UDC 004

Using language models and data processing techniques to optimize scientific research

Khakimullin, R.(ITMO), **Sim, R.** (ITMO),
Zagorulko, O. (ITMO), **Odobesku, R.** (ITMO)
Scientific supervisor – Razlivina, J. (ITMO)

Introduction: Scientific research plays a pivotal role in the advancement of society and science. However, the processing and analysis of scientific papers often pose a complex challenge, demanding significant effort and time. In light of this issue, we present an innovative tool - DiZyme Assistant, which integrates the power of language models with cutting-edge data processing techniques to facilitate easier access to information in scientific articles. The foundational model is based on research in predicting the catalytic properties of nanozymes and operates using the unique DiZyme[1] database, designed to enhance efficiency in scientific exploration.

Main Body: DiZyme Assistant employs ChatGPT[2] and Multilingual-E5-large for processing user queries, accepting a link to an article and a textual question. The assistant analyzes the article, creating vector representations of parts of the article using Multilingual-E5-large[3]. To expedite the output process, DiZyme Assistant utilizes a faiss index. Faiss is a library for efficient similarity search and clustering of dense vectors. With this technology, the assistant can effectively find relevant information in the article, a key component in providing users with accurate and comprehensive responses. However, mere information extraction is not always sufficient. ChatGPT conducts an analysis of semantic similarities between the question vector and the 'article+context' vectors. This enables the identification of the most relevant excerpts and the creation of a response that synthesizes these fragments into a coherent and easily understandable format.

Conclusions: DiZyme Assistant, by integrating the power of language models and modern data processing technologies, plays a crucial role in simplifying access to information in scientific articles. This tool becomes an integral part of contemporary research, accelerating and streamlining the process of data search and analysis. Consequently, the integration of language model capabilities and data processing methods within DiZyme Assistant is vital in simplifying the analysis of scientific literature, thereby contributing to the enhancement of efficiency and quality in scientific research endeavors.

References

[1] Razlivina, J., Serov, N., Shapovalova, O. & Vinogradov, V. DiZyme: Open‐Access Expandable Resource for Quantitative Prediction of Nanozyme Catalytic Activity. Small 18, 2105673 (2022).

[2] OpenAI. GPT-4 Technical Report. OpenAI, San Francisco, CA, USA (2023).

[3] Wang, L., Yang, N., Huang, X., Jiao, B., Yang, L., Jiang, D., Majumder, R., & Wei, F. (2022). Text Embeddings by Weakly-Supervised Contrastive Pre-training. Submitted to arXiv on December 7, 2022.