JOURNAL OF
EMERGING INVESTIGATORS

Article

Assessing large language models for math tutoring effectiveness

Pranav Goel[1], **Sindhu Ghanta**[2]
[1] Hindsdale South High School, Chicago, Illinois
[2] AIClub, Mountain View, California

SUMMARY

The decline in math performance among middle school students in the United States, particularly following the COVID-19 pandemic, has highlighted the need for effective and personalized tutoring solutions. This study explores the potential of Large Language Models (LLMs) as a tool for personalized learning, specifically in the context of math education. Our research question centers on the effectiveness of different LLMs—BERT, MathBERT, and OpenAI GPT-3.5—in providing help with math word problems posed by middle school students without directly answering the question for them. We hypothesized that a more sophisticated model (OpenAI GPT-3.5) and a math-specific LLM (MathBERT) will outperform a generically trained LLM (BERT) in assisting students with math problems. Using the Grade School Math 8k (GSM8K) dataset of math problems, we employed a methodology where student's math questions were matched with questions in the dataset. Then, the closest matching problems and solutions from the data set were provided to the student to help them understand the method to solve the problem posed. The effectiveness of each model was evaluated based on student feedback collected through dedicated web apps. The results showed that the OpenAI GPT-3.5 model received the highest average feedback score (4.72), indicating its superior performance in providing relevant solutions. Statistical analysis further confirmed a significant difference in the effectiveness of the OpenAI model compared to BERT and MathBERT. Overall, the study demonstrates the promising application of LLMs, particularly OpenAI GPT-3.5, in enhancing math education through personalized tutoring.

INTRODUCTION

Mathematics is a widely studied topic across the entire world and is a very useful field due to its applications in majors like engineering, physics, finance, and accounting. The future job output for individuals interested in math or statistics related majors have been predicted to increase by a considerable amount between now and 2028 (1). Yet, despite this increase in the necessity for individuals skilled in math, the math capability of youth in the United States (US) has been declining (2,3). Following COVID-19, math performance by elementary and middle schoolers have fallen by 6-15% on standardized tests when compared to pre-pandemic rates, according to the Northwest Evaluation Association (NWEA) (2). Furthermore, math scores decreased by nine points on the National Assessment of Educational Progress (NAEP) after the pandemic, which was the largest decrease in scores since 1973 (3).

In our modern learning environments where most students learn from a single teacher in a class of 30 students, one-on-one tutoring can be more beneficial to the student. This claim is supported by Benjamin Bloom's 2 Sigma Problem wherein Bloom found that students who were taught in an unconventional, individual-focused, environment performed significantly better statistically than students who were taught in a conventional, large bodied, setting (4). These unconventional learning environments that Bloom evaluated applied a variety of teaching methods to help the students master topics such as clear learning objectives, regular formative assessments, and additional opportunities to review and relearn the material (4). Despite the increasing demand for one-on-one tutors, there is a shortage of tutors focused on a student's individual learning.

One potential solution for this problem can be seen in the works of Large Language Models (LLMs). In recent times, there has been a tremendous amount of research done in the area of LLMs and its capabilities to understand and generate human text (5). Trained on extensive volumes of textual data, an LLM exhibits the capability to produce text that closely resembles human language. LLMs excel in tasks such as generating human-like text, responding to questions, and accomplishing various language-related activities with notable precision (6). Recent strides in LLMs stem from incorporating neural networks with billions of parameters trained extensively from self-supervised learning on massive unlabeled text data sets (6). Models like Bidirectional Encoder Representations from Transformers (BERT) and Generative Pre-trained Transformers (GPT) represent the forefront of recent advancements in LLMs, being the latest and commonly used models (11).

In the last few years, LLMs have been commonly used for a variety of purposes from tasks like essay and code writing, generating weather forecasts, holding interactive conversations with humans via chatbots, or "captioning images and visual scenes" (7). Moreover, interest has spiked in its applications in the fields of education and how LLMs can enhance learning and teaching experiences. An LLM's capability to cater to a student's specific needs allows it to be a potentially useful tool for personalized learning experiences (8).

To help address the declining math test scores and struggles that many elementary and middle schoolers are facing in math, we created web apps to help students with solving math problems. Our web apps were developed using LLMs that are able to find similar math problems and show students a way of solving them. We created three such unique web app tutors with different LLM embeddings, which is a vector representation of words, phrases and text that helps an LLM understand and process data. We developed one tutor with a regular BERT embedding, one with a MathBERT embedding, and a third with OpenAI's GPT-3.5 embedding to determine which LLM would be able to provide the most helpful responses to a middle school student. We hypothesized that using a more sophisticated LLM, such as Open AI's GPT-3.5, would be able to pull math questions that are more beneficial to students, in comparison to a smaller and more limited model such as BERT or MathBERT.

It is important to note that our web apps do not solve the student's problem itself, as that would not aid in the learning experience. By providing a similar problem and solution to the one that a student is looking for assistance with, the model should be able to help guide the student's thoughts on how to solve their own question without directly giving them the answer. This makes our models unique from many other tutors, many of which just give the answer to the posed question with limited explanatory steps.

Our results show that the OpenAI model, with an average feedback score of 4.72, outperformed both BERT (average score of 4.22) and MathBERT (average score of 4.07). Statistical analysis further supported the superior performance of the OpenAI model over the other two. This study displays the promising potential of LLMs in math education as a tool to assist students in their learning without needing a one-on-one human tutor.

RESULTS

In this study, we aimed to understand the effectiveness of three different LLMs—BERT, MathBERT, and OpenAI GPT-3.5—in accurately identifying and providing solutions to math word problems that are most similar to the ones posed by the students. Our approach is grounded in the idea that students can benefit and learn from seeing the solutions to closely related questions (12). Doing so enhances the students' understanding and problem-solving skills as well as provides cognitive support while learning (12).

We used the feedback provided about the effectiveness of each model to explore the impact of the type of model chosen for providing the closest question and answer. By converting the questions into numerical vectors and employing cosine similarity measures, we aimed to quantitatively assess the similarity between the students question and the dataset questions; thereby identifying the most relevant solution. Although cosine similarity, which measures the similarity of two vectors using the angle between them, was the form of measurement that was eventually used for the models, we also tested the efficiency and accuracy of using a Euclidean Distance form of measurement, which measures the straight linear distance between two points. In the end, the cosine similarity was found to be more helpful than a Euclidean Distance similarity because cosine similarity worked better with words encoded to vectors.

The results were collected on a scale between 0 and 6, where 0 represents "actively harmful" and 6 represents "extremely helpful". The rationale behind this scale is borrowed from research conducted on evaluating language models for mathematics through interactions (10). A set of 100 responses were collected for each web app from middle school students in Cook County and Lake County. The students were 6th-8th graders who matched the quality and skill level of the questions that our models were trained for. This helped ensure that varying difficulties and types of questions were asked to the model to ensure representative and holistic feedback on the accuracy and usefulness of the models. This was done in collaboration with teachers in the schools who shared the links to the web app with their students in their classrooms. The average score obtained by the BERT model was 4.22, the MathBERT model was 4.07, and the OpenAI GPT-3.5 model was 4.72 (**Figures 1, 2, 3**).

We performed statistical analysis using a two-sample *t*-test to compare the feedback means for OpenAI and MathBERT as well as for OpenAI and BERT. Both comparisons were found to be significantly different with both *p*-values < 0.001, which were both less than the alpha value of 0.05. However, the feedback means for BERT and MathBERT were also compared and not found to be significantly different, with a *p*-value of 0.38. Additionally, one-sided two-sample *t*-tests indicate that the mean feedback for OpenAI was significantly greater than that of MathBERT (*p*-value < 0.001) and BERT (*p*-values < 0.01). In contrast, the mean feedback for MathBERT was not significantly greater than that for BERT (*p*-value = 0.81). Our results suggest that the OpenAI model outperforms both MathBERT and BERT in terms of student feedback on the effectiveness of providing relevant solutions to their mathematics questions.

DISCUSSION

We created three versions of an LLM-based web app that provides one-on-one mathematics assistance for middle school students. We assessed three different LLMs, which were BERT, MathBERT, and OpenAI GPT-3.5. OpenAI GPT-3.5 outperformed BERT and MathBERT with an average feedback score of 4.72, indicating a high level of student satisfaction. Our results suggest that a more sophisticated model trained on larger and higher quality data, as in OpenAI GPT-3.5, will provide better support and assistance in math tutoring.

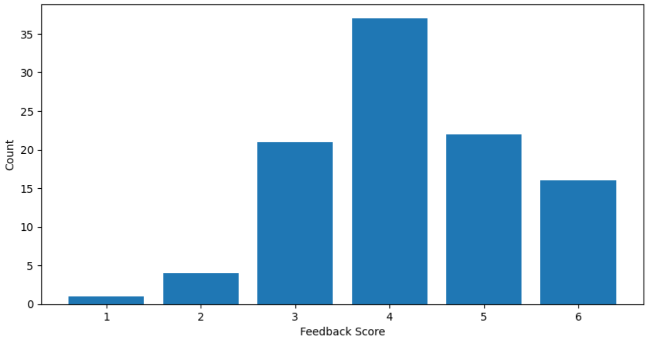


Figure 1: Bert model feedback. Feedback was gathered as ratings on a 0-6 scale, where a rating of 0 indicated that the provided problem was "actively harmful" and a rating of 6 was a "extremely helpful" problem. Feedback values from the mentioned 0-6 scale and the frequency that each rating was selected is displayed with an average score of 4.22.

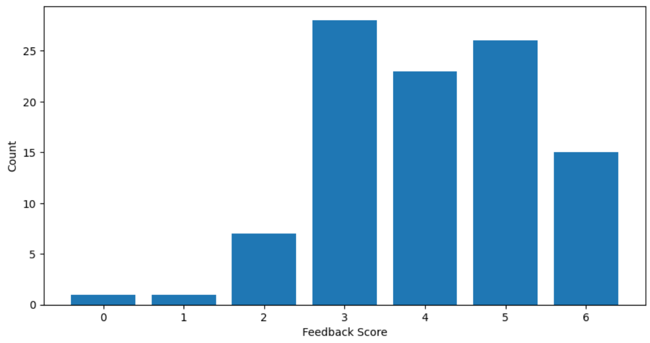


Figure 2: MathBERT model feedback. Feedback was gathered as ratings on a 0-6 scale, where a rating of 0 indicated that the provided problem was "actively harmful" and a rating of 6 was a "extremely helpful" problem. Feedback values from the mentioned 0-6 scale and the frequency that each rating was selected is displayed with an average score of 4.07.

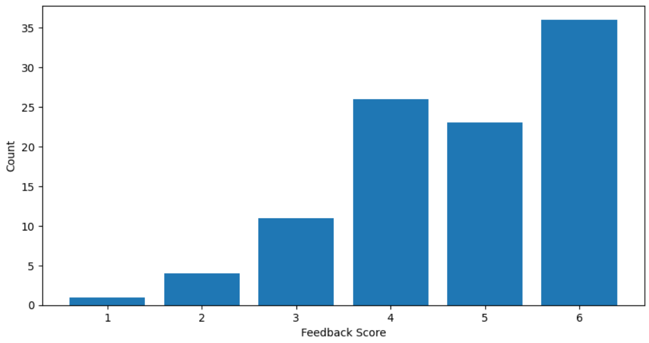


Figure 3: OpenAI model feedback. Feedback was gathered as ratings on a 0-6 scale, where a rating of 0 indicated that the provided problem was "actively harmful" and a rating of 6 was a "extremely helpful" problem. Feedback values from the mentioned 0-6 scale and the frequency that each rating was selected is displayed with an average score of 4.72.

One limitation that could have influenced our results is the reliance on student feedback as the sole measure of effectiveness. While feedback provides a direct assessment of student satisfaction, it may be subjective and influenced by factors such as the student's mood or personal preferences. Additionally, the study was conducted with a specific group of middle school students in Cook and Lake County, which may limit the generalizability of the findings to other populations or age groups.

Future experiments could involve a larger and more diverse sample of students, as well as the inclusion of objective measures of learning outcomes, such as pre-tests and post-tests on similar question style and difficulty to those the students are asking the LLM, to complement the subjective feedback data. Moreover, exploring the integration of these language models into a comprehensive educational platform that provides personalized learning experiences could further enhance their effectiveness in supporting math education.

Another limitation that could have influenced our results and the learning of the students using the web apps is model hallucination. In LLMs, model hallucination is when a model, such as the ones we have created, produces incorrect results. This happens when an LLM model, in attempting to produce coherent responses for the user, fills in gaps with plausible-sounding information that is not necessarily correct. Model hallucination is harmful in every situation but can have a drastic impact in the education sector as a student may be learning something that is incorrect or not applicable to what they are studying. Our models currently have no way of fully addressing model hallucination, but, based on feedback from students, most students were able to identify when a similar question was really not that similar to the one they asked. There still are situations in which a student may not fully recognize an incorrect or mismatched question, which could potentially influence the results of this study and the education of the students.

Future work and research into model hallucination could address this issue. One such solution could be to potentially add a message next to a similar question that was asked frequently but not given a strong feedback rating. Knowing that other users did not find the similar questions helpful to their learning could help the user understand that they should proceed with a bit more caution when looking at the same similar question.

In conclusion, this study demonstrates the promising application of language models in educational technology, particularly in the field of mathematics education. We believe that the benefits and individualized attention that an LLM-based tutor could provide to a student outweighs its potential harmful aspects due to technological errors. The superior performance of our initial tests with OpenAI GPT-3.5 highlights the importance of continued research and development in this area to fully harness the potential of these models for enhancing student learning as supplements to traditional classroom activities and settings.

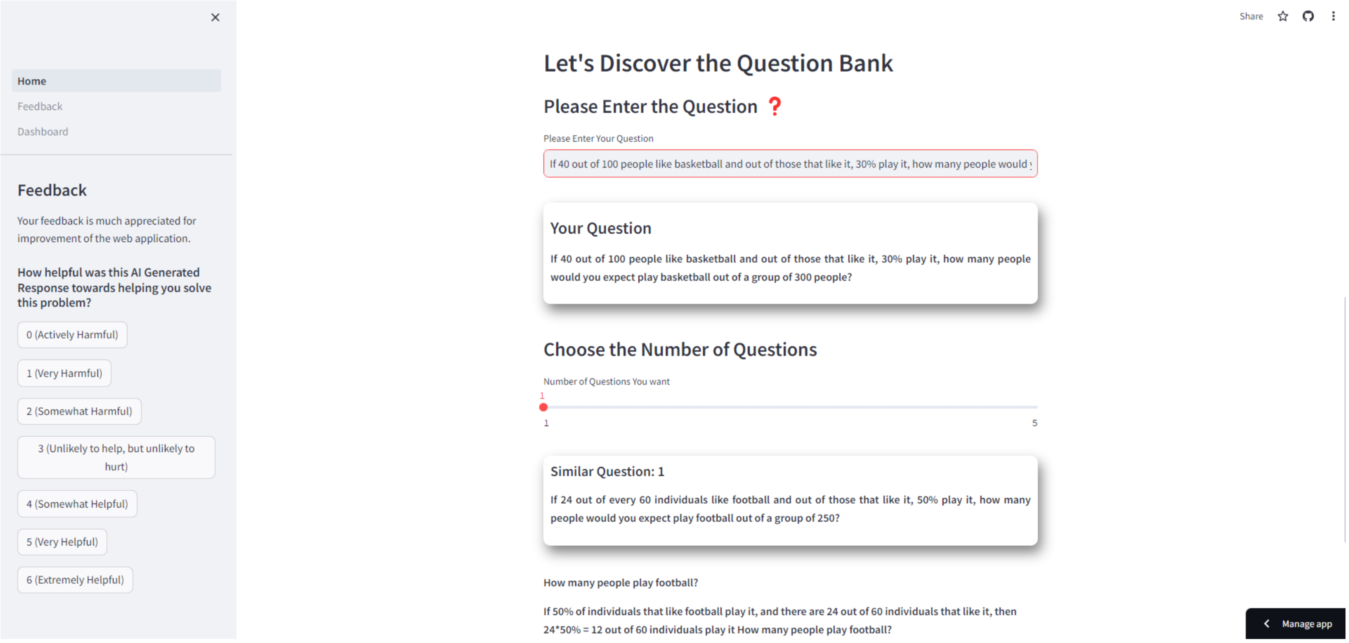
MATERIALS AND METHODS

Dataset

Our study utilized the Grade School Math 8k (GSM8K) dataset, a set of over 8,000 strong and representative middle school math world problems (9). This dataset was subdivided into 7,500 training problems and 1,000 testing problems, designed to challenge bright middle school students. The problems require two to eight steps to solve, and they involve basic arithmetic operations. The dataset was accessed from the grade_school_math/data directory, which contains train_socratic.jsonl and test_socratic.jsonl files, where each line represents a single problem in JavaScript Object Notation (JSON) format with assisting questions to help a student get to the answer. JSON is a data structure used to represent a set of key-value pairs. The JSON used was written in the format of a Microsoft Word document with 957,387 words cut across all the questions.

LLMs and web app creation

The JSON files obtained from the Github source were converted into CSV format with two columns, one for the question and the other for the answer. Each question column of the dataset was converted into numerical vectors such



Feedback

Your feedback is much appreciated for improvement of the web application.

How helpful was this AI Generated Response towards helping you solve this problem?

Let's Discover the Question Bank

Please Enter the Question ❓

Please Enter Your Question

If 40 out of 100 people like basketball and out of those that like it, 30% play it, how many people would y

Your Question

If 40 out of 100 people like basketball and out of those that like it, 30% play it, how many people would you expect play basketball out of a group of 300 people?

Choose the Number of Questions

Number of Questions You want

Similar Question: 1

If 24 out of every 60 individuals like football and out of those that like it, 50% play it, how many people would you expect play football out of a group of 250?

How many people play football?

If 50% of individuals that like football play it, and there are 24 out of 60 individuals that like it, then 24*50% = 12 out of 60 individuals play it How many people play football?

Figure 4: Representative user interface. A sample question on one of the web apps. Key features include the feedback system on the left-hand side of the application, the number of similar questions a user can choose from (ranging from 1-5), and the solution, which appears after the user chooses the number of questions they want.

that any incoming new question could also be converted into a similar numerical vector. This numerical vector was then compared to all the questions in the dataset for determination of the closest question. This conversion of the question into numerical vectors was done using three different model embeddings: BERT, MathBERT, and OpenAI GPT-3.5.

BERT is a generically trained language model with 110 million parameters, uses sub-word-based tokenization and embedding, which results in a numerical vector of 768 dimensions (13). MathBERT, on the other hand, consists of 383 million parameters and is specifically trained for math related text (13). MathBERT uses the "all-MiniLM-L12-v2" tokenizer and embedding, which results in a numerical vector of 384 dimensions (13). GPT3.5 is a large language model released by OpenAI with over 175 billion parameters and an embedding vector of 1536 dimensions (13). As a result of this conversion, we have three separate datasets with different numbers of columns, each column representing a numerical value of the vector embedding for every sample.

These embeddings were used in the following way: When a student posts a question they cannot solve, the question was converted into the embedding space using either BERT, MathBERT, or the OpenAI GPT-3.5 embeddings. A cosine similarity measure was used for measuring the similarity between any math questions and all the questions listed in the dataset. The cosine similarity measure helped with identifying the closest related question and its solution. Please note that the solution column stayed as is and was not embedded into numerical vectors. The closest related question was identified and displayed for the student along with a step-by-step solution.

Three separate web apps, one for each type of embedding, were created and made available online. The BERT-powered model is available at https://bertmathmodel.streamlit.app/. The MathBERT-powered model is available at https://MathBERT.streamlit.app/. And, the OpenAI GPT-3.5-powered model is available at https://MathBERTopenai-sj8vxi5nuwvfcfgebf6skx.streamlit.app/.

Student feedback collection

The 100 students from across Cook and Lake County that tested the apps were asked to input their question in the question box and choose how many similar questions, up to five, they were looking for (**Figure 4**). The web app would run by providing a similar question(s) and solution(s) to that problem with details based on the questions and solutions from our initial data set. Then, after each solution was given, a feedback bar appeared on the left hand of the screen to gather feedback data on the success and usefulness of the models. Feedback was gathered as ratings on a 0-6 scale, where a rating of 0 indicated that the provided problem was "actively harmful" and a rating of 6 was an "extremely helpful" problem (**Figure 4**).

Each student repeated this process three times, using the same question on all three of the web apps. The students did not know which app used which model type, but all students always tested the web apps in the same order: BERT, MathBERT, OpenAI GPT-3.5. By gathering feedback in this manner, bias was reduced as a student would have less opportunity to score a model higher or lower based on what they wanted to be the best model or perceived as the best.

Feedback analysis and statistical tests

After the feedback from all three websites was collected, the effectiveness of each method was calculated as the mean score of the responses received. An alpha value of 0.05 was chosen as it is widely considered the most common alpha value for. Two-sample *t*-tests and one-sided two-sample *t*-tests were conducted on these sample means to determine whether there were statistically significant differences between the feedback received for each model. These tests were conducted using Google Colab and Python code and packages, such as Pandas and Numpy, to calculate the *t*-statistics and the *p*-values for each test.

Received: April 16, 2024
Accepted: July 29, 2024
Published: February 21, 2025

REFERENCES

1. "Mathematicians and Statisticians: Occupational Outlook Handbook." *U.S. Bureau of Labor Statistics*. http://www.bls.gov/ooh/math/mathematicians-and-statisticians.htm. Accessed 6 Sep. 2023
2. Lewis, K. and Megan Kuhfeld. "Education's long COVID: 2022–23 achievement data reveal stalled progress toward pandemic recovery." *NWEA*, July 2023, htts://www.nwea.org/uploads/Educations-long-covid-2022-23-achievement-data-reveal-stalled-progress-toward-pandemic-recovery_NWEA_Research-brief.pdf. Accessed 8 Sep. 2023.
3. "NAEP Long-Term Trend Assessment Results: Reading and Mathematics." *The Nation's Report Card*. https://www.nationsreportcard.gov/highlights/ltt/2023/. Accessed 11 Apr. 2024.
4. Bloom, B. S. "The 2 sigma problem: The search for methods of group instruction as effective as one-to-one tutoring." *Educational researcher,* vol. 13, no. 6, June 1984, pp. 4-16. https://doi.org/10.3102/0013189X013006004
5. Sejnowski, Terrence J. "Large language models and the reverse turing test." *Neural computation*, vol. 35, no. 3, 21 Aug 2022 pp. 309-342. http://doi.org/10.48550/arXiv.2207.14382
6. Nye, B., et al. "Generative large language models for dialog-based tutoring: An early consideration of opportunities and concerns." *AIED Workshops*, 2023. http://ceur-ws.org/Vol-3487/paper4.pdf. Accessed 10 Oct. 2023.
7. Celikyilmaz, Asli, et al. "Evaluation of text generation: A survey." *arXiv,* 26 June 2020, https://doi.org/10.48550/arXiv.2006.14799
8. Kasneci, Enkelejda, et al. "ChatGPT for good? On opportunities and challenges of large language models for education." *Learning and individual differences*, vol. 103, April 2023, pp 102274. https://doi.org/10.1016/j.lindif.2023.102274
9. Cobbe, Karl, et al. "Training verifiers to solve math word problems." *arXiv*, 27 Oct. 2021, https://doi.org/10.48550/arXiv.2110.14168
10. Collins, Katherine M., et al. "Evaluating language models for mathematics through interactions." arXiv, 2 June 2023, https://doi.org/10.48550/arXiv.2306.01694
11. "Introduction to Large Language Models (LLMs): An Overview of BERT, GPT, and Other Popular Models." *John Snow Labs.* http://www.johnsnowlabs.com/introduction-to-large-language-models-llms-an-overview-of-bert-gpt-and-other-popular-models/. Accessed 5 Nov. 2024.
12. Sweller, John, et al. "Cognitive Architecture and Instructional Design." *Educational Psychology Review,* vol. 10, September 1998, https://doi.org/10.1023/A:1022193728205
13. "GPT-3 vs. BERT: Ending The Controversy" *SoftTeco* https://softteco.com/blog/bert-vs-chatgpt?WPACRandom=1736705158877. Accessed 12 Jan. 2025.