Education Sciences (grant number). The authors acknowledge the funding agencies and the project teams for making the data available for analysis. The findings, conclusions, or opinions herein represent the views of the authors and do not necessarily represent the views of personnel affiliated with the funding agencies.

Declaration of Human Subject Approval

According to U.S. federal regulations, human subject research involves obtaining data through interaction with individuals or identifiable private information about them. If

[4] https://ai.google.dev/

the data is completely de-identified and the researcher cannot link it back to individual subjects, it may not be considered human subject research. In our case, the researchers secondarily analyzed existing de-identified data, which have undertaken IRB review. This specific research is not regarded as human subject research and thus is waived from further IRB review.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work the author(s) used ChatGPT in order to check grammar and polish the wordings. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

References

Adiguzel, T., M.H. Kaya, and F.K. Cansu. 2023. Revolutionizing education with ai: Exploring the transformative potential of chatgpt. *Contemporary Educational Technology 15*(3): ep429 .

Bertolini, L., V. Elce, A. Michalak, G. Bernardi, and J. Weeds. 2023. Automatic scoring of dream reports' emotional content with large language models. *arXiv preprint arXiv:2302.14828* .

Bewersdorff, A., K. Seßler, A. Baur, E. Kasneci, and C. Nerdel. 2023. Assessing student errors in experimentation using artificial intelligence and large language models: A comparative study with human raters. *arXiv preprint arXiv:2308.06088* .

Brown, T., B. Mann, N. Ryder, M. Subbiah, J.D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33: 1877–1901 .

Caines, A., L. Benedetto, S. Taslimipoor, C. Davis, Y. Gao, O. Andersen, Z. Yuan, M. Elliott, R. Moore, C. Bryant, et al. 2023. On the application of large language models for language teaching and assessment technology. *arXiv preprint arXiv:2307.08393* .

Candel, A., J. McKinney, P. Singer, P. Pfeiffer, M. Jeblick, P. Prabhu, J. Gambera, M. Landry, S. Bansal, R. Chesler, et al. 2023. h2ogpt: Democratizing large language models. *arXiv preprint arXiv:2306.08161* .

Cochran, K., C. Cohn, J.F. Rouet, and P. Hastings 2023. Improving automated evaluation of student text responses using gpt-3.5 for text data augmentation. In *International Conference on Artificial Intelligence in Education*, pp. 217–228. Springer.

Council, N.R. et al. 2013. Next generation science standards: For states, by states .

Dai, H., Z. Liu, W. Liao, X. Huang, Z. Wu, L. Zhao, W. Liu, N. Liu, S. Li, D. Zhu, et al. 2023. Chataug: Leveraging chatgpt for text data augmentation. *arXiv preprint arXiv:2302.13007* .

Devlin, J., M.W. Chang, K. Lee, and K. Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* .

ETS-MTS, T. November, 2023. Learning progression-based and ngss-aligned formative assessment for using mathematical thinking in science. *http://ets-cls.org/mts/index.php/assessment/* .

Floridi, L. and M. Chiriatti. 2020. Gpt-3: Its nature, scope, limits, and consequences. *Minds and Machines* 30: 681–694 .

Fu, S., H. Gu, and B. Yang. 2020. The affordances of ai-enabled automatic scoring applications on learners' continuous learning intention: An empirical study in china. *British Journal of Educational Technology 51*(5): 1674–1692 .

Gerard, L.F. and M.C. Linn. 2016. Using automated scores of student essays to support teacher guidance in classroom inquiry. *Journal of Science Teacher Education 27*(1): 111–129 .

Ghojogh, B. and A. Ghodsi. 2020. Attention mechanism, transformers, bert, and gpt: tutorial and survey .

Hahn, M.G., S.M.B. Navarro, L.D.L.F. Valentín, and D. Burgos. 2021. A systematic review of the effects of automatic scoring and automatic feedback in educational settings. *IEEE Access* 9: 108190–108198 .

Halaweh, M. 2023. Chatgpt in education: Strategies for responsible implementation .

Harris, C.J., J.S. Krajcik, and J.W. Pellegrino. 2024. *Creating and using instructionally supportive assessments in NGSS classrooms*. NSTA Press.

He, P., N. Shin, X. Zhai, and J. Krajcik. 2024. Guiding teacher use of artificial intelligence-based knowledge-in-use assessment to improve instructional decisions: A conceptual framework, In *Uses of Artificial Intelligence in STEM Education*, eds. Zhai, X. and J. Krajcik, xx–xx. Oxford University Press.

Holmes, W. and I. Tuomi. 2022. State of the art and practice in ai in education. *European Journal of Education 57*(4): 542–570 .

Horbach, A. and T. Zesch 2019. The influence of variance in learner answers on automatic content scoring. In *Frontiers in Education*, Volume 4, pp. 28. Frontiers Media SA.

Huallpa, J.J. et al. 2023. Exploring the ethical considerations of using chat gpt in university education. *Periodicals of Engineering and Natural Sciences 11*(4): 105–115 .

Jin, H., C. Delgado, M.I. Bauer, E.C. Wylie, D. Cisterna, and K.F. Llort. 2019. A hypothetical learning progression for quantifying phenomena in science. *Science & Education*: 1181–1208 .

Jin, H., P. van Rijn, J.C. Moore, M.I. Bauer, Y. Pressler, and N. Yestness. 2019. A validation framework for science learning progression research. *International Journal of Science Education 41*(10): 1324–1346 .

Kahn, J. 2010. Reporting statistics in apa style. *American Psychological Association* .

Kasneci, E., K. Seßler, S. Küchemann, M. Bannert, D. Dementieva, F. Fischer, U. Gasser, G. Groh, S. Günnemann, E. Hüllermeier, et al. 2023. Chatgpt for good? on opportunities and challenges of large language models for education. *Learning and individual differences* 103: 102274 .

Kung, T.H., M. Cheatham, A. Medenilla, C. Sillos, L. De Leon, C. Elepaño, M. Madriaga, R. Aggabao, G. Diaz-Candido, J. Maningo, et al. 2023. Performance of chatgpt on usmle: Potential for ai-assisted medical education using large language models. *PLoS digital health 2*(2): e0000198 .

Kurdi, G., J. Leo, B. Parsia, U. Sattler, and S. Al-Emari. 2020. A systematic review of automatic question generation for educational purposes. *International Journal of Artificial Intelligence in Education* 30: 121–204 .

Latif, E., G. Mai, M. Nyaaba, X. Wu, N. Liu, G. Lu, S. Li, T. Liu, and X. Zhai. 2023. Artificial general intelligence (agi) for education. *arXiv preprint arXiv:2304.12479* .

Latif, E. and X. Zhai. 2023. Automatic scoring of students' science writing using hybrid neural network. *arXiv preprint arXiv:2312.03752* .

Latif, E., X. Zhai, and L. Liu. 2023. Ai gender bias, disparities, and fairness: Does training data matter? *arXiv preprint arXiv:2312.10833* .

Lee, G.G., E. Latif, X. Wu, N. Liu, and X. Zhai. 2023. Applying large language models and chain-of-thought for automatic scoring. *arXiv preprint arXiv:2312.03748* .

Lee, G.G., L. Shi, E. Latif, Y. Gao, A. Bewersdorf, M. Nyaaba, S. Guo, Z. Wu, Z. Liu, H. Wang, et al. 2023. Multimodality of ai for education: Towards artificial general intelligence. *arXiv preprint arXiv:2312.06037* .

Limna, P., S. Jakwatanatham, S. Siripipattanakul, P. Kaewpuang, and P. Sriboonruang. 2022. A review of artificial intelligence (ai) in education during the digital era. *Advance Knowledge for Executives 1*(1): 1–9 .

Linn, M.C., L. Gerard, K. Ryoo, K. McElhaney, O.L. Liu, and A.N. Rafferty. 2014. Computer-guided inquiry to improve science learning. *Science 344*(6180): 155–156 .

Liu, J., C. Liu, R. Lv, K. Zhou, and Y. Zhang. 2023. Is chatgpt a good recommender? a preliminary study. *arXiv preprint arXiv:2304.10149* .

Liu, Z., X. He, L. Liu, T. Liu, and X. Zhai. 2023. Context matters: A strategy to pre-train language model for science education. *arXiv preprint arXiv:2301.12031* .

Manyika, J. and S. Hsiao. 2023. An overview of bard: an early experiment with generative ai. *https://ai.google/static/documents/google-about-bard.pdf* .

Matcha, W., D. Gašević, A. Pardo, et al. 2019. A systematic review of empirical studies on learning analytics dashboards: A self-regulated learning perspective. *IEEE transactions on learning technologies 13*(2): 226–245 .

Mhlanga, D. 2023. Open ai in education, the responsible and ethical use of chatgpt towards lifelong learning. *Education, the Responsible and Ethical Use of ChatGPT Towards Lifelong Learning (February 11, 2023)* .

Namatherdhala, B., N. Mazher, and G.K. Sriram. 2022. A comprehensive overview of artificial intelligence tends in education. *International Research Journal of Modernization in Engineering Technology and Science 4*(7) .

Ng, D.T.K., M. Lee, R.J.Y. Tan, X. Hu, J.S. Downie, and S.K.W. Chu. 2023. A review of ai teaching and learning from 2000 to 2020. *Education and Information Technologies 28*(7): 8445–8501 .

Organisciak, P., S. Acar, D. Dumas, and K. Berthiaume. 2023. Beyond semantic distance: automated scoring of divergent thinking greatly improves with large language models. *Thinking Skills and Creativity*: 101356 .

Ormerod, C.M., A. Malhotra, and A. Jafari. 2021. Automated essay scoring using efficient transformer-based language models. *arXiv preprint arXiv:2102.13136* .

PASTA, P.T. November, 2023. Supporting instructional decision making: Potential of an automatically scored three-dimensional assessment system. *https:/ai4stem.org/pasta/* .

Ren, P., L. Yang, and F. Luo. 2023. Automatic scoring of student feedback for teaching evaluation based on aspect-level sentiment analysis. *Education and information technologies 28*(1): 797–814 .

Riordan, B., S. Bichler, A. Bradford, J.K. Chen, K. Wiley, L. Gerard, and M.C. Linn 2020. An empirical investigation of neural methods for content scoring of science explanations. In *Proceedings of the fifteenth workshop on innovative use of NLP for building educational applications*, pp. 135–144.

Rodriguez, P.U., A. Jafari, and C.M. Ormerod. 2019. Language models and automated essay scoring. *arXiv preprint arXiv:1909.09482* .

Shen, J.T., M. Yamashita, E. Prihar, N. Heffernan, X. Wu, B. Graff, and D. Lee. 2021. Mathbert: A pre-trained language model for general nlp tasks in mathematics education. *arXiv preprint arXiv:2106.07340* .

Sukkarieh, J.Z. and J. Blackmore 2009. c-rater: Automatic content scoring for short constructed responses. In *Flairs conference*, pp. 290–295.

Susanti, M.N.I., A. Ramadhan, and H.L.H.S. Warnars. 2023. Automatic essay exam scoring system: A systematic literature review. *Procedia Computer Science* 216: 531–538 .

Tahiru, F. 2021. Ai in education: A systematic literature review. *Journal of Cases on Information Technology (JCIT) 23*(1): 1–20 .

Team, G. 2023. Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805* .

Teasley, S.D. 2017. Student facing dashboards: One size fits all? *Technology, Knowledge and Learning 22*(3): 377–384 .

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, and I. Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 .

Wang, Y., Y. Kordi, S. Mishra, A. Liu, N.A. Smith, D. Khashabi, and H. Hajishirzi. 2022. Self-instruct: Aligning language model with self generated instructions. *arXiv preprint arXiv:2212.10560* .

Wang, Y., Z. Yu, Z. Zeng, L. Yang, C. Wang, H. Chen, C. Jiang, R. Xie, J. Wang, X. Xie, et al. 2023. Pandalm: An automatic evaluation benchmark for llm instruction tuning optimization. *arXiv preprint arXiv:2306.05087* .

Wei, L., Z. Jiang, W. Huang, and L. Sun. 2023. Instructiongpt-4: A 200-instruction paradigm for fine-tuning minigpt-4. *arXiv preprint arXiv:2308.12067* .

Wei, X., X. Cui, N. Cheng, X. Wang, X. Zhang, S. Huang, P. Xie, J. Xu, Y. Chen, M. Zhang, et al. 2023. Zero-shot information extraction via chatting with chatgpt. *arXiv preprint arXiv:2302.10205* .

Wu, D., M. Wang, and X. Li. 2022. Automatic scoring for translations based on language models. *Computational Intelligence and Neuroscience* 2022 .

Wu, X., X. He, T. Liu, N. Liu, and X. Zhai 2023. Matching exemplar as next sentence prediction (mensp): Zero-shot prompt learning for automatic scoring in science education. In *International Conference on Artificial Intelligence in Education*, pp.

401–413. Springer.

Yan, L., L. Sha, L. Zhao, Y. Li, R. Martinez-Maldonado, G. Chen, X. Li, Y. Jin, and D. Gašević. 2023. Practical and ethical challenges of large language models in education: A systematic literature review. *arXiv preprint arXiv:2303.13379* .

Zhai, X. 2022. Chatgpt user experience: Implications for education. *Available at SSRN 4312418, from http://dx.doi.org/10.2139/ssrn.4312418*: 1–10 .

Zhai, X. 2023a. Chatgpt and ai: The game changer for education. *Shanghai Education*: 1–4 .

Zhai, X. 2023b. Chatgpt for next generation science learning. *XRDS: Crossroads, The ACM Magazine for Students 29*(3): 42–46 .

Zhai, X., K. C Haudek, L. Shi, R. H Nehm, and M. Urban-Lurain. 2020. From substitution to redefinition: A framework of machine learning-based science assessment. *Journal of Research in Science Teaching 57*(9): 1430–1459 .

Zhai, X., X. Chu, C.S. Chai, M.S.Y. Jong, A. Istenic, M. Spector, J.B. Liu, J. Yuan, and Y. Li. 2021. A review of artificial intelligence (ai) in education from 2010 to 2020. *Complexity* 2021: 1–18 .

Zhai, X., P. He, and J. Krajcik. 2022. Applying machine learning to automatically assess scientific models. *Journal of Research in Science Teaching 59*(10): 1765–1794 .

Zhai, X. and J. Krajcik. 2024. Pseudo artificial intelligence bias, In *Uses of Artificial Intelligence in STEM Education*, eds. Zhai, X. and J. Krajcik, xx–xx. Oxford University Press.

Zhai, X. and R.H. Nehm. 2023. Ai and formative assessment: The train has left the station. *Journal of Research in Science Teaching* .

Zhai, X., L. Shi, and R.H. Nehm. 2021. A meta-analysis of machine learning-based science assessments: Factors impacting machine-human score agreements. *Journal of Science Education and Technology* 30: 361–379 .

Zhai, X. and E. Wiebe. 2023. Technology-based innovative assessment. *Classroom-Based STEM Assessment*: 99–125 .

Zhai, X., Y. Yin, J.W. Pellegrino, K.C. Haudek, and L. Shi. 2020. Applying machine learning in science assessment: a systematic review. *Studies in Science Education 56*(1): 111–151 .