International Journal of Artificial Intelligence in Education
https://doi.org/10.1007/s40593-024-00416-y

ARTICLE

Math-LLMs: AI Cyberinfrastructure with Pre-trained Transformers for Math Education

Fan Zhang[1] · Chenglu Li[2] · Owen Henkel[3] · Wanli Xing[1] · Sami Baral[4] · Neil Heffernan[4] · Hai Li[1]

Accepted: 20 June 2024
© International Artificial Intelligence in Education Society 2024

Abstract

In recent years, the pre-training of Large Language Models (LLMs) in the educational domain has garnered significant attention. However, a discernible gap exists in the application of these models to mathematics education. This study aims to bridge this gap by pre-training LLMs on authentic K-12 mathematical dialogue datasets. Our research is structured around three primary research questions (RQs) that investigate the impact of fine-tuning data size and pre-training in downstream Natural Language Processing (NLP) tasks, and the efficacy of LLMs in text generation tasks within the mathematical context. Our findings indicate that data size plays a pivotal role in the performance of LLMs in downstream NLP tasks, with larger datasets yielding more consistent and improved results. Furthermore, pre-trained models consistently outperformed their non-pre-trained counterparts, emphasizing the importance of leveraging prior knowledge in LLMs. In the realm of text generation, we found that our model can not only enhance mathematical understanding and performance on downstream math tasks but also generate more engaging and human-like language.

Keywords LLMs · Math education · Pre-train

Introduction

Mathematics stands as not only a foundational subject but also a cornerstone in the educational landscape, critical for both academic pursuits and everyday problem-solving (D'Ambrosio, 2007). Its importance is further magnified given its applicability across diverse disciplines and real-world scenarios (Ernest et al., 2016). However, the inherently challenging nature of math, coupled with its abstract concepts, often poses significant hurdles for learners. To better support students' math learning and prepare for future workforce development, educational research in mathematics has examined one-on-one tutoring, curricular innovations, and meta-cognitive

Extended author information available on the last page of the article

Published online: 09 July 2024

skill training to effectively and successfully cater to the diverse needs of students' math learning (Callender et al., 2016; Gao et al., 2017; Kim & Lee, 2010). However, prior educational attempts to facilitate math learning mainly focus on manual support, highlighted by their limitations in scalability (e.g., supporting hundreds of thousands of students), flexibility (e.g., personalizing learning), and cost constraints (e.g., high cost of tutor training and associated service fees) (Xing & Du, 2019; Callender et al., 2016; Gao et al., 2017; Guill & Bos, 2014).

To address the aforementioned limitations in math learning studies, educators have examined automated support using emerging learning technologies such as deep neural networks to promote a paradigm shift towards supporting teaching and learning driven by artificial intelligence (AI). For instance, Hussain et al. (2019) employed deep learning to predict student performance, enabling educators to provide real-time, personalized feedback, which in turn enhances student success. Safarov et al. (2023) combined deep learning with clustering to deliver tailored e-education content, ensuring students receive content aligned with their unique learning styles. Lastly, Bannert et al. (2014) employed process mining to analyze the sequences of students' self-regulated learning activities, allowing educators to identify and address specific areas where students may struggle, thereby improving their overall learning experience. As an advancement of deep learning, Large Language Models (LLMs) enable a more accurate, contextual, and adaptive analysis of student artifacts. An LLM often is pre-trained on vast amounts of textual data,[1] allowing it to understand, generate, and interact with human language. LLMs' unparalleled precision in natural language understanding stands in stark contrast to traditional NLP tools and earlier deep learning models like Recurrent Neural Networks, which, while valuable in their time, lack the depth and breadth of understanding that LLMs demonstrate.

LLMs have been successfully adopted to provide real-time feedback, generate contextually relevant content, and facilitate interactive learning sessions (Xing et al., 2015; Li & Xing, 2021; Matelsky et al., 2023; Sallam et al., 2023). For example, Matelsky et al. (2023) introduced a tool that utilizes LLMs to provide automated feedback on open-ended questions. Their findings affirmed the tool's efficacy in delivering quick and tailored feedback, assisting students in pinpointing areas that require further attention and improvement. Moore et al. (2023) implemented LLMs to co-create educational content with students, enhancing the quality and range of learning resources. This integration has resulted in improved learning analytics and personalized educational experiences. In math learning settings, given the intricate and specialized nature of mathematics, which encompasses abstract concepts, domain-specific vocabulary, and symbolic representations, conventional models often face challenges when applied to mathematical contexts. Therefore, educators have proposed pre-trained LLMs specifically designed and fine-tuned for mathematical contexts as a solution. These LLMs utilize a wealth of mathematical data, ranging from basic arithmetic to advanced calculus, to better comprehend and process mathematical language. For instance,

[1] LLMs in our study are scoped in text-based models.

Shen et al. (2021) introduced MathBERT which was pre-trained on an extensive mathematical corpus and demonstrated superior performance in mathematical tasks compared to its counterparts without specific exposure to large corpus of mathematics. Wang et al. (2023) introduced Math-Shepherd, trained on automatically constructed supervision data, significantly enhancing the accuracy of LLMs on mathematical tasks without relying on human annotations. Additionally, Yu et al. (2023) introduced MetaMath, a fine-tuned language model specialized in mathematical reasoning, which was trained on the MetaMathQA Dataset–a collection of diverse, rephrased mathematical questions– and demonstrated superior performance on solving math word problems. Similarly, Nakamoto et al. (2023) explored enhancing automated scoring of mathematical explanations using LLM-generated datasets, showing that this semi-supervised approach significantly improves evaluation metrics.

Currently, mainstream training datasets for mathematical models can be categorized into three types: the first category uses mathematical problem-solving datasets such as GSM8K and MATH, or augmented versions of these datasets such as the MetaMathQA dataset, as exemplified by the MetaMath. The second category involves training models on scientific articles from platforms such as arXiv.org, such as MathBERT. The third category uses datasets generated by LLMs, such as in the Math-Shepherd project. These datasets have been shown to improve the mathematical understanding or performance on mathematical downstream tasks to some extent. However, few studies have focused on enhancing the quality of LLM outputs using naturalistic and authentic math learning datasets in K-12 settings. Such authenticity of math learning is important as it tends to capture common conceptual and procedural problem-solving patterns among students, teen language use, and math expert guidance (Bunch & Martin, 2021). To address the gap, we propose a unique approach: Utilizing large-scale authentic K-12 student mathematical discussions facilitated by paid math professionals for LLM pre-training. Our study's goal is bi-fold: (1) Investigate whether training our model on this authentic discussion dataset enhances its capability in understanding mathematics and handling downstream tasks and (2) whether such pre-training improves LLMs' ability to mimic the expressive ways of K-12 students, thereby producing outputs that are more natural.

Additionally, given the intense computational demands of LLMs, mainstream pre-training attempts are primarily focused on medium to small models or smaller datasets. Taking these constraints into account, this study explores the potential of leveraging recently introduced larger language models for mathematical tasks. We propose the pre-training of expansive models such as Llama, Llama-2, and GPT-J to cater to mathematical contexts. To evaluate our pre-trained models, we establish three distinct tasks: (1) single-label classification, (2) multi-label classification, and (3) text generation. In this setup, Tasks 1 and 2 aim to investigate the performance of the pre- trained model in downstream tasks, while Task 3 focuses on the text generation ability in a mathematical context. For a comprehensive comparison, we use Llama, Llama-2, and GPT-J models as our base models. These models serve as a baseline, allowing us to measure the improvements gained by our pre-trained models. Simultaneously, GPT 3.5 is also employed as the benchmark model for further evaluation and comparison. Our objective is to robustly explore the feasibility of

training models on this forum discussion dataset and assess the performance of the trained models under various potential influencing factors. We make the following contributions in this work:

(1) We build pre-trained Llama, GPT-J, and Llama-2 by pre-training the base Llama, GPT-J, and Llama-2 on authentic K-12 mathematical dialogue datasets. We publicly release pre-trained Llama, GPT-J, and Llama-2 as a community resource at: https://huggingface.co/uf-aice-lab.
(2) We evaluate the performance of pre-trained Llama, GPT-J, and Llama-2 for three general NLP tasks: (a) single-label classification, (b) multi-label classification, and (c) text generation and compare its performance to the baseline models. In our experiments, all pre-trained models demonstrated a noticeable improvement in performance across the three tasks when compared to their respective base models. To demonstrate our contribution, we have conducted research on the following research questions (RQs):

    1. To what extent does fine-tuning data size influence LLMs performance in text classification in math learning settings?
    2. To what extent does pre-training influence LLMs performance in text classification in math learning settings?
    3. To what extent does pre-training influence LLMs performance in NLP text generation tasks?

Related Work

Pre-training and fine-tuning represent the dual facets of the reusability paradigm inherent to LLMs. At its core, pre-training involves training a model on a vast corpus of text, enabling it to learn general linguistic patterns, structures, and knowledge from diverse domains. The pre-training phase typically employs objectives to predict missing words in a sentence, thereby facilitating the model's understanding of context and semantics (Devlin et al., 2019). In contrast, fine-tuning is the subsequent phase where a pre-trained model is further trained on a specific task or relatively small dataset, allowing the model to specialize and adapt its previously acquired knowledge to the nuances of the target task (Howard & Ruder, 2018). The synergy between these phases offers multiple benefits. Firstly, it capitalizes on the vast knowledge captured during pre-training, reducing the need for extensive labeled data in the fine-tuning phase. Secondly, it accelerates the training process for specific tasks, as the model starts with a robust foundational understanding. Lastly, research has shown that this two-step process often leads to state-of-the-art performance across a plethora of NLP tasks (Raffel et al., 2020).

The quantum of pre-training, or the extent to which a model is pre-trained, is influenced by a myriad of factors that can significantly shape the model's subsequent performance and adaptability. One of the most pivotal factors is the size of the data on which the model is pre-trained. Large-scale datasets encompassing diverse

linguistic patterns and domains empower the model to capture a broader spectrum of knowledge, leading to a more robust and generalized understanding (Veyseh et al., 2022). However, the benefits of increasing data size may exhibit diminishing returns, especially when the data becomes redundant or less relevant to the target tasks (Wang et al., 2020). Context, another crucial determinant, pertains to the nature and quality of the data used for pre-training. Models pre-trained on context-rich datasets that mirror real-world scenarios or specific domains tend to exhibit superior performance when fine-tuned for related tasks, as they have been exposed to relevant semantic structures and nuances (Gururangan et al., 2020). Thus, while the sheer volume of data is vital, the contextual relevance and diversity of the pre-training data play an equally, if not more, significant role in determining the efficacy of the pre-trained model.

Educational studies also emphasize the pre-training of LLMs for specific downstream tasks. So far, a niche yet significant segment of research has ventured into the pre-training of LLMs within educational contexts. For instance, Xiao et al. (2023) have ventured into the pre-training of LLMs within educational contexts, specifically in generating reading comprehension exercises for middle school students in China. Nakamoto et al. (2023) pre-trained LLMs to enhance the automated scoring of mathematical self-explanations using a semi-supervised approach that integrates LLM-generated datasets, significantly improving the model's accuracy and evaluation capabilities. Similarly, Leinonen et al. (2023) have dived into the pre-training of LLMs in enhancing programming error messages to make them more comprehensible for novice programmers, demonstrating significant improvements in the interpretability and actionability of these messages. These pioneering efforts can be categorized based on three primary dimensions: model architecture, pre-training data size, and domain specificity.

Regarding model architecture, two predominant paradigms emerge: the Masked Language Model (MLM) and the Causal Language Model (CLM). MLMs, exemplified by BERT, are trained to predict masked or concealed words within a sentence, thereby learning rich contextual representations (Devlin et al., 2019). On the other hand, CLMs, as embodied by models such as GPT, predict subsequent words in a sequence, inherently adopting an autoregressive approach (Radford et al., 2019). Recent educational research underscores the advantages of CLMs over MLMs, particularly for tasks necessitating sequential comprehension and generation, such as essay writing or narrative creation (MacAvaney et al., 2021).

In terms of pre-training data size, a salient observation in the realm of educational LLMs is the prevalent reliance on relatively smaller training datasets. For example, Ogueji et al. (2021) trained a multilingual language model on small data with just about 12,000 sentences, offering a scalable and efficient solution for linguistic diversity in NLP applications. While these datasets can be effective, they might not fully harness the expansive potential of LLMs (Zhang & Wallace, 2015). The burgeoning field of edu- cational big data accentuates the significance of leveraging larger, diverse datasets for pre-training. Such datasets ensure that models are comprehensively equipped to tackle a spectrum of educational challenges, from basic comprehension tasks to complex problem-solving.

For domain, currently, a substantial portion of pre-training endeavors is anchored in open domains or specialized fields, such as science and literature. For instance, in the domain of science education, Liu et al. (2023b) developed the SciEdBERT model by pre-training BERT and SciBERT with domain-specific data, which significantly enhanced the models' ability to automatically score scientific writings, thereby automating educational tasks with high accuracy. Niklaus and Giofré (2022) created the BudgetLongformer by training Longformer models with the Replaced Token Detection task on legal texts, achieving state-of-the-art performance on legislation summarization and demonstrating cost-effective methods for developing high-performance models in specialized fields. While domains such as science and literature are undeniably pivotal, there is a growing consensus on the imperative to diversify pre-training initiatives, ensuring they cater to a broader educational spectrum and more varied student demographics.

Experiment Setup

Pre-training Model

In this experiment, we selected Llama, Llama-2, and GPT-J models for pre-training. Their basic information can be referred to in Table 1. The reasons for selecting these models are multifaceted. First, their open-source nature ensures transparency, accessibility, and the potential for community-driven improvements. Their open-sourced nature helps democratize AI advancements, allowing researchers and developers from various backgrounds to access, modify, and build upon these models (Liu et al., 2023c). Secondly, their powerful performance is indicative of the cutting-edge advancements in the field of NLP (Touvron et al., 2023a, b; Wang & Komatsuzaki, 2022). These models have demonstrated state-of-the-art results in various benchmarks, making them prime candidates for rigorous tasks. Additionally, factors such as community support, adaptability to diverse tasks, and their scalability also played a crucial role in their selection.

Table 1 Description of models

Models	Description
Llama by Meta Touvron et al. (2023a)	Language Model Meta-AI (Llama) is introduced as a collection of foundation language models ranging from 7 to 65B parameters. Llama-2
Llama-2 by Meta Touvron et al. (2023b)	Llama-2 is developed as a family of pre-trained and fine-tuned Touvron, LLMs, with versions Llama-2 and Llama-2-Chat, available at scales up to 70B parameters. It is an updated version of Llama and variants of Llama-2 with 7B, 13B, and 70B parameters have been released
GPT-J by Eleuther AI Wang and Komatsuzaki (2022)	GPT-J 6B is a state-of-the-art transformer language model developed Wang using Ben Wang's Mesh Transformer JAX framework with 6 billion trainable parameters

Pre-training Data

We executed a comprehensive study within Algebra Nation (AN),[2] an online mathematics digital platform with more than one million active users annually. Our research involved gathering all discussion threads and responses from AN's MySQL database spanning the period from September 1, 2015, to September 1, 2021. After filtering out threads without any responses, our dataset comprised 3,097,139 discussion-response combinations, contributed by 71,918 AN participants. We defined a 'post' as the start of a continuous discussion and a 'reply' as a reaction to such a post. Interaction on AN primarily involved two groups: paid or volunteer educators and the students they assist. Of the dataset, 97% were students, and 3% were educators. Specifically, students contributed 95% of the posts and 73% of the replies, while educators accounted for 5% of the posts and 27% of the replies. This rich dataset played a crucial role in developing and assessing AI-driven tools tailored for mathematics education, as shown in Table 2. From these responses, we can observe subtle differences between human and machine replies. For example, samples 1 and 2 not

Table 2 Pre-training and task 3 data

Post		Reply
Sample 1	Can someone please help me get the answer to this problem!!! I'm stuck. Rearrange the formula $F=mg+ma$ for a	You need to get the a by itself Do you know how to do that? To get a by itself, start with subtracting mg! Try watching one of the videos to see if that could help you:)I promise their video literally ALWAYS help
Sample 2	in my class I'm learning about slope and everything that follow slope Yesterday I took a test and there was this one question that confused me. instead of $y=mx+b$ it gave me $2x+3y+6$. how do I do this to get it equal to $y=mx+b$?	do you mean $2x+3y=6$? First subtract $2\times$ on both sides Post what you get
Sample 3	when you have $3x-18=7x+63$, would it matter what you combined first?	Basically you are just getting variables on one side and constants on the other and then combining like terms
Sample 4	$45=2y+5x$?????	hey, what are you solving for? If you're solving for y subtract $5\times$ from both sides and divide by 2. If you're solving for x subtract 2y from both sides and divide by 5

[2] https://www.mathnation.com/. AN was officially renamed to Math Nation in late 2022. However, data in this study was collected in 2021, and we intend to use AN as a reference to differentiate our future studies involving the latest data.

only provide solutions but also offer educational facilitation, such as recommending videos or providing step-by-step prompts in their responses. Not all responses follow this pattern; for instance, samples 3 and 4 offer a direct answer. Nevertheless, their mode of expression seems to be more acceptable to students. Combined with the data set, we can further elaborate on our two research directions: (1) Traditional mathematical training materials, such as the GSM8K and MATH datasets, typically present mathematical concepts through a straightforward format of problems and solutions. These datasets effectively enhance models' understanding of mathematical concepts and their performance on downstream mathematical tasks, both from a human cognitive perspective and empirically demonstrated viewpoints. There will be common conceptual and procedural problem-solving patterns among students, teen language use, and expert guidance. We hypothesize that such discussions, rather than traditional materials, could potentially enhance a model's ability to grasp mathematical concepts. We primarily evaluate this through NLP downstream tasks in Task 1 and 2. (2) Typically, when posing a question to models like GPT-3.5, the model provides a direct answer using academic language. While this direct response approach can be efficient, it might not always be conducive to student learning. This type of direct response is closely related to the nature of the training datasets used in model development. However, our dataset differs significantly from these well-crafted datasets such as GSM8K and MATH which are derived from real-world students' mathematical discussions and boast a substantial volume of data. Compared to the academic language generated by models, the language used in our dataset may be more accessible and appealing to K-12 students. Therefore, we aim to explore whether this dataset not only enhances the model's ability to understand mathematics but also enables it to produce more humanlike language that engages students which is evaluated through NLP text generation tasks in Task 3.

Pre-training Methods

In our research, we adopted three distinct methodologies for pre-training our models: the conventional direct training approach, alongside the more innovative LoRA and Q-LoRA techniques. A notable observation was that when juxtaposed with LoRA and Q-LoRA, the traditional method not only exhibits a higher GPU memory consumption but also demands more resources and takes longer. Our rationale for integrating these tripartite techniques was twofold: Firstly, we were keen on ascertaining whether models pre-trained under each of these paradigms could consistently yield satisfactory and robust outcomes. Secondly, by leveraging the efficiencies of LoRA and Q-LoRA, we position ourselves to seamlessly integrate even larger and more complex models in future experiments, such as the formidable 60B Llama model. This strategic approach not only underscores our commitment to optimizing model performance but also paves the way for potential breakthroughs in large-scale model training and deployment.

LoRA (Localized Re-Adaptation) is a novel adaptation technique tailored for Large Language Models (LLMs), addressing the complexities of pre-training them

(Hu et al., 2021). This strategy emphasizes efficiency, avoiding inference latency and maintaining input sequence length. In other words, instead of training the entire model from scratch, LoRA allows us to adjust only specific parts of the model for different tasks, which saves time and resources. Q-LoRA (Quantized LoRA) takes the efficiency of LoRA a step further by reducing the amount of memory needed during training (Dettmers et al., 2024). This is done by compressing the model data into smaller, more manageable sizes without losing important information. For instance, with Q-LoRA, we can train very large models (such as those with 65 billion parameters) on a regular consumer-grade GPU, something that was previously impossible. This means more people can work with advanced models without needing expensive hardware.

Model Evaluation

Task 1 Single-label Classification

Task 1 focuses on a single-label classification problem based on the question body and answer. The data for this experiment is sourced from ASSISTments, an online learning platform that specializes in K-12 mathematics education. ASSISTments allows teachers to assign coursework and evaluate their students' performance through grading and detailed reports. The dataset consists of teacher-graded responses from students answering open-ended questions, with a total of 1,000,000 pairs of problems and responses. These pairs cover 5,699 unique mathematical problems spanning 273 different topics within the K-12 curriculum. The grading system in ASSISTments is binary, where a grade of 0 indicates an incorrect or improvement-needed response and a grade of 1 signifies a correct or acceptable answer. Specific examples of data samples are illustrated in Table 3, which showcases a selection of problem-response pairs and their corresponding correctness. The primary metric for evaluation is accuracy. For the experiment setting, 2,000 data points from the dataset were designated as the test set. The remaining data was sampled in increasing batch sizes: 100, 200, 500, 1,000, 2,000, and 5,000, with each size sampled five times. Various model architectures were fine-tuned across the training set for five epochs, followed by a comprehensive evaluation using the test set.

Table 3 Task 1 data

	Question	Correctness	Correctness
Sample 1	What is 1% of 75? Anna charges $8.50 per hour to babysit	0.75	1
Sample 2	How long will it take Anna to earn $51.00?	6	1
Sample 3	Naomi's allowance is $2.00 per week. If she convinces her parents to double her allowance each week for two months, what will her weekly allowance be at the end of the second month (week 8)?	16	0
Sample 4	Write the division expression in words: g/h + 12	The quotient of g and 12 divided by h	0

Task 2 Multi-label Classification

Task 2 is centered on classifying dialogues, with each dialogue potentially associated with multiple categories, highlighting a multi-label classification challenge. In Task 2, we utilized the paired annotations dataset from NCTE Classroom Transcript Analysis dataset (Demszky & Hill, 2023). The paired annotations dataset provides turn-level annotations for student_on_task, teacher_on_task, high_uptake, and focusing_question on 2,349 dialogues between teachers and students, using the majority rater labels. Each metric is a single label, where 1 indicates a match or presence, and 0 indicates a non-match or absence. Specific examples of data samples are illustrated in Table 4, which showcases a selection of student–teacher pairs and their corresponding annotations. The model's performance is gauged using three pivotal metrics: precision, recall, and F-1 scores (Zhang & Zhou, 2013). Precision score, ranging from 0 to 1, measures the ratio of correctly predicted positive observations to the total predicted positives. A higher precision, approaching 1, indicates fewer false positives and greater accuracy in identifying true positives. Recall, also known as sensitivity, calculates the ratio of correctly predicted positive observations to all actual positives. Its value also ranges between 0 and 1, with a higher recall indicating that the model captures a larger proportion of actual positives. F-1 Score is the harmonic mean of precision and recall, providing a balance between the two. An F-1 score closer to 1 indicates a better balance between precision and recall, ensuring that both false positives and false negatives are minimized. The experimental setup mirrors that of Experiment 1, where, due to data volume constraints, training data is segmented into batches of sizes 100, 200, 500, 1000, and 1800. 500 data entries were reserved in advance to serve as the test dataset.

Task 3 Text Generation

The Task 3 experiment is centered on a model specifically crafted to generate textual content. The experimental data for Task 3 consists of 1,000 paired discussions from AN discussion-response pairs that were reserved before pre-training and did not participate in the pre-training process. We employed all models to undertake generation and evaluation tasks on these dialogues. The assessment criteria for the generated text are both comprehensive and multifaceted: BLEURT, BERTSCORE, Readability, and Coherence Score. BLEURT evaluates the caliber of the generated text, with a particular emphasis on machine translation and other related text generation fields. Anchored in the BERT model, BLEURT's objective is to resonate more with human evaluations by harnessing the power of pre-trained contextual embeddings (Sellam et al., 2020).

$$BLEURT\ Score = \sum (BERT_{embeddings}(c_i, r_i) \cdot parameters) \tag{1}$$

where c_i and r_i are the candidate and reference sentences, respectively. Readability determines the ease with which a reader can decipher the generated text. A myriad of time-tested formulas and metrics, such as the Flesch-Kincaid Grade Level and the Gunning Fog Index, are deployed to quantify the readability of the text (DuBay, 2004). In this experiment, we used Flesch Reading Ease (Farr et al., 1951) which

Table 4 Task 2 data

	student text	teacher text	student on task	teacher on task	high uptake	focusing question
Sample 1	We could divide by, like if we put a cup on his desk and another cup on his desk and stuff like that	We could. But listen to your question, Student A. Do you think any of the estimates up here are too high or too low?	1	1	1	1
Sample 2	So, not in the envelope?	I 'm sorry, what? It does go on the envelope. We need to get moving here.It's the noise. We don 't need to do that.	0	0	0	0
Sample 3	It's something it	It 's something about it. It 's a detail, right? Remember in reading this morning, what word did we use to describe that?	1	1	0	0
Sample 4	I know how to do it, but every time I try to do that I get confused	And that's fine as long as you understand which order it goes in when you put it in the house. Yes, sir?	1	1	1	0

ensures that the content remains approachable to its target audience. The Flesch Reading Ease score is calculated as follows:

$$FRE = 206.835 - 1.015\left(\frac{total\ word}{total\ sentences}\right) - 84.6\left(\frac{total\ syllables}{total\ words}\right) \quad (2)$$

This formula outputs a number from 0 to 100, where higher scores indicate material that is easier to read. BERTSCORE delves deep into the quality of text generation by calculating the similarity between candidate and reference sentences within the BERT embedding realm (Zhang et al., 2019). Historically, BERTSCORE has showcased a strong alignment with human evaluations across a spectrum of benchmark datasets. BERTSCORE calculates the quality of text generation by computing cosine similarity within the BERT embedding space:

$$BERTSCORE = \frac{I}{N}\sum_{i=1}^{N}\max_{j}\cos\left(BERT_{embed}\left(c_i\right), BERT_{embed}\left(r_j\right)\right) \quad (3)$$

where c_i and r_j are tokens from the candidate and reference sentences, respectively, and N is the number of tokens in the candidate. Coherence Score typically evaluates the logical and structural consistency embedded within the generated text. This ensures a natural flow and logical sequence in the content, making it reader-friendly. The coherence score in our model is calculated using the *BERTScore*, which measures the semantic similarity between the generated text (candidates) and the original questions (references). The Coherence Score (F_{1coh}) using BERTScore is computed as the harmonic mean of precision (P_{coh}) and recall (R_{coh}), which are calculated from theembeddings of the candidate and reference sentences:

$$F_{1_{coh}} = 2 \times \frac{P_{coh} \times R_{coh}}{P_{coh} + R_{coh}} \quad (4)$$

where P_{coh} (precision) is the cosine similarity between each token in the candidate sentence and its closest match in the reference sentence, R_{coh} (recall) is the cosine similarity between each token in the reference sentence and its closest match in the candidate sentence. Collectively, these metrics present a well-rounded perspective on the quality, readability, and coherence of the text produced by the models in Task 3.

Experiment Workfolw

Figure 1 demonstrates how we set up and conducted the experiment.

1. We first split the full dataset into training ($n \approx 3{,}000{,}000$) and evaluating ($n = 1{,}000$) sets. Llama, Llama-2, and GPT-J were pre-trained with the training set and evaluated with the evaluation set in Task 3.
2. We used a powerful machine with 128 gigabytes of CPU RAM and eight NVIDIA A100 80G GPUs to train LLM. Python packages, PyTorch (Paszke

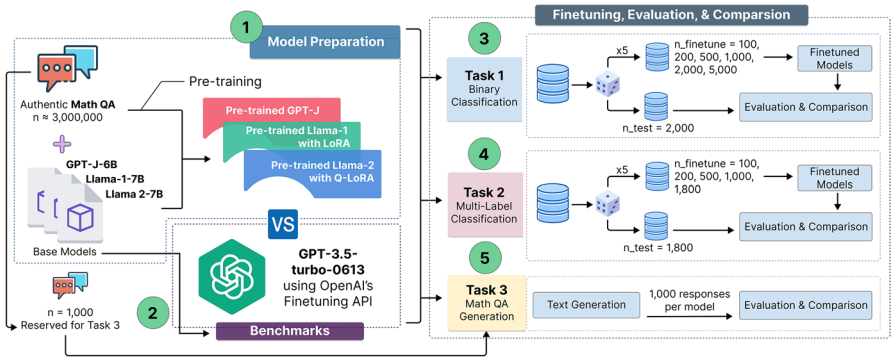

Fig. 1 Experiment workflow

et al., 2019) and HuggingFace's Transformers (Wolf et al., 2020) were used to provide the computing infrastructure with GPUs. During the pre-training phase, we used an initial learning rate of 1e-4 and trained on the complete dataset for 3 epochs using the Adam optimizer. Additionally, we employed a batch size of 16, adjustable between 8 and 32 depending on the model size.

3. We evaluated and compared the pre-trained models, base models, and benchmark models. For this purpose, we designed three specific tasks for these models. Tasks one and two are downstream classification tasks. We first fine-tuned all models on datasets of varying sizes and then evaluated and compared them on a unified evaluation dataset. In the fine-tuning phase, we similarly used an initial learning rate of 1e-4 and trained on the complete dataset for 5 epochs using the Adam optimizer. The model largely converged and remained stable between the 3rd and 5th epochs. Task 3 is a text generation task. We used the 1,000 paired discussion data set retained from phase 1 to generate, evaluate, and compare all models. For detailed task content and data descriptions, please refer to the subsequent sections.

Results

The experimental results for Task 1 are shown in Table 5,[3] and Fig. 2 illustrates the variation in accuracy of different models for Task 1 as the number of fine-tune instances changes. Based on the chart and table, we can observe that GPT-3.5 leads in this task, achieving the best predictive accuracy across different fine-tune instances. The results for Task 2 are presented in Table 6. Considering the composite

[3] Additionally, we have examined tasks 1 and 2 using a bag-of-words method with SVM, linear regression, and RandomForest, trained with the max sample size in each task (n task1 = 5000, n task2 = 1800). Their results yielded accuracies around 0.75 in task 1 and 0.45 in task 2. Since these two tasks focus on investigating the effectiveness of pretraining LLMs with authentic mathematical learning data, results of these traditional NLP models are not included in the results and discussion.

Table 5 Task 1 single-label classification outcomes

Fine-tune Instances	100	200	500	1000	2000	5000
GPT-J	0.61 ± 0.02	0.60 ± 0.04	0.66 ± 0.01	0.72 ± 0.01	0.77 ± 0.01	0.84 ± 0.01
Pre-trained-GPT-J	0.61 ± 0.02	0.61 ± 0.03	0.67 ± 0.01	0.72 ± 0.01	0.78 ± 0.02	0.85 ± 0.00
Llama	0.59 ± 0.04	0.60 ± 0.04	0.63 ± 0.01	0.67 ± 0.01	0.69 ± 0.03	0.73 ± 0.05
Pre-trained-Llama	0.61 ± 0.02	0.62 ± 0.03	0.63 ± 0.01	0.68 ± 0.03	0.72 ± 0.03	0.74 ± 0.04
Llama-2	0.61 ± 0.02	0.61 ± 0.02	0.65 ± 0.02	0.71 ± 0.02	0.72 ± 0.05	0.81 ± 0.06
Pre-trained-Llama-2	0.60 ± 0.03	0.60 ± 0.04	0.65 ± 0.02	0.67 ± 0.03	0.75 ± 0.02	0.83 ± 0.02
GPT-3.5	**0.71 ± 0.03**	**0.72 ± 0.02**	**0.79 ± 0.02**	**0.85 ± 0.01**	**0.90 ± 0.00**	**0.94 ± 0.01**

Bolded values represent the best results in the respective columns

metrics of precision, recall, and F-1 score, we notice that for Task 2, GPT-3.5 loses its leading position. Instead, Pre-trained GPT-J excels on relatively smaller fine-tune datasets, while pre-trained Llama-2 performs better on larger datasets. The results for Task 3 are stored in Table 7, and Fig. 3 shows the bar distribution of various evaluation metrics and models for Task 3. Through the experimental data, we can observe that Pre-trained Llama-2 achieves the best performance. Furthermore, by synthesizing the results from all three tasks, we can see that the pre-trained models generally outperform the non-pre-trained models. Additionally, as the number of fine-tune instances for the task increases, all models show a certain degree of improvement in their performance on that task.

Discussion and Conclusion

Educational research focusing on the pre-training of Large Language Models (LLMs) has predominantly targeted disciplines like science and literature, as highlighted by McNamara et al. (2017), while mathematics education has often been overlooked. This oversight is critical because mathematics possesses unique

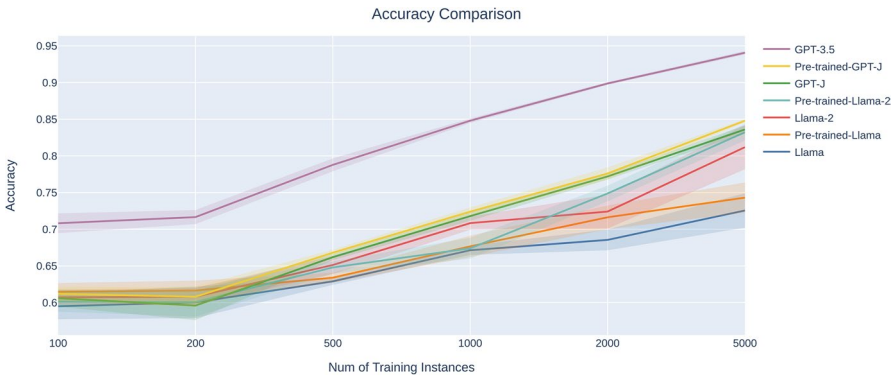

Fig. 2 Task 1 single-label classification accuracy

Table 6 Task 2 multi-label classification outcomes

Pre-training Data Instances	100			200		
Evaluation Metrics	precision	recall	f-1	precision	recall	f-1
GPT-3.5	0.762 ± 0.033	0.690 ± 0.065	0.708 ± 0.041	0.754 ± 0.019	0.720 ± 0.010	0.722 ± 0.013
GPT-J	0.782 ± 0.044	**0.760 ± 0.024**	**0.754 ± 0.019**	0.796 ± 0.017	0.744 ± 0.030	0.754 ± 0.015
Pre-trained-GPT-J	**0.796 ± 0.034**	0.742 ± 0.050	0.748 ± 0.016	**0.810 ± 0.010**	**0.744 ± 0.021**	**0.762 ± 0.008**
Llama	0.776 ± 0.062	0.688 ± 0.057	0.726 ± 0.023	0.780 ± 0.072	0.734 ± 0.037	0.740 ± 0.019
Pre-trained-Llama	0.794 ± 0.030	0.746 ± 0.019	0.746 ± 0.015	0.808 ± 0.046	0.734 ± 0.036	0.754 ± 0.013
Llama-2	0.764 ± 0.026	0.740 ± 0.019	0.736 ± 0.013	0.786 ± 0.009	0.726 ± 0.025	0.736 ± 0.015
Pre-trained-Llama-2	0.774 ± 0.027	0.744 ± 0.033	0.742 ± 0.013	0.794 ± 0.015	0.742 ± 0.008	0.750 ± 0.012
Fine-tuning Data Size	500			1000		
Evaluation Metrics	precision	recall	f-1	precision	recall	f-1
GPT-3.5	0.772 ± 0.011	0.742 ± 0.011	0.744 ± 0.011	0.776 ± 0.005	0.760 ± 0.010	0.756 ± 0.005
GPT-J	0.794 ± 0.011	0.730 ± 0.027	0.744 ± 0.015	0.796 ± 0.009	0.758 ± 0.018	0.762 ± 0.013
Pre-trained-GPT-J	0.802 ± 0.008	**0.752 ± 0.008**	**0.758 ± 0.008**	**0.806 ± 0.009**	0.760 ± 0.010	0.768 ± 0.008
Llama	0.746 ± 0.042	0.748 ± 0.022	0.738 ± 0.026	0.774 ± 0.018	0.748 ± 0.016	0.756 ± 0.009
Pre-trained-Llama	0.774 ± 0.027	0.746 ± 0.039	0.742 ± 0.026	0.782 ± 0.011	0.756 ± 0.017	0.760 ± 0.010
Llama-2	0.784 ± 0.015	0.742 ± 0.011	0.746 ± 0.015	0.792 ± 0.015	0.750 ± 0.017	0.756 ± 0.015
Pre-trained-Llama-2	**0.802 ± 0.004**	0.742 ± 0.008	0.756 ± 0.005	0.804 ± 0.009	**0.766 ± 0.011**	**0.770 ± 0.014**
Fine-tuning Data Size	1800					
Evaluation Metrics	precision	recall	f-1			
GPT-3.5	0.770 ± 0.007	0.752 ± 0.008	0.754 ± 0.009			
GPT-J	0.794 ± 0.005	0.760 ± 0.000	0.762 ± 0.004			
Pre-trained-GPT-J	**0.808 ± 0.008**	0.764 ± 0.005	**0.772 ± 0.004**			
Llama	0.794 ± 0.013	0.750 ± 0.016	0.758 ± 0.005			
Pre-trained-Llama	0.796 ± 0.011	0.758 ± 0.012	0.762 ± 0.008			
Llama-2	0.796 ± 0.009	0.760 ± 0.010	0.762 ± 0.011			
Pre-trained-Llama-2	0.804 ± 0.009	**0.768 ± 0.008**	0.772 ± 0.011			

Bolded values represent the best results in the respective columns

Table 7 Task 3 Text generation outcomes

Evaluation Metrics	BLEURT	BERTSCORE	Readability	Coherence
Llama	0.33	0.421	**79.327**	0.441
Pre-trained-Llama	0.388	0.511	79.147	0.432
Llama-2	0.343	0.457	76.642	0.470
Pre-trained-Llama-2	**0.394**	**0.609**	76.646	0.480
GPT-J	0.320	0.427	77.621	0.381
Pre-trained-GPT-J	0.355	0.475	78.595	0.439
GPT-3.5	0.354	0.366	69.627	**0.502**

Bolded values represent the best results in the respective columns

characteristics such as sequential reasoning, the use of symbolic notations, and abstract concepts, which are not as prevalent in other disciplines. In addition, further delving into the pre-training process, Zhang and Wallace (2015) observed that most educational datasets used for pre-training are relatively small, which restricts the models' ability to capture and understand nuanced semantic details, particularly within mathematical contexts.

To address these challenges and better serve K-12 mathematics education, our strategy involves pre-training on vast authentic K-12 mathematical dialogue datasets. This approach aims to enhance the models' comprehension of mathematics by exposing them to realistic educational interactions. We have structured our assessment regimen around three distinct tasks, each utilizing different evaluation metrics. This approach ensures a thorough evaluation of the models' capabilities, allowing us to assess their performance in mathematical reasoning, problem-solving and

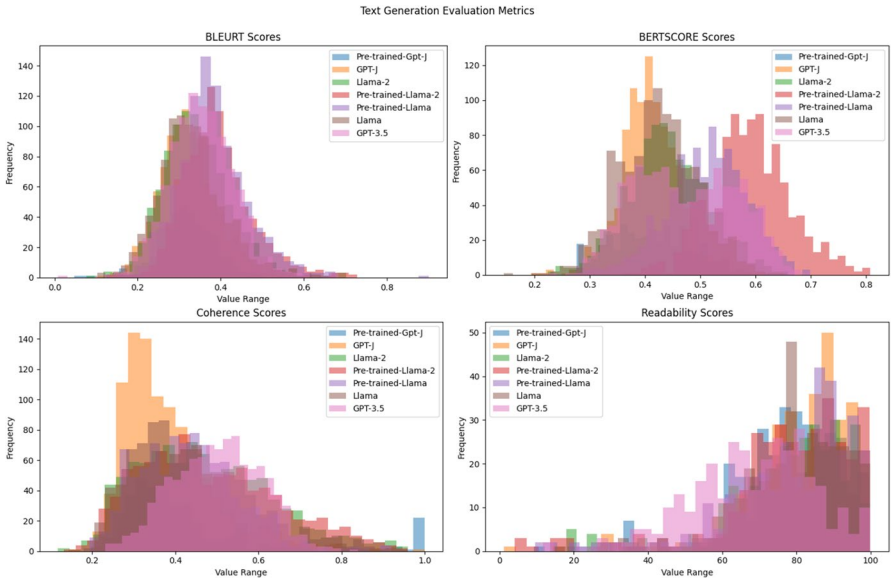

Fig. 3 Task 3 text generation evaluation metrics distribution

conversation. By doing so, we aim to bridge the gap in LLM applications within the field of mathematics education, providing more robust tools for educators and learners.

RQ1: To What Extent does Fine-tuning Data Size Influence LLMs Performance in Text Classification in Math Learning Settings?

The findings from Task 1 and Task 2 provide insightful evidence regarding the impact of dataset size on the performance of Large Language Models (LLMs) in various natural language processing (NLP) downstream tasks. Specifically, we observed that as the dataset size increased, the performance of both pre-trained and non-pre-trained models such as GPT-J, Llama, Llama-2, and GPT-3.5 consistently improved. This improvement across models suggests that larger datasets enable the models to learn from a broader range of examples, enhancing their ability to generalize across different tasks and scenarios.

Contrary to the phenomenon of 'Diminishing Returns' reported in previous studies like that by Wang et al. (2020), where increases in data size eventually lead to smaller incremental improvements in model performance, our results did not align with this pattern. Instead, we found no clear point at which adding more data ceased to contribute significantly to model improvement. This divergence could be due to the specific cap we set on the dataset size—5,000 examples—which might not have been large enough to reach the threshold where diminishing returns typically set in. It is possible that with LLMs, especially the newer iterations such as GPT-3.5, the threshold for diminishing returns is much higher, suggesting that these models have a greater capacity for learning from larger volumes of data without experiencing saturation.

In addition to these findings, we also discovered that dataset size has a substantial impact on the stability of LLMs in performing NLP downstream tasks. Analysis of the data from Task 2 revealed a noteworthy trend: as the dataset size increased, the variability in model performance, as indicated by the standard deviation of performance metrics, tended to decrease. This suggests that larger datasets not only enhance model performance but also contribute to more consistent and reliable outputs. Such stability is crucial for practical applications where predictability and dependability are key. This observation aligns with findings from recent research by Naveed et al. (2023), which also underscored the importance of robust training datasets for achieving performance consistency.

In Task 1, while the trend of improved performance with larger datasets was evident in models like GPT-3.5 and GPT-J, Llama and Llama-2 did not consistently show this behavior. This inconsistency could potentially be attributed to insufficient training iterations for these particular models, suggesting that they may require more extensive training to fully exploit larger datasets. This insight points to the need for tailored training strategies that consider the specific characteristics and requirements of different LLMs to optimize their learning from expanded datasets effectively.

RQ2: To What Extent does Pre-training Influence LLMs Performance in Text Classification in Math Learning Settings?

From the experimental data gathered in Task 1 and Task 2, it was evident that nearly all pre-trained versions of the models demonstrated superior performance compared to their non-pre-trained counterparts across various training scenarios. This marked difference clearly illustrates the crucial role that pre-training plays in enhancing the capabilities of Large Language Models (LLMs) for tackling NLP downstream tasks, which aligns with findings in the field, such as those presented by Ladhak et al. (2023) and Liu et al. (2023a). Both studies affirm the pivotal importance of pre-training in preparing LLMs for successful deployment in NLP downstream tasks. According to these researchers, pre-training not only enhances model performance but also contributes to a more stable and reliable adaptation to task-specific demands. This body of research supports the conclusion that pre-training is not merely beneficial but rather essential for achieving optimal performance in complex NLP downstream tasks, providing a compelling case for its continued and expanded use in developing state-of-the-art language models.

In addition, building on the robust findings from our experiments, we can confidently assert that with sufficient data, pre-training utilizing forum discussion datasets within a mathematical context could enhance a model's ability to comprehend math- ematics concepts and improve its performance on downstream mathematics tasks to some extent. Same with Feng et al. (2021), we believe that the enhanced performance of models that underwent pre-training underscores the value of incorporating extensive prior knowledge and a robust base of general language understanding. This preparatory step seems to effectively prime the models, equipping them with a depth of mathematics linguistic insight that is lacking in models without such a foundation. When these pre-trained models are subsequently fine-tuned for specific tasks, they are not starting from scratch but are instead refining and adapting an already rich linguistic framework to meet particular requirements. This dual approach of broad initial training followed by targeted refinement ensures that the models not only grasp the nuances of specific tasks more effectively but also apply their pre-acquired language skills in a focused manner.

Finally, When pre-training large language models (LLMs) using grade-level specific datasets, it is crucial to ensure that the training data is diverse and representative of the various educational stages within K-12. Grade-specific datasets may contain nuanced language, context, and knowledge that vary significantly across different educational levels. Ensuring that these datasets capture a wide range of topics, linguistic structures, and contextual variations is essential for the model to generalize well across all K-12 grades. For example, in our experiment, we used a dataset of over 3 million entries that effectively covered all K-12 grade levels. This extensive dataset ensures that the model can perform well across various K-12 math tasks, providing a strong foundation for its generalization capabilities. Also, the alignment of pre-training objectives with downstream tasks is essential. We believe that through pre-training, the model can enhance its understanding of mathematical concepts, including formulas and problem-solving methods. Consequently, the trained model is expected to achieve better performance on related math tasks. Finally, continuous evaluation and

iterative refinement are necessary to ensure that the model performs robustly across all grade levels. Regular assessments on diverse and representative datasets from different grades can help identify potential biases or gaps in the model's knowledge, enabling targeted improvements. By addressing these considerations, we can develop LLMs that not only excel in specific grade-level tasks but also provide consistent and reliable performance across the entire K-12 educational spectrum.

RQ3: To What Extent does Pre-training Influence LLMs Performance in NLP Text Generation Tasks

The experimental results from Task 3 provide significant insights into the performance enhancement of pre-trained models compared to their non-pre-trained counterparts across various evaluation metrics such as BLEURT, BERTScore, readability, and coherence. Focusing on benchmark model GPT-3.5, its lower scores in BLEURT and BERTScore suggest a notable deviation between the generated texts and the reference texts. This trend is consistent across all models evaluated, including GPT-J and Llama variants, and aligns with our expectations. The nature of the datasets used—predominantly student dialogues in mathematics settings—likely contributes to this outcome. Unlike straightforward answer generation, these dialogues encompass a richer, more human-like interaction, which poses a greater challenge for text generation models. These interactions often require the models to grasp subtler nuances of conversational language, which are crucial for generating responses that are not only correct but also contextually rich and engaging. In addition, the comparison of models before and after pre-training revealed that while there was no significant change in readability and coherence, suggesting that the basic understanding and structure of the text were maintained, there was a certain degree of improvement in BLEURT and BERTScore. This indicates that the pre-training has indeed enhanced the models' ability to generate more human-like responses. This phase of our research, is still preliminary, we do not expect the models to fully mimic human conversational abilities at this point. However, it provides encouraging evidence that our pre-training approach is valid and effective. Looking ahead, we plan to refine our strategies further by incorporating actual teacher responses into the training process. By doing so, we hope to leverage the nuanced, pedagogical insights that experienced educators bring to discussions, thereby enriching the models' ability to generate not just correct, but pedagogically sound and engaging mathematics discussions.

Implication

The exploration into mathematics discussions, as opposed to traditional methods, has profound implications for the field of NLP and education. By emphasizing contextual discussions and guided problem-solving, models can be trained to generate responses that are not only accurate but also pedagogically beneficial. This approach can revolutionize online learning platforms, where the emphasis is often on rote learning and direct answers. In traditional online learning platforms, students are often provided with direct solutions without much context or guidance on the problem-solving

process. This can lead to superficial understanding and reliance on memorization. However, with models trained on mathematics discussions and guided problem-solving, learners can be taken through a step-by-step process, mirroring the guidance a human tutor might provide. This not only aids in understanding the current problem but also equips learners with the skills and strategies to tackle similar problems in the future. Meanwhile, the findings of this research emphasize the pivotal role of the dataset's nature in determining the behavior and performance of NLP models. It's not just about the sheer volume of data but the quality, context, and structure that matter. The content of the training data, its organization, and the nuances it carries can significantly influence the model's understanding and subsequent outputs. For instance, a model trained on data that emphasizes contextual discussions and guided problem-solving will likely generate responses that mirror such guidance, as opposed to one trained on direct answers. In light of these insights, it becomes evident that the curation and selection of training data are not merely preliminary steps but are central to achieving desired model outcomes and behaviors. It underscores the need for meticulous attention to the content, structure, and quality of datasets used in training NLP models.

Limitations and Future Work

While our findings offer a promising outlook, certain limitations warrant attention. First and foremost, our exploration of Task 3 might not have been exhaustive enough. Generating larger volumes of text could potentially lead to more rigorous generalizations. Moreover, our reliance on computational evaluation metrics in Task 3, while valuable, might not capture the full spectrum of the model's capabilities. Introducing human expert evaluations could offer a more nuanced and persuasive assessment of the model's performance. The other significant area of focus is the quality and structure of the pre-training dataset. While the depth and integrity of the dataset are undeniably crucial for model performance, our current pre-training dataset, though vast, offers room for expansion. There's potential to bolster the fine-tuning samples, either by sourcing from a broader range or by employing grammatical augmentation techniques. Enhancing the dataset's quality is equally vital, and tools like error detection and data filtering could be invaluable in this regard. Our future research endeavors will prioritize both the volume and quality of data, ensuring optimal pre-training and fine-tuning. Moreover, we plan to introduce a rigorous manual inspection and enhancement process for the pre-training dataset, aiming to guarantee not just data quality but also its relevance to the tasks in focus. Lastly, the presence of inherent biases in the dataset, particularly those stemming from personal opinions, raises ethical concerns (Zhang et al., 2023; Li et al., 2024; Song et al., 2024). It becomes paramount for subsequent studies to employ interpretable machine learning techniques to assess model fairness. Additionally, the development of text preprocessing methodologies to mitigate potential ethical challenges will be crucial.

Data Availability The data that support the findings of this study are available from the corresponding author, WX, upon reasonable request.

References

Bannert, M., Reimann, P., & Sonnenberg, C. (2014). Process mining techniques for analysing patterns and strategies in students' self-regulated learning. *Metacognition and Learning, 9*, 161–185.

Bunch, G. C., & Martin, D. (2021). From "academic language" to the "language of ideas": A disciplinary perspective on using language in k-12 settings. *Language and Education, 35*(6), 539–556.

Callender, A. A., Franco-Watkins, A. M., & Roberts, A. S. (2016). Improving metacognition in the classroom through instruction, training, and feedback. *Metacognition and Learning, 11*, 215–235.

D'Ambrosio, U. (2007). The role of mathematics in educational systems. *ZDM Mathematics Education, 39*, 173–181.

Demszky, D., & Hill, H. (2023). The NCTE Transcripts: A Dataset of Elementary Math Classroom Transcripts. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)* (pp. 528-538).

Dettmers, T., Pagnoni, A., Holtzman, A., & Zettlemoyer, L. (2024). Qlora: Efficient finetuning of quantized llms. *Advances in Neural Information Processing Systems, 36*.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Long and Short Papers)* (Vol. 1, pp. 4171–4186).

DuBay, W. H. (2004). The principles of readability. *Online Submission*.

Ernest, P., Skovsmose, O., Van Bendegem, J. P., Bicudo, M., Miarka, R., Kvasz, L., & Moeller, R. (2016). *The philosophy of mathematics education*. Springer Nature.

Farr, J. N., Jenkins, J. J., & Paterson, D. G. (1951). Simplification of flesch reading ease formula. *Journal of Applied Psychology, 35*(5), 333.

Feng, Y., Jiang, J., Tang, M., Jin, R., & Gao, Y. (2021). Rethinking Supervised Pre-Training for Better Downstream Transferring. In *International Conference on Learning Representations*.

Gao, Y., Zhang, P. P., Wen, S. F., & Chen, Y. G. (2017). Challenge, opportunity and development: Influencing factors and tendencies of curriculum innovation on undergraduate nursing education in the mainland of china. *Chinese Nursing Research, 4*(3), 113–116.

Guill, K., & Bos, W. (2014). Effectiveness of private tutoring in mathematics with regard to subjective and objective indicators of academic achievement. *Journal for Educational Research Online, 6*(1), 34–67.

Gururangan, S., Marasović, A., Swayamdipta, S., Lo, K., Beltagy, I., Downey, D., & Smith, N. A. (2020). Don't Stop Pretraining: Adapt Language Models to Domains and Tasks. *In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 8342-8360).

Howard, J., & Ruder, S. (2018). Universal language model fine-tuning for text classification. In *ACL 2018-56th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers)* (Vol. 1, pp. 328-339). Association for Computational Linguistics.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., et al. (2021). Lora: Low- rank adaptation of large language models. arXiv preprint arXiv:2106.09685.

Hussain, S., Muhsin, Z., Salal, Y., Theodorou, P., Kurtoğlu, F., & Hazarika, G. (2019). Prediction model on student performance based on internal assessment using deep learning. *International Journal of Emerging Technologies in Learning, 14*(8).

Kim, S., & Lee, J.-H. (2010). Private tutoring and demand for education in south korea. *Economic Development and Cultural Change, 58*(2), 259–296.

Ladhak, F., Durmus, E., Suzgun, M., Zhang, T., Jurafsky, D., McKeown, K., & Hashimoto, T. B. (2023). When do pre-training biases propagate to downstream tasks? a case study in text summarization. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistic*s (pp. 3206-3219).

Leinonen, J., Hellas, A., Sarsa, S., Reeves, B., Denny, P., Prather, J., & Becker, B. A. (2023). Using large language models to enhance programming error messages. *Proceedings of the 54th ACM Technical Symposium on Computer Science Education V. 1*, 563–569.

Li, C., & Xing, W. (2021). Natural language generation using deep learning to support MOOC learners. *International Journal of Artificial Intelligence in Education, 31*, 186–214.

Li, C., Xing, W., & Leite, W. (2024). Using fair AI to predict students' math learning outcomes in an online platform. *Interactive Learning Environments, 32*(3), 1117–1136.

Liu, H., Xie, S. M., Li, Z., & Ma, T. (2023a). Same pre-training loss, better down- stream: Implicit bias matters for language models. *International Conference on Machine Learning*, 22188–22214.

Liu, Z., He, X., Liu, L., Liu, T., & Zhai, X. (2023b). Context matters: A strategy to pre-train language model for science education. *International Conference on Artificial Intelligence in Education*, 666–674.

Liu, Z., Qiao, A., Neiswanger, W., Wang, H., Tan, B., Tao, T., Li, J., Wang, Y., Sun, S., Pangarkar, O., et al. (2023c). Llm360: Towards fully transparent open-source llms. arXiv preprint arXiv:2312.06550.

MacAvaney, S., Macdonald, C., Murray-Smith, R., & Ounis, I. (2021). IntenT5: Search Result Diversification using Causal Language Models. arXiv e-prints, arXiv-2108.

Matelsky, J. K., et al. (2023). A large language model-assisted education tool to provide feedback on open-ended responses. arXiv preprint arXiv:2308.02439.

McNamara, D. S., Allen, L. K., Crossley, S. A., Dascalu, M., & Perret, C. A. (2017). Natural Language Processing and Learning Analytics. *Grantee Submission*.

Moore, S., Tong, R., Singh, A., Liu, Z., Hu, X., Lu, Y., Liang, J., Cao, C., Khosravi, H., Denny, P., et al. (2023). Empowering education with llms-the next- gen interface and content generation. *International Conference on Artificial Intelligence in Education*, 32–37.

Nakamoto, R., Flanagan, B., Yamauchi, T., Dai, Y., Takami, K., & Ogata, H. (2023). Enhancing automated scoring of math self-explanation quality using llm-generated datasets: A semi-supervised approach. *Computers, 12*(11), 217.

Naveed, H., Khan, A. U., Qiu, S., Saqib, M., Anwar, S., Usman, M., et al. (2023). A comprehensive overview of large language models. arXiv preprint arXiv:2307.06435.

Niklaus, J., & Giofr´e, D. (2022). Budgetlongformer: Can we cheaply pretrain a sota legal language model from scratch? arXiv preprint arXiv:2211.17135.

Ogueji, K., Zhu, Y., & Lin, J. (2021). Small data? no problem! exploring the viability of pretrained multilingual language models for low-resourced languages. In *Proceedings of the 1st Workshop on Multilingual Representation Learning* (pp. 116-126).

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., ... & Chintala, S. (2019). Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems, 32*.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI Blog, 1*(8), 9.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., et al. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research, 21*(1), 5485–5551.

Safarov, F., Kutlimuratov, A., Abdusalomov, A. B., Nasimov, R., & Cho, Y. I. (2023). Deep learning recommendations of e-education based on clustering and sequence. *Electronics, 12*(4), 809.

Sallam, M., et al. (2023). Chatgpt applications in medical, dental, pharmacy, and public health education: A descriptive study highlighting the advantages and limitations. *Narra J, 3*(1), e103–e103.

Sellam, T., Das, D., & Parikh, A. P. (2020). Bleurt: Learning robust metrics for text generation. arXiv preprint arXiv:2004.04696.

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

Song, Y., Li, C., Xing, W., Li, S., & Lee, H. H. (2024, March). A Fair Clustering Approach to Self-Regulated Learning Behaviors in a Virtual Learning Environment. In *Proceedings of the 14th Learning Analytics and Knowledge Conference* (pp. 771-778).

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M. A., Lacroix, T., ... & Lample, G. (2023a). Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., et al. (2023b). Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288.

Veyseh, A. P. B., Meister, N., Yoon, S., Jain, R., Dernoncourt, F., & Nguyen, T. H. (2022). Macronym: A large-scale dataset for multilingual and multi-domain acronym extraction. arXiv preprint arXiv:2202.09694.

Wang, S., Khabsa, M., & Ma, H. (2020). To Pretrain or Not to Pretrain: Examining the Benefits of Pretrainng on Resource Rich Tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 2209-2213).

Wang, B., & Komatsuzaki, A. (2022). GPT-J-6B: a 6 billion parameter autoregressive language model (2021). URL https://github.com/kingoflolz/mesh-transformer-jax.

Wang, P., Li, L., Shao, Z., Xu, R. X., Dai, D., Li, Y., ... & Sui, Z. (2023). Math-shepherd: Verify and reinforce llms step-by-step without human annotations. *CoRR, abs/2312.08935*.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., ... & Rush, A. M. (2020). Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations* (pp. 38-45).

Xiao, C., Xu, S. X., Zhang, K., Wang, Y., & Xia, L. (2023). Evaluating reading com- prehension exercises generated by llms: A showcase of chatgpt in education applications. *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, 610–625.

Xing, W., & Du, D. (2019). Dropout prediction in MOOCs: Using deep learning for personalized intervention. *Journal of Educational Computing Research, 57*(3), 547–570.

Xing, W., Guo, R., Petakovic, E., & Goggins, S. (2015). Participation-based student final performance prediction model through interpretable Genetic Programming: Integrating learning analytics, educational data mining and theory. *Computers in human behavior, 47*, 168–181.

Yu, L., Jiang, W., Shi, H., Yu, J., Liu, Z., Zhang, Y., ... & Liu, W. (2023). Metamath: Bootstrap your own mathematical questions for large language models. arXiv preprint arXiv:2309.12284.

Zhang, F., Xing, W., & Li, C. (2023, March). Predicting Students' Algebra I Performance using Reinforcement Learning with Multi-Group Fairness. In *LAK23: 13th International Learning Analytics and Knowledge Conference* (pp. 657-662).

Zhang, M. L., & Zhou, Z. H. (2013). A review on multi-label learning algorithms. *IEEE Transactions on Knowledge and Data Engineering, 26*(8), 1819–1837.

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., & Artzi, Y. (2019). Bertscore: Evaluating text generation with bert.arXiv preprint arXiv:1904.09675.

Zhang, Y., & Wallace, B. (2015). A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification. arXiv preprint arXiv:1510.03820.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

Authors and Affiliations

Fan Zhang[1] · Chenglu Li[2] · Owen Henkel[3] · Wanli Xing[1] · Sami Baral[4] · Neil Heffernan[4] · Hai Li[1]

✉ Wanli Xing
wanli.xing@coe.ufl.edu

Fan Zhang
f.zhang1@ufl.edu

Chenglu Li
chenglu.li@utah.edu

Owen Henkel
owen.henkel@risingacademies.com

Sami Baral
sbaral@wpi.edu

Neil Heffernan
nth@wpi.edu

Hai Li
lihai@ufl.edu

[1] University of Florida, Florida, USA

[2] University of Utah, Utah, USA

[3] Rising Academy Network, Freetown, Sierra Leone

[4] Worcester Polytechnic Institute, Worcester, USA