Generating math learning materials with LLMs

Overview

01

One-sentence focus

Exploring how NLP and Large Language Models can assist in creating accurate and relevant learning materials for mathematics education.

02

Keywords

Education, Studying, Coursework, Preparation, Generation

03

Research Question & Motivation

Current AI tools tend to deviate from specific course topics when asked to generate material. There's a need for reliable and accurate AI-generated educational materials. This would have the potential to support teachers and students in personalized learning

Methodology & Tools

Tools

Hugging Face
Provide an accessible interface potentially to interfacing with our own model

Python
Ease of development on model

PyTorch
Ease of model building

Model Capabilities

Marked Up Exam Labeling
Exam question and answers will be paired and used to train the model

Question Generation
Potentially use another model to compare if question generated matches topics

Grading
Give model answers to question and compare output grade to human generated grade

Training

Answer Checking
Have model generate answers and compare to human answers to same questions

Datasets
Archives exist of old mathematics exams across colleges and various government exam reports!

Evaluation
Compare grade, topic, and answer accuracy in materials generated and processed

Evaluation, Risks, and Benefits

As the project goes on, how do we evaluate the model? And what benefits are expected from this?

Evaluation

- Compare model graded exam questions v. human grades for same questions
- Analyze deviations and patterns in model errors
- Determine performance thresholds for practical classroom use

Risks

- Lack of labeled/graded exam data
- Poor model generalization or low performance
- Difficulty aligning model output with course-specific topics

Benefits

- A framework for evaluating LLM-generated learning materials
- Insight into how LLMs align with human graders
- Potential tool for teachers/students to generate study materials

Timeline

11/8/2025

Begin Data Labeling

We will begin gathering our data and labeling the data by question, answer, and overall grade.

11/21/2025

Transition to Model Training

Now that the data has been labeled, we will take our gold standard data and begin using test data to train a model

11/25/2025

Finalize Model Training

Nearing Final Project submission time, this will allow us to find where we're struggling. This is a pulse check.

11/30/2025

Complete Report and Training

Training will be completed at this point, and the output and results will be compiled into a detailed report

Source:
Can LLMs Master Math? Investigating Large Language Models on Math Stack Exchange

01

11 July 2024

Overall, in many models they tend to handle basic mathematics questions and early calculus very well. Anything beyond this and they begin to have issues.

This study tried out multiple models, and found GPT-o4 had the best performance.

They used Math stack exchange questions and answers on the model.

Source:
Grading Exams Using Large Language Models: A Comparison between Human and AI Grading of Exams in Higher Education Using ChatGPT

02

16 September 2024

This paper sought out to compare the differences between human graders and AI graders on real world exams.

This group used old exams from the School of Business at the University of Gothenburg, Sweden. As well it was tried out on many different exam topics.

It found the models tended to score around 70% of processed exams an average of +/- 10% from what a human grader scored. The more creative or longer an answer was, the more deviations occurred. As well, if English was not the exam taker's first language, it would also deviate. These deviations almost always were not in favor of the exam taker.

Source:
Large Language Models for Education: A Survey and Outlook

03

1 April 2024

This paper sought to prove that AI models have benefits when used in tutoring for school work and exam preparation.

It found that students that used models tended to have more of a bond to the model due to it learning over time quirks and weak points of the student.

However, a dependency would develop if relied upon too long.

The study wasn't large, but left room to see if it could potentially be useful or harmful in a follow up study.

Source:
A Model for Improving the Accuracy of Educational Content Created by Generative AI

04

14 May 2024

This study addresses the hallucinations or misinformation that can be present with using some language models by creating a tool for text processing and factual claims verification, focusing on extracting factual claims, retrieving evidence from authoritative sources, verifying content, and rewriting it to ensure accuracy.

The authors use a multi-stage NLP pipeline that includes claim detection, retrieval from verified sources, claim verification, and rewriting modules; they evaluated this on news-style text using a benchmark of factual errors and trusted source databases.

This is relevant to our project because it directly addresses AI hallucinations and misinformation, showing a method to systematically improve factuality, which is key for any use of AI in education, grading, or feedback where truthfulness and trust are critical.

Conclusions?

We have found that there is a great potential for models to act as educational tools in classrooms or studying.

The use of models to generate exam material prep work could easily help out with the struggle of determining what to look over in preparation for an exam.

As well, datasets exist, but we are very aware of the unique nature of what we are looking for and are aware of the risk.

Questions?