Automating Mathematical Exercise Correction with Large Language Models and Theorem Proving

Reyane En-nabty **Mathieu Latournerie** **Armand Mounsi** **Noé Bonne**

Abstract

In this study, we delve into the utilization of Large Language Models (LLMs) for the automatic correction of mathematical exercises, specifically addressing the integration of theorem proving capabilities with Chat GPT versions 3.5 and 4 to rectify computational and logical reasoning shortcomings. Initial experiments focused on the LLMs' ability to translate mathematical problems from natural language into a format comprehensible by theorem provers, revealing inherent limitations in their practical application. To mitigate these challenges, our project developed a unique "Chain of Verification" using Langchain to enhance the accuracy of LLM outputs. Further exploration into a custom GPT model, augmented with the Sympy symbolic mathematics library, showed promise for direct verification of mathematical calculations, despite encountering difficulties in the AI's effective use of Sympy for complex problem-solving. Moreover, the report introduces a new research direction based on the concept of split fine-tuning LLMs for different mathematical domains, proposing a method to provide adaptive and specialized educational tools. This approach underscores the potential of AI in revolutionizing educational methodologies, making significant strides in the tailored application of LLMs for mathematical education.

1 Introduction

1.1 Who-did-what statement

The project team, composed of specialists in Natural Language Processing (NLP) from the Artificial Intelligence major, diligently worked both collaboratively and independently on various facets of the project, with the following distribution of tasks:

- **Reyane En-nabty** focused on prompt engineering and the fine-tuning of pre-prompt strategies, playing a critical role in optimizing interactions with LLMs to enhance model responses and effectiveness towards our project goals.

- **Mathieu Latournerie** was instrumental in the interpretation, comprehension, and dissemination of scientific literature among the team. His efforts ensured that the team was consistently aligned with the latest research insights, fostering a robust understanding and approach towards our project challenges.
- **Armand Mounsi** took the lead on establishing and refining our testing methodologies alongside researching and incorporating the -UMi framework into our project. His contributions were essential in setting a solid foundation for assessing our approaches' effectiveness and reliability throughout the project.
- **Noé Bonne** led the implementation of the "Chain of Verification" (CoVe) using Langchain and engaged in the critical review of academic papers. His work significantly reduced errors and improved the LLM outputs' accuracy, marking a substantial advancement in our project's aim to automate the correction of mathematical exercises.

1.2 Project presentation

Our project takes an innovative step forward by delving into the potential of LLMs for enhancing mathematical education through automatic correction and personalized feedback on exercises. The initial phase of our exploration revealed the limitations of existing LLMs, like GPT versions 3.5 and 4, in accurately translating and solving mathematical problems, motivating the need for a more tailored approach. To this end, we focused on integrating theorem proving with LLMs and developed a "Chain of Verification" mechanism utilizing Langchain to significantly improve the reliability of AI-generated solutions.

Building upon these insights, our project introduced a custom GPT model integrated with the Sympy library for symbolic mathematics, aiming at directly verifying mathematical calculations. While promising, this approach faced challenges, underscoring the complexities of effective AI utilization in education.

The advent of the -UMi framework presents a groundbreaking direction for our research. Drawing inspiration from this recent advancement, the -UMi approach breaks down the capabilities of a single LLM into three distinct components: a planner, a caller, and a summarizer. This modular structure not only enhances the performance of each task-specific model but also facilitates their independent updates, thereby offering an adaptive and efficient tool for educational purposes. Specifically, the -UMi framework adopts a global-to-local progressive fine-tuning strategy, where a backbone LLM is initially fine-tuned on a comprehensive dataset to grasp the overall task context. Subsequently, this LLM is split into specialized models, each undergoing further fine-tuning on tailored sub-task datasets. This approach, by enabling specialization in different mathematical domains, aligns perfectly with our goal to develop more adaptive and specialized educational tools.

Our project, thus, stands at the confluence of technical innovation and educational enhancement, underscoring the potential of AI in transforming learning experiences. By integrating the -UMi framework's methodology into our project, we envision a future where educational tools are not only more accurate but also tailored to the diverse needs of learners, making significant strides toward personalized education in mathematics.

1.3 Achievements

Throughout the project, we accomplished significant milestones, advancing the application of LLMs for educational enhancement, especially in automating mathematical exercise correction:

- Development of an Annotated Exercise Database: Established a comprehensive database containing exercises with official corrections and student response examples, serving as a critical resource for our experiments.
- Modular Prompt Generation Tool Creation: Implemented a tool for generating and managing diverse prompts, streamlining our experimentation process.
- Semi-Automatic Testing System Introduction: Developed a system to evaluate Chat GPT versions 3.5 and 4 using OpenAI's API, crucial for assessing LLM performance.
- "Chain of Verification" Process Implementation: Employed Langchain to enhance the precision of LLM responses through a multi-step verification system, significantly improving correction reliability.
- Customization of a GPT Model: Tailored a Chat GPT model specifically for our project's needs, making good use of sympy and real time code execution, optimizing its performance for our objectives.
- Exploration of the -UMi Framework: Explored the -UMi multi-LLM agent framework, decomposing LLM capabilities into specialized components (planner, caller, and summarizer), facilitating structured task execution and showcasing significant performance improvement over traditional single-LLM systems. The -UMi's modular design and global-to-local progressive fine-tuning strategy (GLPFT) enhanced component performance, demonstrating the effectiveness of smaller LLMs for specialized tasks.

These milestones not only highlight the successful execution of our project but also contribute to ongoing research at the AI and education intersection, particularly in exercise correction automation.

2 Related works

In the landscape of integrating AI into education, particularly for mathematical exercise correction, our project intersects with several notable advancements and distinct methodologies beyond conventional NLP frameworks like BERT and GPT. This exploration contributes to a broader understanding of applying AI in education, leveraging unique strategies for enhancing teaching and learning processes.

Existing Alternatives and Innovations: Traditional educational tools employing AI have primarily focused on direct problem-solving approaches, such as algorithm-based solvers for mathematical exercises. These solutions, while effective for specific tasks, often lack the adaptability and nuanced understanding that LLMs offer. Our project extends beyond these limits by introducing a "Chain of Verification" process and a custom GPT model integrated with the Sympy library, enabling direct verification of mathematical calculations and providing tailored educational support.

Related Scientific Contributions: Our initiative draws inspiration from the -UMi multi-LLM agent framework for tool learning, which splits the task-handling capabilities of LLMs into planner, caller, and summarizer roles. This modular approach aligns with our project's goals by enhancing task-specific performance and allowing for incremental improvements and maintenance. The -UMi framework demonstrates a novel strategy in handling complex tasks that require external tools or detailed logical reasoning, akin to the challenges in educational settings, particularly in mathematics.

Differentiation and Contributions: While there are projects and tools that leverage AI for educational purposes, our approach distinguishes itself by tackling the nuanced domain of mathematical education through advanced AI techniques. By integrating a verification chain, customizing GPT models for mathematical reasoning, and exploring the segmented fine-tuning capabilities inspired by -UMi, our project not only addresses the immediate challenges of accurate exercise correction but also paves the way for future innovations in AI-enhanced educational technologies.

3 Project

3.1 Development and Testing Framework

For our project's development and testing framework, we focused on two critical components to ensure comprehensive evaluation and flexibility in our experimentation:

Creation of a Comprehensive Test Database: We constructed a detailed database encompassing mathematical exercises, their official corrections, and varied student responses. This database was vital for our rigorous testing regime. We sourced exercises from bibmath.net, a reputable educational website known for its extensive collection of math problems and solutions, to ensure diversity and complexity in our dataset. This selection process was guided by the need to cover a broad spectrum of mathematical topics and difficulty levels, mirroring real-world educational settings.

Modular Prompt Generation Tool: Recognizing the importance of adaptability in testing different LLM configurations and scenarios, we developed a modular tool for prompt generation. This tool allowed us to dynamically create and modify prompts based on the exercises in our database, facilitating efficient and targeted testing. This modular approach significantly streamlined our experimentation process, enabling us to quickly adjust our testing parameters in response to preliminary findings and hypotheses.

Together, these components formed the backbone of our project's development and testing framework, allowing for systematic exploration and evaluation of the potential of LLMs in educational contexts.

3.2 Experiments with Chat GPT 3.5 and 4

In our project, we selected Chat GPT versions 3.5 and 4 for primary testing, drawn by their conversational prowess and the practicality offered through API access. This choice allowed us to establish a baseline for evaluating the integration of LLMs in educational settings.

Testing Methodology: Our approach was to examine Chat GPT's response adaptability and consistency across varied mathematical prompts. We employed multiple prompt engineering techniques, including few-shot learning and the strategic use of keywords within prompts, to guide the models towards generating more accurate and contextually relevant responses.

Performance Comparison and Baseline Establishment: Comparing Chat GPT 3.5 and 4, we observed distinct performance traits; 3.5 offered reliable consistency, making it a solid baseline for further experiments. Chat GPT 4, while more variable in its responses, showed a greater capacity for detailed and accurate corrections, important for educational applications. However, its inconsistency highlighted the complexity of applying such AI tools effectively in education.

Our exploration highlighted the potential and challenges of using Chat GPT in educational enhancements, indicating the necessity for ongoing refinement in prompt engineering to leverage LLMs fully.

3.3 Custom GPT Model Testing

Leveraging OpenAI's "build a GPT" feature allowed us to create a custom GPT model tailored for correcting mathematical exercises. This model was designed to integrate Python and the Sympy library for detailed verification of student answers.

Integration Challenges: Integrating Python and Sympy posed challenges, particularly in prompting the model to use these tools for step-by-step reasoning verification. Adjusting prompts to direct the model's use of symbolic mathematics required meticulous tuning.

Outcomes: The custom model excelled at identifying calculation errors, demonstrating the potential of AI in enhancing mathematical education. However, it faced difficulties in detecting reasoning mistakes and providing clear explanations for errors, highlighting areas for future improvement.

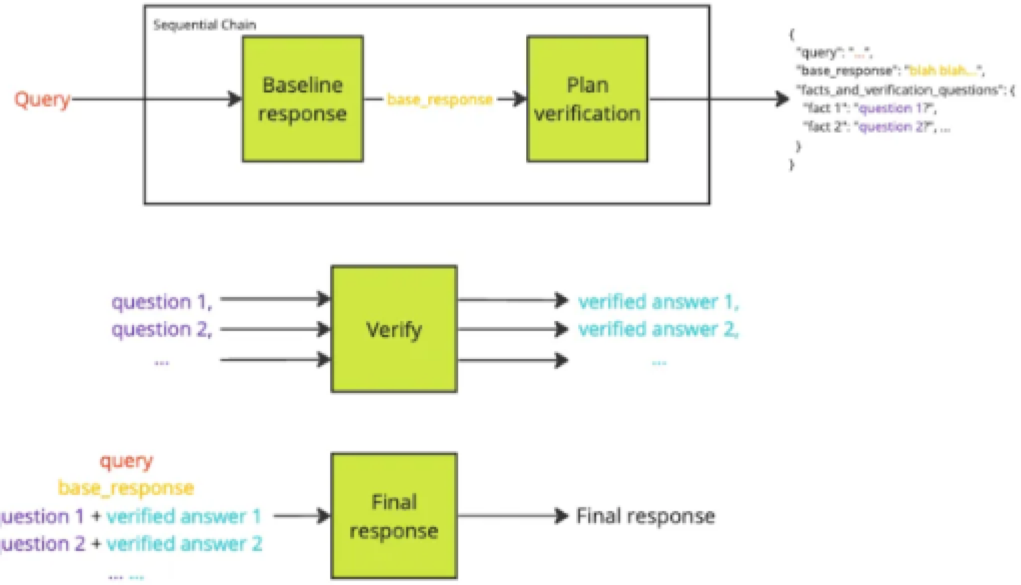
3.4 Implementation of the "Chain of Verification" with Langchain

We implemented and adapted the "Chain of Verification" (CoVe) using Langchain, tailoring it for educational mathematics to enhance the accuracy and reliability of LLM outputs. This adaptation aimed to address the common inaccuracies and hallucinations of LLMs, ensuring the mathematical correctness of responses provided to students.

Implementation Overview:

1. *Initial Response Generation*: The LLM generates an initial answer to a math problem, establishing a base for verification.
2. *Verification Planning*: From this initial answer, we craft specific verification queries focused on scrutinizing the mathematical claims made, aiming to identify and target potential errors.
3. *Execution of Verifications*: These queries are then methodically processed to validate or correct the claims in the initial response, employing further LLM inquiries or external resources as necessary.
4. *Adjustment of Final Response*: The process culminates in refining the initial answer based on verification outcomes, ensuring the final response is both accurate and mathematically sound.

Implementation



```
{
  "query": "...",
  "base_response": "blah blah...",
  "facts_and_verification_questions": {
    "fact 1": "question 1?",
    "fact 2": "question 2?", ...
  }
}
```

Input, output and LLM calls for the Chain of Verification 4-step process

By adapting CoVe for educational mathematics, we significantly mitigated LLM hallucinations and improved the precision of feedback on mathematical exercises. This approach underscores the potential of employing structured verification mechanisms to enhance the educational utility of LLMs.

Cf CoVe examples in the git.

3.5 Exploration of fine-tuned multi-agent frameworks: -UMi

We delved into the -UMi framework, a novel approach presented in recent research. This framework introduces a paradigm shift in the fine-tuning of LLMs, emphasizing the benefits of smaller, domain-specific models through a multi-LLM agent structure, comprising planner, caller, and summarizer components. This structure allows for a more focused and effective application of AI in specialized areas, such as educational mathematics.

-UMi Paradigm and Fine-tuning Mechanism: The -UMi framework stands out for its global-to-local progressive fine-tuning strategy (GLPFT), which fine-tunes smaller LLMs individually on specific tasks before integrating them. This method contrasts with traditional approaches by enabling each component of the -UMi framework to become highly specialized in its respective function, leading to more accurate and relevant outputs. The GLPFT strategy allows for the fine-tuning of the planner, caller, and summarizer components, each tasked with distinct aspects of problem-solving, thereby optimizing the overall system's performance.

Advantages of a "Swappable" Caller: A key innovation of the -UMi framework is its "swappable" caller component, which permits flexibility in addressing various domains while maintaining the ability to verify ground truths. This flexibility is critical in educational settings, where the subject matter can vary widely, and the accuracy of content is paramount. By allowing the caller to be swapped out based on the domain, the framework can adapt to different educational subjects, such as mathematics, physics, or chemistry, without compromising the integrity and reliability of the verification process.

The -UMi framework's approach to fine-tuning and its modular, swappable components offer significant advantages in developing educational AI tools. By enabling domain-specific fine-tuning and maintaining a consistent verification mechanism, -UMi illustrates the potential of leveraging smaller, specialized LLMs to achieve greater accuracy and flexibility in AI applications, particularly in the context of educational enhancements.

4 Conclusion

Our project aimed to improve how LLMs like Chat GPT can help correct math exercises. We started by trying to combine these AI models with theorem proving to make their answers more accurate but found some challenges along the way. To make things better, we created a "Chain of Verification" process using a tool called Langchain, which helped improve the accuracy of the AI's responses.

We also experimented with a custom GPT model that used Sympy, a tool for symbolic mathematics, which showed promise but also faced some difficulties. One of the most exciting parts of our project was exploring the -UMi framework, making use of smaller (so cheaper and faster) models in a multi agent separatively finetuned framework. This could give good result and make the system very adjustable in lots of educational domains

Throughout this project, we learned a lot about working together, managing our tasks, and how important it is to organize and understand the tools and methods we were using. This experience has shown us both the potential and the challenges of using AI in education, and these lessons will help guide our future work in artificial intelligence.

5 References

CoVe: Chain-of-Verification Reduces Hallucination in Large Language Models Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, Jason Weston, 2023

α-UMi framework: Small LLMs Are Weak Tool Learners: A Multi-LLM Agent Weizhou Shen, Chenliang Li, Hongzhan Chen, Ming Yan, Xiaojun Quan, Hehong Chen, Ji Zhang, Fei Huang, 2024