Enhancing Large Language Models for Natural Language Mathematical Reasoning via Formal Proof AutoInformalization

Rishi Padmanabhan
Stanford University
riship1@stanford.edu

Shane Mion
Stanford University
smion@stanford.edu

Ameya Jadhav
Sanford University
ajadhav@stanford.edu

Brando Miranda
Stanford University
brando9@stanford.edu

May 12, 2024

Abstract

This study introduces a method to improve Large Language Models' (LLMs) mathematical reasoning capabilities by integrating formal proofs from Interactive Theorem Provers (ITPs) into their training. We fine-tune GPT-3.5, Mistral-7B, and Gemma-7B models with datasets pairing formal and informal proofs. The effectiveness of this approach is assessed using the Hendrycks MATH dataset and Massive Multitask Language Understanding (MMLU) benchmark. Results show improvements in LLMs' performance on various mathematical categories, suggesting the potential of formal proofs to advance LLMs' reasoning abilities. Further exploration of diverse formal proofs and advanced fine-tuning techniques is necessary to bolster LLMs' formal mathematics comprehension.

1 Introduction

Enhancing the mathematical reasoning capabilities of Large Language Models (LLMs) is crucial for advancing artificial intelligence and automated theorem proving. Current LLMs demonstrate an impressive understanding of language tasks but lack proficiency in deciphering and formulating rigorous formal mathematics [1, 2]. This research proposes leveraging verified proof libraries from Interactive Theorem Provers (ITPs) to catalyze enhancements in LLMs' ability to understand and generate formal mathematical proofs.

2 Related Work

Several studies explore using deep learning and neural networks for theorem proving [1, 5, 11, 15], formalizing and mechanizing mathematical proofs [6, 8, 9, 14], and advanced machine learning techniques for theorem proving [3, 4, 10, 12, 16]. Our research differentiates itself by directly leveraging verified proofs from ITPs to enhance LLMs' reasoning and autoformalization abilities.

3 Methodology

Our approach involves fine-tuning GPT-3.5, Mistral-7B, and Gemma-7B models with datasets of formal-informal proof pairs. We constructed two datasets, ClaudeJson and MistralJson, using the LeanDojo proof library and LLMs for informal proof generation. The models were fine-tuned on these datasets and evaluated on the Hendrycks MATH dataset and MMLU benchmark.

Model	MMLU Category	Standard [1]	Fine-tuned [2]
Gemma-7b	Humanities/Social Sciences	46.5%	**48.2%**
	STEM	35.9%	**38.4%**
Mistrab-7b	Humanities/Social Sciences	41.5%	38.6%
	STEM	25.6%	**28.4%**
GPT-3.5	Humanities/Social Sciences	43.2%	36.7%
	STEM	34.8%	**37.6%**

Table 1: Comparison of performance in MMLU categories across different models, AutoInformalization improves the reasoning capabilities of LLMs generally from high quality formal source code, with minimal degradation in non-mathematicla reasoning.

Hendrycks Category	Gemma-7b		Mistral-7b		GPT-3.5	
	Std [1]	FT [2]	Std [1]	FT [2]	Std [1]	FT [2]
Count	42.9%	42.8%	19.1%	**23.8%**	38.5%	34.6%
Algebra	40.0%	33.3%	16.7%	**33.3%**	33.3%	**58.3%**
Geometry	69.2%	**76.9%**	18.2%	9.1%	78.6%	42.9%
Intermediate Algebra	7.6%	**38.5%**	9.5%	**14.3%**	6.7%	**33.3%**
Number Theory	38.5%	**53.9%**	13.6%	**22.7%**	26.3%	**40.0%**
Pre-algebra	46.6%	**73.3%**	23.8%	**28.6%**	66.7%	**83.3%**

Table 2: Comparison of Hendryck's category performance across different models, AutoInformalization improves the reasoning capabilities of LLMs generally from high quality formal source code.

4 Results

5 Discussion

The results provide evidence that models trained with formal/informal proof pairs can improve performance on mathematical tasks. Improvements are seen in higher-level categories such as geometry, algebra, and number theory. However, there is mixed evidence for improved performance on non-STEM subjects.

Further research should explore integrating a wider variety of formal proofs, employing more advanced models, and refining evaluation metrics. Additionally, investigating multi-stage fine-tuning processes could yield further improvements.

Mathematics is founded on logical principles and rigorous reasoning. Advanced mathematical concepts require the ability to understand and work with complex symbolic representations, formulate precise definitions, and construct logically valid proofs. By training LLMs on these types of mathematical domains, they are forced to develop strong skills in formal logic, deductive reasoning, and manipulating abstract symbolic structures. Our research highlights the potential for formal mathematical proofs to enrich the training datasets of LLMs, potentially leading to broader applications in fields that require the interpretation and understanding of complex mathematical concepts.

6 Conclusion

Our research highlights the potential for formal mathematical proofs to enrich the training of LLMs, leading to enhanced mathematical reasoning capabilities. The observed gains warrant further investigation into the integration of diverse formal proofs and the use of advanced fine-tuning techniques to bolster LLMs' comprehension of formal mathematics.

References

[1] Kshitij Bansal, Sarah M. Loos, Markus N. Rabe, Christian Szegedy, and Stewart Wilcox. Holist: An environment for machine learning of higher-order theorem proving. 2019.

[2] Jesse Michael Han, Igor Babuschkin, Harrison Edwards, Arvind Neelakantan, Tao Xu, Stanislas Polu, Alex Ray, Pranav Shyam, Aditya Ramesh, Alec Radford, and others. Unsupervised neural machine translation with generative language models only. 2021.

[3] Jesse Michael Han, Jason Rute, Yuhuai Wu, Edward W. Ayers, and Stanislas Polu. Proof artifact co-training for theorem proving with language models. 2022.

[4] Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. 2021.

[5] Daniel Huang, Prafulla Dhariwal, Dawn Song, and Ilya Sutskever. Gamepad: A learning environment for theorem proving. 2018.

[6] Wenda Li, Lei Yu, Yuhuai Wu, and Lawrence Paulson. Modelling high-level mathematical reasoning in mechanised declarative proofs. 2020.

[7] Felipe Maia Polo, Lucas Weber, Leshem Choshen, Yuekai Sun, Gongjun Xu, and Mikhail Yurochkin. Tinybenchmarks: evaluating llms with fewer examples. 2024.

[8] The mathlib Community. The lean mathematical library. In *Proceedings of the 9th ACM SIGPLAN International Conference on Certified Programs and Proofs*, pages 367–381, 2020.

[9] Dennis Müller, Florian Rabe, Colin Rothgang, and Michael Kohlhase. Representing structural language features in formal meta-languages. In Christoph Benzmüller and Bruce Miller, editors, *Intelligent Computer Mathematics*, pages 206–221. Springer, 2020.

[10] Stanislas Polu, Jesse Michael Han, Kunhao Zheng, Mantas Baksys, Igor Babuschkin, and Ilya Sutskever. Formal mathematics statement curriculum learning. 2022.

[11] Stanislas Polu and Ilya Sutskever. Generative language modeling for automated theorem proving. 2020.

[12] Markus N. Rabe, Dennis Lee, Kshitij Bansal, and Christian Szegedy. Mathematical reasoning via self-supervised skip-tree training. 2020.

[13] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. Sequence to sequence learning with neural networks. 2014.

[14] Sean Welleck, Jiacheng Liu, Ronan Le Bras, Hannaneh Hajishirzi, Yejin Choi, and Kyunghyun Cho. Naturalproofs: Mathematical theorem proving in natural language. 2021.

[15] Daniel Whalen. Holophrasm: a neural automated theorem prover for higher-order logic. 2016.

[16] Yuhuai Wu, Markus Rabe, Wenda Li, Jimmy Ba, Roger Grosse, and Christian Szegedy. Lime: Learning inductive bias for primitives of mathematical reasoning. 2022.