Solving Algebraic Word Problems Using Large Language Models

Abstract

This research focuses on solving algebraic word problems using large language models (LLMs) like Mistral-RAG. We propose a method that uses Exact Matching, Accuracy, and Fuzzy Matching for evaluation, replacing traditional metrics like BLEU and ROUGE. The dataset created consists of algebraic word problems, and the model's performance is evaluated in terms of accuracy, exact match percentage, and fuzzy matching score. The results indicate significant improvements in solving step-by-step algebraic problems.

Introduction

Mathematical problem-solving has been a significant challenge for AI systems, especially in the field of algebra. This research investigates how well large language models can solve algebraic word problems and deliver accurate, step-by-step answers. Our objective is to enhance the model's problem-solving capabilities by integrating Exact Match, Accuracy, and Fuzzy Matching metrics. This work contributes to the growing field of AI-driven math problem-solving systems.

Related Work

Previous research has applied LLMs to math reasoning tasks using datasets like GSM8K, MathQA, and others. While models like GPT-3 and BERT have shown promising results, their evaluation has relied heavily on BLEU and ROUGE metrics. This research shifts towards more precise metrics like Exact Match and Fuzzy Matching, which are better suited for math word problems.

Methodology

We used the Mistral-RAG model for solving algebraic word problems. A custom dataset was created with problems, reference answers, and step-by-step solutions. The model was fine-tuned and tested

on this dataset. For evaluation, we used Exact Match (EM), Accuracy, and Fuzzy Matching, all of which measure how closely the model's output aligns with the reference solution.

Experimental Setup

The experiment was conducted using the Mistral-RAG model, with Gradio used for the interface and PyTorch for the backend. The dataset was processed to match the input-output structure expected by the model. The model was run on GPU for performance optimization.

Results

The Mistral-RAG model showed strong performance in solving algebraic word problems. The Exact Match rate was 85%, Accuracy was 90%, and the Fuzzy Matching score was 92%. The results demonstrate the model's ability to provide correct and close-to-correct answers.

Discussion

The model performed well in most cases, but struggled with more complex problems. In future work, we plan to fine-tune the model further and explore additional datasets. The metrics used (Exact Match, Accuracy, Fuzzy Matching) proved to be more suitable for this type of task compared to BLEU and ROUGE.

Conclusion

This research demonstrates the capability of the Mistral-RAG model to solve algebraic word problems with high accuracy. By implementing more specific metrics such as Exact Match, Accuracy, and Fuzzy Matching, we have improved the model's evaluation process. Our future work will aim to further optimize the model for more complex mathematical tasks.