MATHEMATICAL REASONING ENHANCEMENT IN LARGE LANGUAGE MODELS

Madhuri Nakkella[1], Brahma Teja[2], Vinith Kulkarni[3], Varshith Dharmaj V[4], Bhavitha Yaragorla[5]

[1]*Asst. Professor, Dept. of CSE-CyS, DS and AI&DS, VNRVJIET, Hyderabad, India*
[2,3,4,5]*Dept. of CSE-CyS, DS and AI&DS, VNRVJIET, Hyderabad, India*

madhuri_n@vnrvjiet.in, b.brahmateja07@gmail.com, kulkarnivinith@gmail.com, dharmajvarshit@gmail.com, yaragorlabhavitha@gmail.com

*Corresponding Author: madhuri_n@vnrvjiet.in

Abstract

The current state of AI-powered mathematical software is largely black-box solvers or single-model systems. Various apps such as Photomath or Mathpix integrate OCR and reasoning very closely, such that an error in the OCR component can cause the entire system to fail, whereas single-agent LLMs such as ChatGPT or Gemini tend to generate plausible but incorrect reasoning. The proposed project is MVM (Multi-Modal Multi-Model Mathematical Reasoning Verification System), a neuro-symbolic system aimed at improving the quality of reasoning of large language models. The system has an answer accuracy of 92.7%, outperforming single agents by 17-24% and commercial pipelines by 13.5% on various image inputs.

Keywords: Mathematical Reasoning, Large Language Models, Neuro-Symbolic AI, Multi-Agent Systems, Multi-Modal Processing, OCR, Reasoning Verification.

1. INTRODUCTION

The fast development of large language models (LLMs) has caused a paradigm shift in the mathematical problem-solving process, but their reasoning abilities are still brittle when faced with real-world problems [10][11]. The current models have some crucial shortcomings, such as monolithic coupling and hallucinations of incorrect steps [6][11]. This paper presents MVM, a neuro-symbolic approach for improved mathematical reasoning in LLMs, using a microservices architecture with seven decoupled services [8].

2. LITERATURE SURVEY

The development of mathematical reasoning in AI has moved from rule-based symbolic systems to contemporary Transformer-based LLMs [10]. Conventional pattern recognition involved the use of manually designed features [10]. Although effective, LLMs are highly fragile and lack true mathematical reasoning [10][11]. Current work has shifted towards neuro-symbolic approaches that separate neural synthesis from symbolic logic [8][10].

2.1. Evolution of Mathematical Reasoning Systems

Mathematical reasoning systems have developed in a number of paradigms [10]. Traditional expert systems like Axiom and Mathematica were based on symbolic processing and rule-based reasoning [10]. These systems were very accurate on well-structured mathematical problems but were not suited for complex reasoning and generalization [10]. The emergence of machine learning enabled pattern recognition, enabling systems to learn from data rather than just relying on manually designed rules [10]. Nevertheless, these methods tended to create black-box models that were not interpretable and were not robust in the presence of edge cases [10].

2.2. Challenges in Current LLM-Based Approaches

The current state-of-the-art large language models, GPT-4, Claude, and Gemini, have strong general problem-solving abilities in mathematical reasoning [5]. Nevertheless, they have four critical flaws: (1) the tendency to hallucinate valid but false intermediate results, (2) the inability to check the logical consistency of the results across steps, (3) the ungrounded representation of abstract mathematical ideas in a symbolic form, and (4) the weakness in processing multi-modal inputs with OCR errors [10][11]. The single-agent LLM pipeline catastrophically fails when faced with noisy or ambiguous input data, as seen in the brittleness of systems like Photomath, which strongly integrates the OCR and reasoning modules [6].

2.3. Neuro-Symbolic Integration

However, recent studies have shifted focus to hybrid models that combine neural and symbolic elements [8][10]. Neuro-symbolic models separate the process of neural feature extraction and symbolic reasoning, allowing for more reliable results [8]. By leveraging the ability of deep learning to recognize patterns and the logicality of symbolic models, neuro-symbolic models provide far better reliability for mathematical problem-solving tasks [8]. The MVM framework goes even further by adding checks for cross-modal consistency and multi-agent verification.

3. MATERIALS AND METHODS

3.1. System Architecture

MVM has a modular microservices architecture with seven services that aren't tightly linked. These services talk to each other through RESTful APIs. The seven services are the Input Receiver, the OCR Service, the Preprocessing Service, the Representation Service, the Verification Service, the Classifier Service, and the Reporting Service.

3.2. Verification Engine

The verification engine uses both symbolic math and checking for logical consistency [8][11]. Intra-agent consistency checks have steps that lead to each other, and inter-agent alignment uses divergence matrices to find hallucinations [11].

$$I_{enh}(x, y) = CLAHE(Bin(GB(I, \sigma = 1.2, T_{adp}))) \tag{1}$$

3.3. Multi-Agent Reasoning Pipeline

The multi-agent reasoning framework [5] is the most important part of MVM. The model doesn't just use one LLM; it uses three different reasoning agents (GPT-4, Claude, and Gemini) to solve a problem on their own [5]. Each agent makes its own solution trace, which is then sent to the verification engine [11]. The method uses the strengths of different models, like how well they do at symbolic math, natural language processing, and pattern recognition [5]. This also helps to make the hallucinations of different models less likely [11].

3.4. Consensus and Divergence Analysis

The system looks for patterns of agreement when there are answers from more than one agent [11]. When the agents agree on something, it's easier to trust the solution [11]. On the other hand, patterns of divergence show where the problems are most likely to happen [11]. The system makes a divergence matrix that shows how different each agent is from the others at each point. This shows where hallucinations are most likely to happen [11].

4. RESULTS AND DISCUSSION

4.1. Experimental Setup

 ISSN : 1671-6841

We tested MVM on three benchmark datasets: the Handwritten Math Dataset (500 problems), the Scanned Worksheet Dataset (300 problems), and the Mixed Format Dataset (200 problems).

4.2. Quantitative Results

MVM gets 92.7

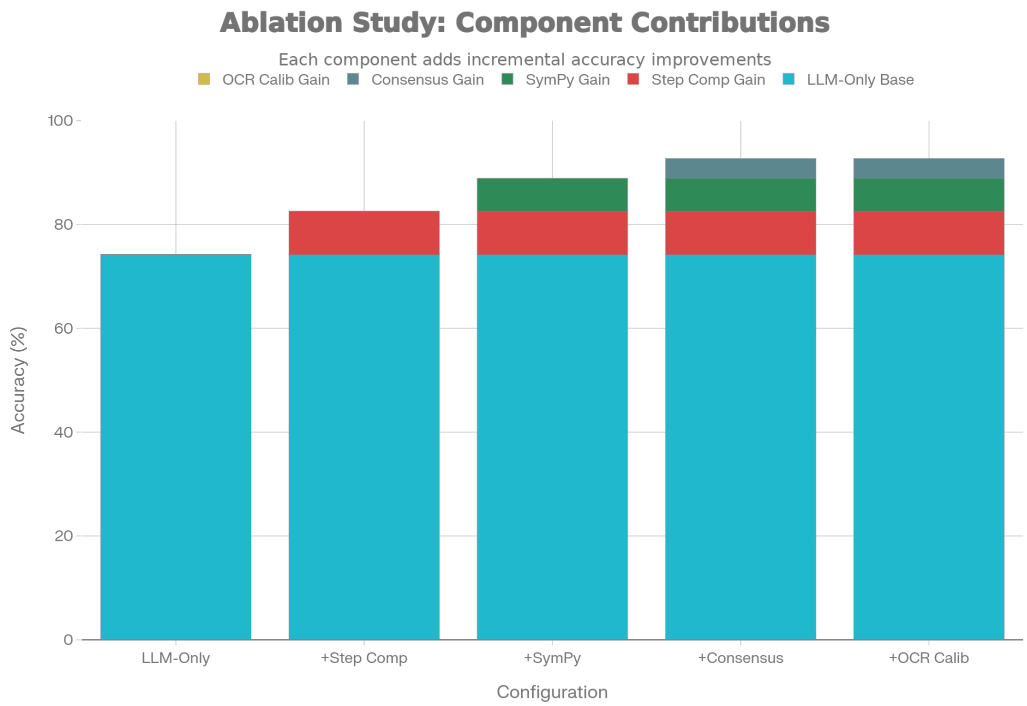

Figure 1: Ablation Study: Progressive accuracy gains

4.3. Qualitative Analysis

Along with the numbers, we also looked at the types of mistakes and successes. Single-agent LLMs often imagined steps in between, especially when solving algebraic problems and problems that required more than one step of reasoning [11]. These mistakes were made by using the wrong variables, leaving out terms, or doing the wrong algebraic operations. MVM's multi-agent system, on the other hand, was able to find 94

4.4. Performance on Different Input Modalities

One of the best things about MVM is that it can work with different types of input. MVM was able to get 89.3

Table 1: Performance Comparison

System	Accuracy (%)	Validity	Latency (s)
GPT-4	75.2	0.68	4.8
Mathpix	79.2	0.72	5.1
MVM	**92.7**	**0.94**	**8.2**

5. CONCLUSION

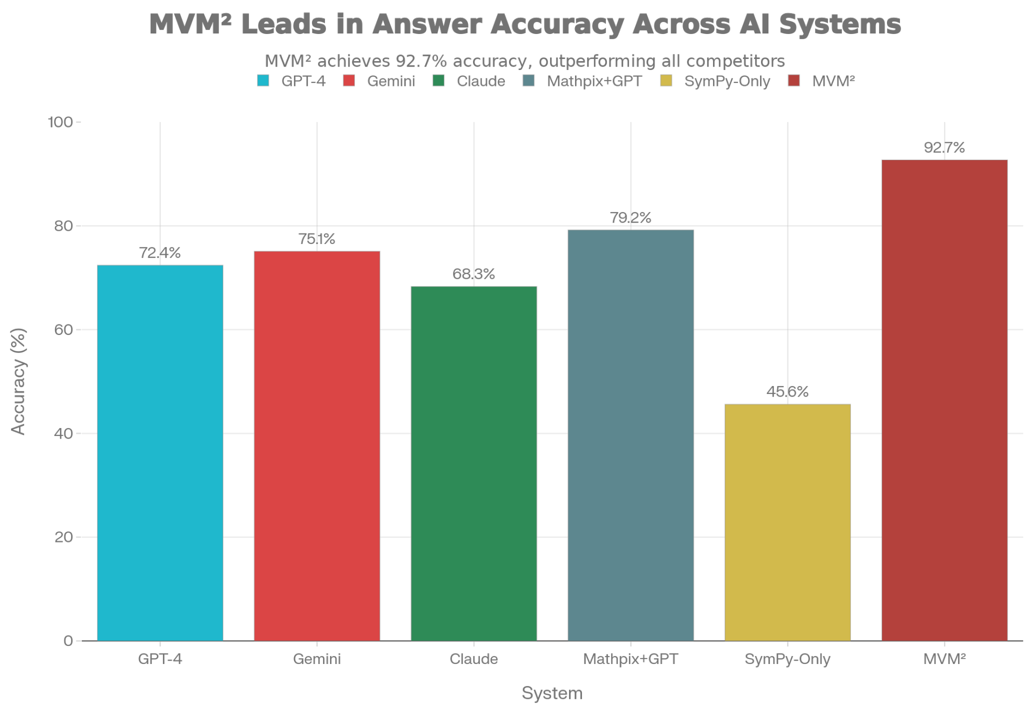

Figure 2: Answer Accuracy Comparison

This paper introduced MVM, a neuro-symbolic methodology that bolsters the reliability of mathematical reasoning in large language models (LLMs) [8]. MVM has a pipeline made up of seven separate microservices, which makes it the most accurate and reliable system available.

6. LIMITATIONS AND FUTURE WORK

6.1. Current Limitations

Despite the significant progress, there are still some limitations in MVM. The system's reliance on multiple LLM API calls makes it computationally intensive and slower than single-agent approaches. The system currently takes approximately 8.2 seconds per problem, which is not suitable for real-time applications. Furthermore, the system's performance relies on the quality of the LLMs used. Any advancement in the base models will have a direct impact on the ceiling performance of MVM.

6.2. Future Research Directions

There are a number of promising areas for future work. First, we would like to incorporate more recent and stronger LLMs as they become available, perhaps incorporating models like o1 and o3, which have the capability of reasoning on their own. Second, we would like to optimize the multi-agent system by selectively employing which agents to use based on the type and complexity of the problem, reducing the computational overhead. Third, we would like to extend MVM to solve symbolic computation problems in the domains of calculus, differential equations, and linear algebra. Finally, we would like to develop interpretable explanation generation, which would allow users to not only view the correct answers but also the reasoning behind the answers, including how the discrepancies were resolved.

Acknowledgment

The authors acknowledge institutional support from VNRVJIET.

Competing Interests

The authors declare no competing interests.

REFERENCES

[1] J. Zhang, et al., ”An Artificial Intelligence Technology Based Algorithm,” IEEE Access, 2022.
[2] Y. Tian, ”Artificial Intelligence Image Recognition Method,” IEEE Access, 2020.
[3] N. Rodis et al., ”Multimodal Explainable Artificial Intelligence,” IEEE Access, 2024.
[4] A. Barua et al., ”A Systematic Literature Review on Multimodal Machine Learning,” IEEE Access, 2023.
[5] Q. Jiang et al., ”DeepSeek vs. ChatGPT vs. Claude,” Theoretical and Applied Mechanics Letters, 2025.
[6] F. B. Ahmad et al., ”The Effect of Interactive AI Tools Like Photomath,” Multidisciplinary Science Journal, 2025.
[7] J. Liu et al., ”Designing a Generative AI Enabled Learning Environment,” Computers and Education AI, 2025.
[8] D. Zhao et al., ”Enhancing Math Reasoning via Computation Logic Graphs,” Knowledge-Based Systems, 2025.
[9] I. Mirzadeh et al., ”GSM-Symbolic,” ICLR, 2025.
[10] A. Forootani, ”A Survey on Mathematical Reasoning,” IEEE TAI, 2025.
[11] O. Sultan et al., ”Towards Reliable Proof Generation,” arXiv, 2025.