Semantic Search in Olympiad Math Problems

Erbol Esengulov, Filip Illievski

Vrije Universiteit Amsterdam, The Netherlands
erba.omu@gmail.com, f.ilievski@vu.nl

1 Introduction

Recent advances in transformer architectures [8,4] have introduced powerful new ways of representing textual information as meaningful numerical vectors, enabling models to process information at a semantic level and capture relationships beyond simple surface features. Such approaches are already integrated into common search engines [5] and demonstrate great potential across a variety of domains. In textual proofs in olympiad-style mathematics problems, where solutions often rely on novel reasoning, it remains difficult to systematically identify problems that train students to recognize where a particular line of reasoning should be applied. To address this challenge, we investigate transformer-based pipelines for searching similar problems through the semantic understanding of their solutions. While earlier attempts in this direction [3] primarily grouped problems by topic, we believe that solving strategy is a more fine-tuned and informative criterion. Accordingly, our approach explores Number Theory olympiad problems, specifically those from *104 Number Theory Problems* by *Titu Andreescu* [1], and evaluates how effectively transformer models can group problems based on the similarity of their solutions.

Problems from *104 Number Theory Problems* were extracted along with their solutions in order to evaluate the strategy recognition capabilities of BERT-family models, which are well known for their effectiveness in summarization and semantic understanding [7,2]. We designed and tested three distinct datasets: **Textbook Dataset**, **Text Piece Dataset**, and **Ranking Dataset**.

The **Textbook Dataset** consists of problem–solution pairs labeled by book chapter, with the task of clustering problems based on similarity scores derived from BERT embeddings. The **Text Piece Dataset** is an engineered collection of short text segments representing specific mathematical strategies, where the task was again clustering, but this time grouping strategy-oriented text pieces rather than entire problems. Finally, the **Ranking Dataset** is also engineered and consists of tuples of four problems, where one problem serves as the anchor and the model's objective is to rank the remaining three according to their similarity to this anchor.

As shown in Table 1, problems can be clustered into groups, with the correct grouping being $[1, 4]$ and $[2, 3]$. Similarly, in the ranking task from Table 2, the correct output would be $[2, 3, 4]$.

In our experiments, we evaluated the impact of domain adaptation and context window sizes on the ability of LLMs to perform these tasks, using a variety of

ID	Problem Text	True Label
1	Problem. Compute $17 \mod 5$. Solution. Since $17 \div 5 = 3$ with remainder 2, $17 \mod 5 = 2$.	Modular Arithmetic
2	Problem. Find the remainder when 1234 is divided by 7. Solution. $1234 \div 7 = 176$ with remainder 2, so the remainder is 2.	Modular Arithmetic
3	Problem. Using Euler's theorem, find $3^{100} \mod 7$. Solution. Since $\varphi(7) = 6$ and $3^6 \equiv 1 \mod 7$, $100 \mod 6 = 4$, thus $3^{100} \equiv 3^4 = 81 \equiv 4 \mod 7$.	Euler Theorem
4	Problem. Calculate $5^{40} \mod 13$ using Euler's theorem. Solution. Since $\varphi(13) = 12$, $5^{12} \equiv 1 \mod 13$. Because $40 \mod 12 = 4$, $5^{40} \equiv 5^4 = 625 \equiv 1 \mod 13$.	Euler Theorem

Table 1. Example Problems for Clustering Task

ID	Group ID	Problem Text	Label
1	1	Problem. Determine if 72 is divisible by 8. Solution. Since $72 \div 8 = 9$ with no remainder, 72 is divisible by 8.	Anchor
2	1	Problem. Check whether 56 is divisible by 7. Solution. $56 \div 7 = 8$ exactly, so 56 is divisible by 7.	Golden
3	1	Problem. Find the remainder when 56 is divided by 7. Solution. $56 \div 7 = 8$ remainder 0, so the remainder is 0.	Silver
4	1	Problem. Calculate Euler's totient function $\varphi(12)$. Solution. The prime factors of 12 are 2 and 3, so $\varphi(12) = 12 \times (1 - \frac{1}{2}) \times (1 - \frac{1}{3}) = 4$.	Wrong

Table 2. Example Problems for Ranking Task

BERT models. For similarity computation, each dataset element was tokenized into chunks, and the average of the [CLS] token embeddings was recorded. Problem similarity was then measured using cosine similarity, allowing us to identify and compare semantically related instances.

2 Results and discussion

Our investigation shows that current encoder models struggle with mathematical strategy recognition reaching an accuracy of at most 48%. We also found that, in our setup, the context window has a greater impact than domain adaptation. Finally, we observed that training with textbook chapter annotations, such as those in *104 Number Theory Problems* [1], is ineffective, while human-engineered datasets yield more meaningful results. Future work could focus on constructing larger benchmarks that reward a deeper understanding of problem-solving structure for training purposes, as well as on exploring hybrid models that combine symbolic reasoning [6] with encoders. Additionally, improved approaches for handling texts that exceed the model's context window could be investigated. Possible directions include developing a tailored encoder model with a larger context window or exploring alternative methods for embedding long texts beyond simple average pooling.

References

1. Andreescu, T., Andrica, D., Feng, Z.: 104 Number Theory Problems: From the Training of the USA IMO Team. Birkhäuser, Boston, 2007th edn. (2007), print edition (Jan 8, 2007); originally published Dec 19, 2006
2. Aviri, A.W., et al.: The performance of BERT as data representation of text clustering. Journal of Big Data **9**, 15 (2022). https://doi.org/10.1186/s40537-022-00564-9
3. Chen, W., Zhu, X., Zhou, C.: A mathematical question matching and ranking method based on knowledge graph and semantic similarity. In: 2022 IEEE 2nd International Conference on Data Science and Computer Application (ICDSCA). pp. 688–691 (2022). https://doi.org/10.1109/ICDSCA56264.2022.9988532
4. Devlin, J., Chang, M., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. CoRR **abs/1810.04805** (2018), http://arxiv.org/abs/1810.04805
5. Google Cloud: What is semantic search? https://cloud.google.com/discover/what-is-semantic-search (2025), accessed: 2025-06-16
6. Greiner-Petter, A., Youssef, A., Ruas, T., Miller, B.R., Schubotz, M., Aizawa, A., Gipp, B.: Math-word embedding in math search and semantic extraction. Scientometrics **125**(3), 3017–3046 (Dec 2020). https://doi.org/10.1007/s11192-020-03502-9, https://doi.org/10.1007/s11192-020-03502-9
7. Qiao, Y., Xiong, C., Liu, Z., Liu, Z.: Understanding the behaviors of BERT in ranking. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP) (2019)
8. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need (2023), https://arxiv.org/abs/1706.03762