MathRAG: Retrieval-Augmented Generation for Verifying Math Solutions.

Kseniia Petukhova, Van-Hoang Nguyen

{kapetukhova, nguyenvanhoang7398}@gmail.com

Introduction

- Current evaluation of Large Language Models (LLMs) in mathematical tasks focus solely on the final results while overlook the quality of intermediate steps.
- Existing research on using LLM-based evaluator solely relying on its parametric reasoning ability.
- In this work, we explore the application of Retrieval-Augmented Generation to improve the capability of LLM-based evaluator by considering external sources of mathematical knowledge.

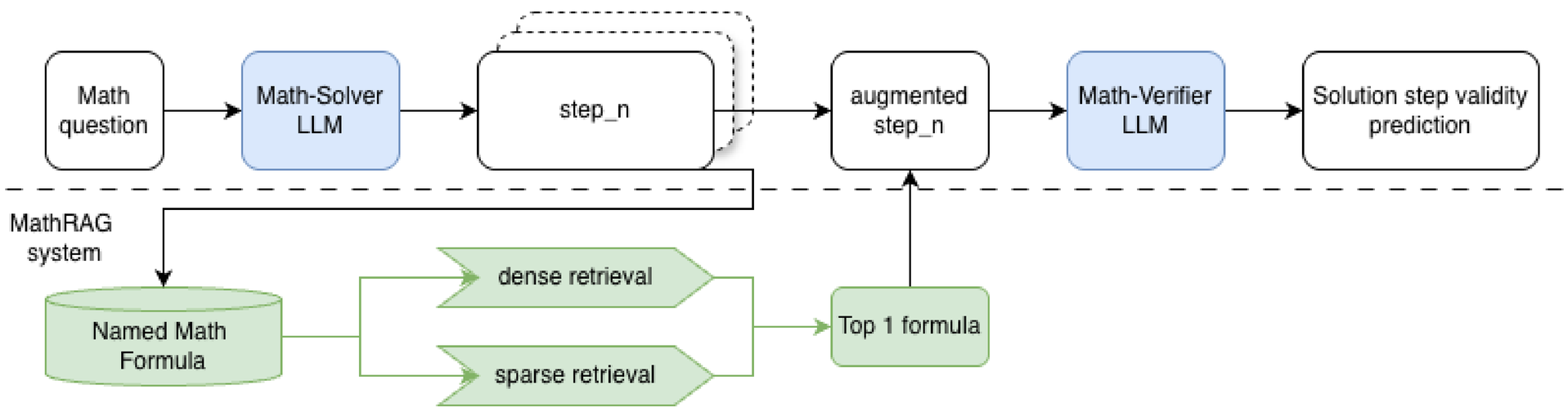
Methodology

RAG

- Using the Named Mathematical Formulas (NMF) dataset [1], we built a retrieval system with 2 methods of retrieve: (1) Dense-retrieval using Sentence Transformers [2] (all-MiniLM-L6-v2) and (2) Term-retrieval using BM25 [3].
- Given each step of the mathematical solution as a query, we retrieve the top 1 math formula from (1) Dense-retrieval with the highest cosine similarity and the top 1 math formula from (2) Term-retrieval with the highest BM25 score.

LLM-based Verifier

- Based on ReasonEval [4], our goal is to improve the AUC score in predicting the validity of each step-level solution of the PRM800K dataset.
- Using the ReasonEval-7B model as the verifier, we prepend each retrieved math formula from Dense- and Term-retrieval to each step-level solution.



Macro-scopic Results

	PRM800K	
	AUC	Macro F1
ReasonEval-7B	0.617	0.458
ReasonEval-7B + RAG	**0.654**	**0.483**

Adding RAG context to solution step increase AUC and F1 of the LLM-based verifier by 3.7% and 2.5%.

Micro-scopic Analysis

Q: The quadratic $x^2 + (2.6)x + 3.6$ can be written in the form $(x+b)^2+c$, where b and c are constants. What is $b+c$ (as a decimal)?
Solution step: Then, comparing the constant terms, I need $b^2 + c = 3.6$, so $c = 3.6 - b^2 = 3.6 - 1.3^2 = 1.31$.
RAG: square of a difference is $c^2 - 2bc + b^2 = (c - b)^2$.
ReasonEval-7B: Valid
ReasonEval-7B + RAG: Invalid

References

[1] Jonathan Drechsel, Anja Reusch, and Steffen Herbold. Mamut: A novel framework for modifying mathematical formulas for the generation of specialized datasets for language model training. *arXiv preprint arXiv:2502.20855*, 2025.

[2] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*, 2019.

[3] Stephen Robertson, Hugo Zaragoza, et al. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389, 2009.

[4] Shijie Xia, Xuefeng Li, Yixin Liu, Tongshuang Wu, and Pengfei Liu. Evaluating mathematical reasoning beyond accuracy. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 27723–27730, 2025.