Smart Reads: RAG and LLMs for Efficient Scientific Insights

Máté Nándor Szladek (szladekmate@gmail.com)

Budapest University of Technology and Economics, Budapest, Hungary

❖ ABSTRACT

This thesis presents a RAG-based AI system for processing scientific documents in Hungarian and English. It ensures high accuracy by minimizing hallucinations and providing precise, page-level citations—critical in scientific and educational fields. Users can upload curricula and engage in an interactive, tutor-like environment for reliable learning. The system features automatic language detection and a modular design, including document preprocessing, vector storage, question classification, and response generation, optimized for Hungarian and multilingual support. Future improvements aim to enhance storage efficiency, accuracy, and dynamic document updates.

❖ INTRODUCTION AND MOTIVATION

The rapid growth of unstructured text, including scientific literature, presents challenges for NLP, particularly in underrepresented languages like Hungarian. Generative AI often produces hallucinations—false or unverifiable information—posing risks in scientific and academic contexts where accuracy is essential. RAG technology addresses this by grounding responses in user-provided documents, significantly improving reliability. This thesis introduces a RAG-based [fig.1] AI system designed for Hungarian and English, featuring automatic language detection, modular design, and precise citations [fig.2]. By allowing users to upload curricula and engage in tutor-like interactions, the system enhances educational experiences while reducing the risks of misinformation.

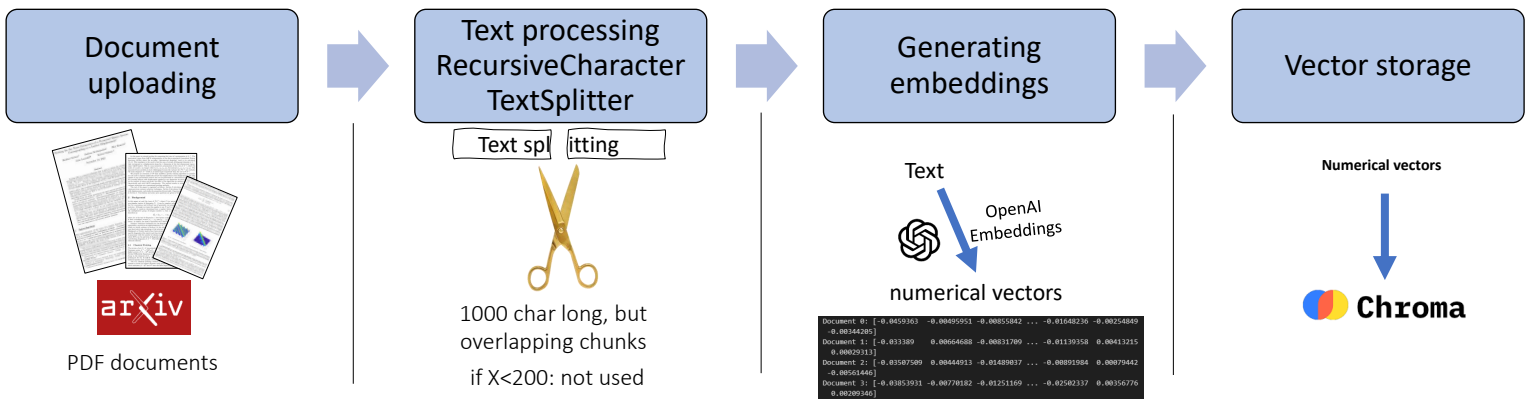

Fig. 1. RAG Pipeline.

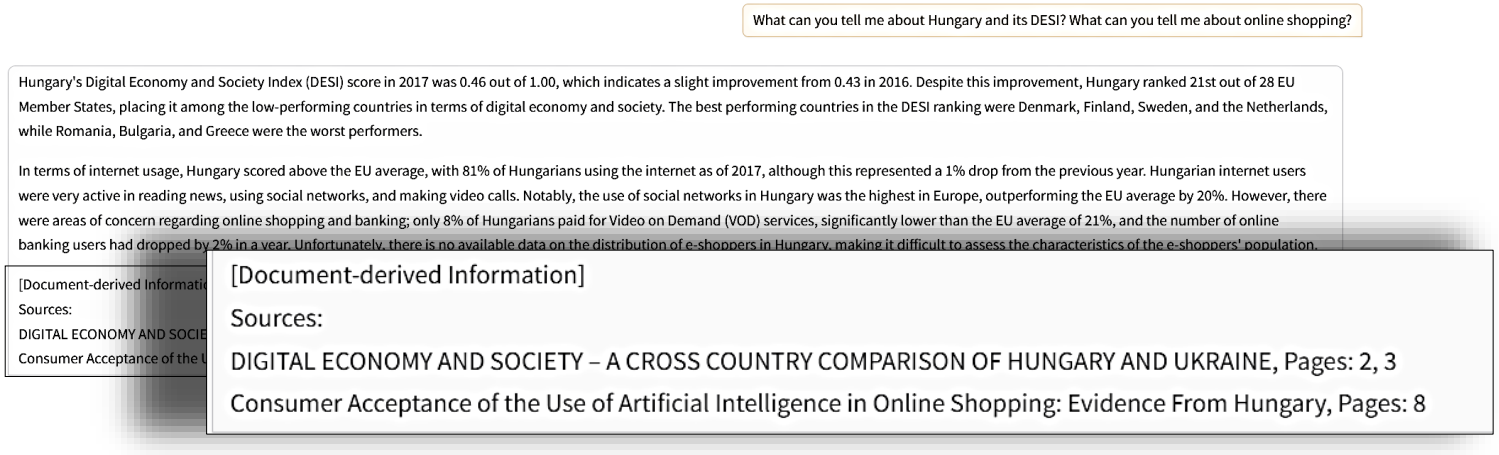

Fig. 2. Example output

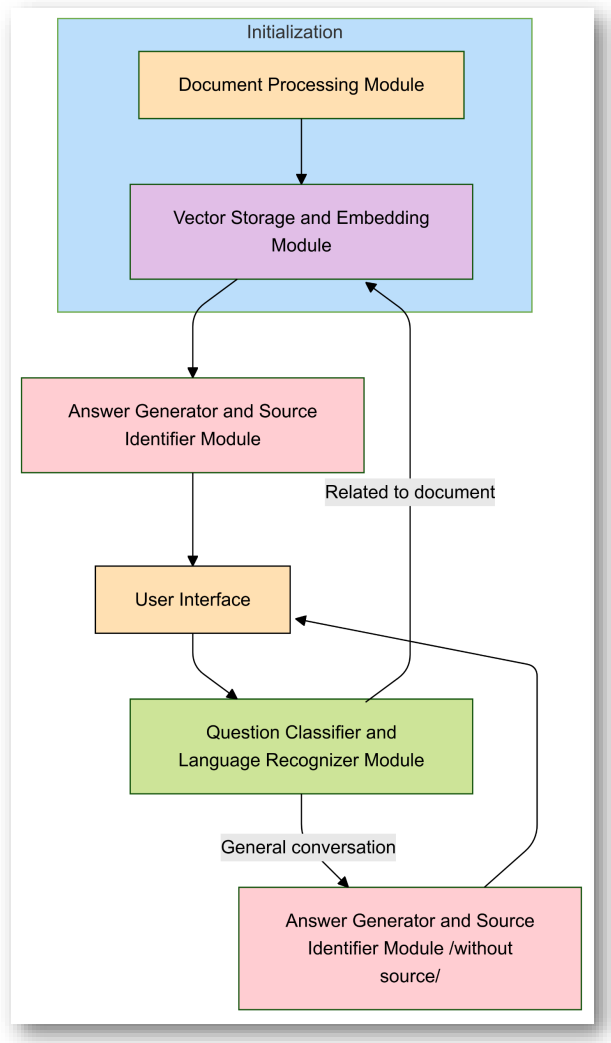

Fig. 3. High-level flowchart

❖ LITERATURE REVIEW

Recent advances in NLP, mostly driven by LLMs like GPT-4 and BERT, have improved applications such as text generation, translation, and summarization (Brown et al., 2020; Reimers & Gurevych, 2019). However, Hungarian's agglutinative nature and complex grammar require tailored approaches (Rehm & Uszkoreit, 2012). Retrieval-Augmented Generation (RAG) combines information retrieval with generative modeling, enhancing accuracy and reducing errors (Lewis et al., 2020). Vector storage technologies like Chroma and FAISS are crucial for managing embeddings in RAG frameworks (Johnson et al., 2021). Despite progress, challenges persist in multilingual support and ethical AI use (Bender et al., 2021; Gehman et al., 2020).

❖ IMPLEMENTATION METHODOLOGY [fig.3]

The system follows a modular design for efficiency and accuracy. Document processing extracts text, segments content, and generates embeddings with OpenAI models. Vector storage is managed with Chroma, [fig.1] enabling fast retrieval via cosine similarity (/BM25/IF-IDF/WMD). Question classification and language detection determine if queries require document-based retrieval or a general response. Response generation integrates retrieved content using RAG, ensuring accuracy (page-level citations). A user-friendly, but prototype interface built with Gradio allows seamless interaction [fig.2]. Future improvements include RAG ReRanking for better retrieval accuracy and automated document updates to enhance system scalability and reliability.

❖ CONCLUSION & PLANS

This RAG-based AI system enhances reliable information retrieval by minimizing hallucinations and providing precise citations. Its modular design ensures efficiency, accuracy, and user accessibility. Future improvements include RAG ReRanking, automated document updates, and expanded multilingual support. Additionally, the system can preload peer-reviewed textbooks and studies, giving students access to trusted materials for reliable learning and research.

❖ HUMAN IMPACT

This initiative takes a human-centered approach to AI, prioritizing safety, clarity, and practicality. Using retrieval-augmented generation (RAG), it minimizes inaccuracies by grounding responses in reliable sources with proper citations, allowing users to trace information back to its origin. This is crucial in education and science, where misinformation can have serious consequences. By focusing on user-provided documents, the system avoids risks from uncontrolled external data. Its interactive, tutor-like features support learning and help users engage with complex information, making it a dependable and accessible resource.

References

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 610–623. https://doi.org/10.1145/3442188.3445922

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., … Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems (NeurIPS)*, 33, 1877–1901.

Gehman, S., Gururangan, S., Sap, M., Choi, Y., & Smith, N. A. (2020). RealToxicityPrompts: Evaluating neural toxic degeneration in language models. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 3356–3369. https://doi.org/10.18653/v1/2020.emnlp-main.265

Johnson, J., Douze, M., & Jégou, H. (2021). Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3), 535–547. https://doi.org/10.1109/TBDATA.2019.2921572

Lewis, P., Oguz, B., Rinott, R., Riedel, S., & Stoyanov, V. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems (NeurIPS)*, 33, 9459–9474.

Rehm, G., & Uszkoreit, H. (2012). The META-NET strategic research agenda for multilingual Europe. *Springer*.

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 3982–3992. https://doi.org/10.18653/v1/D19-1410

Acknowledgements.

The HCAIM (the Human-Centred AI Master's Programme) Project is Co-Financed by the Connecting Europe Facility of the European Union Under Grant NoCEF-TC-2020-1 Digital Skills 2020-EU-IA-0068. This poster was created as part of the Blended Intensive Programme organised under the Erasmus + Programme of the European Union.