EPiC Series in Computing

Volume 107, 2025, Pages 60–69

Proceedings of EUNIS 2025 annual congress in Belfast

Applications of LLM and NLP in the Retrieval and Analysis of Institutional Publications

Luis Filipe de Araujo Pessoa and Raimund Vogl

Center for Information Technology (CIT), University of Münster, Germany
filipe.pessoa@uni-muenster.de, rvogl@uni-muenster.de

Abstract

Modern higher education institutions are increasingly requiring advanced tools to understand and analyze their research output effectively. This paper explores the integration of Large Language Models (LLMs) and Natural Language Processing (NLP) to develop an advanced system for institutional publication analysis at the University of Münster. We demonstrate how Retrieval Augmented Generation (RAG) pipelines can enhance the discovery and analysis of research documents through semantic search and intelligent information processing. Our current approach integrates vector-based document representation with graph-based modeling to construct a RAG pipeline for publication retrieval and analysis. Although still in early development, the system is designed to support several use cases, including the identification of research trends and collaboration patterns, as well as the generation of summaries and reports for specific research topics. We share practical insights from the implementation of this system and discuss technical solutions to common challenges related to the processing and analysis of scientific publications.

1 Introduction

The analysis of institutional scientific and bibliographic production, grounded in scientometrics and the science of science methods [11, 8, 9, 4], has become essential in the context of higher education.

In addition to enabling a better understanding of the evolution of research fields and internal, regional, and international collaboration networks, this analysis allows the identification of areas of excellence, emerging topics, and the tracking of knowledge dissemination and impact metrics. Mapping these thematic and collaborative patterns, combined with specific indicators, is valuable for data-driven strategic decision-making processes, such as planning initiatives to strengthen or expand certain research areas and collaboration networks. Its relevance extends to university administration and planning, benefiting both the central administration and faculties, as well as internationalization and collaboration centers.

Additionally, the retrieval of institutional publications and research projects related to specific topics enables not only the identification of their leading researchers, the most active research groups, and established inter- and intra-institutional collaborations, but also the institution's contribution to different fields of knowledge. The application can also identify potential

L. Desnos, R. Vogl, L. Merakos, C. Diaz, J. Mincer-Daszkiewicz and S. Mclellan (eds.), EUNIS 2025 (EPiC Series in Computing, vol. 107), pp. 60–69

synergies between various research groups and departments within the institution, in addition to highlighting strengths and possible gaps in the scientific production on a given subject. Moreover, the ability to retrieve and analyze institutional publications over time enables a deeper understanding of the evolution, relevance, and impact of different research topics.

In recent years, Large Language Models (LLMs) have transformed the way information is indexed, retrieved, and processed. The rapid advancements in LLMs have led to models with ever-increasing capabilities to represent unstructured information and process complex tasks described in natural language. Embedding models have driven the development of new methodologies for retrieving relevant information based on semantic similarity between user queries and the content of documents in a database, complementing traditional full-text indexing approaches (e.g., TF-IDF or BM25). These transformer-based sentence embeddings offer richer, context-aware representations that can better capture complex semantic information (e.g., synonyms and polysemy) than leaner methods such as Latent Semantic Indexing.

Generative models, in turn, have been employed to generate appropriate responses to user queries. In domains and contexts for which the models were not explicitly trained, generative models can use relevant documents retrieved from a knowledge base to provide more accurate and grounded responses to users, besides reducing hallucinations. This method is known as Retrieval-Augmented Generation (RAG).

This paper reports on experiences and design decisions related to the application of LLMs and Natural Language Processing (NLP) tools in the development of RAG pipelines, explored in the context of institutional publication analysis at the University of Münster. First, we present an overview of the processes involved and discuss some inherent challenges and design choices. Next, we describe investigations and experiments related to information retrieval and the use of generative models to produce appropriate responses to user queries. The paper concludes with directions for future work and final considerations.

2 Overview of the Processes

Figure 1 provides an overview of the processes and elements related to RAG applications, particularly related to the analysis of institutional publications and projects.

First, publication data must be appropriately prepared and integrated into the knowledge base to enable efficient retrieval by the downstream applications. After extracting institutional information on publications and scientific projects through APIs [8] and available internal systems (e.g., Current Research Information System (CRIS), OpenAlex, SemanticScholar, etc.), the data is processed to ensure consistency and accuracy.

This process primarily involves standardizing and aggregating data from different sources, filtering out unnecessary records (e.g., errata publications or retracted papers), handling missing information (e.g., language, identifiers, affiliations, locations, etc.), and detecting and eliminating duplicate entries (i.e., publications, projects, authors, institutions, etc.).

Next, the available metadata related to publication content (primarily titles and abstracts) is transformed into vector representations using an embedding model, enabling semantic search. Additionally, supplementary features may be generated to enhance the existing metadata, facilitating, for instance, information retrieval and analysis of publications (e.g., assigning topics or keywords to represent the subjects covered).

The data can then be stored in the knowledge base using different schemas to retain both raw and processed versions before being made available for the final application. The knowledge base may consist of multiple databases. For example, a relational database might be used

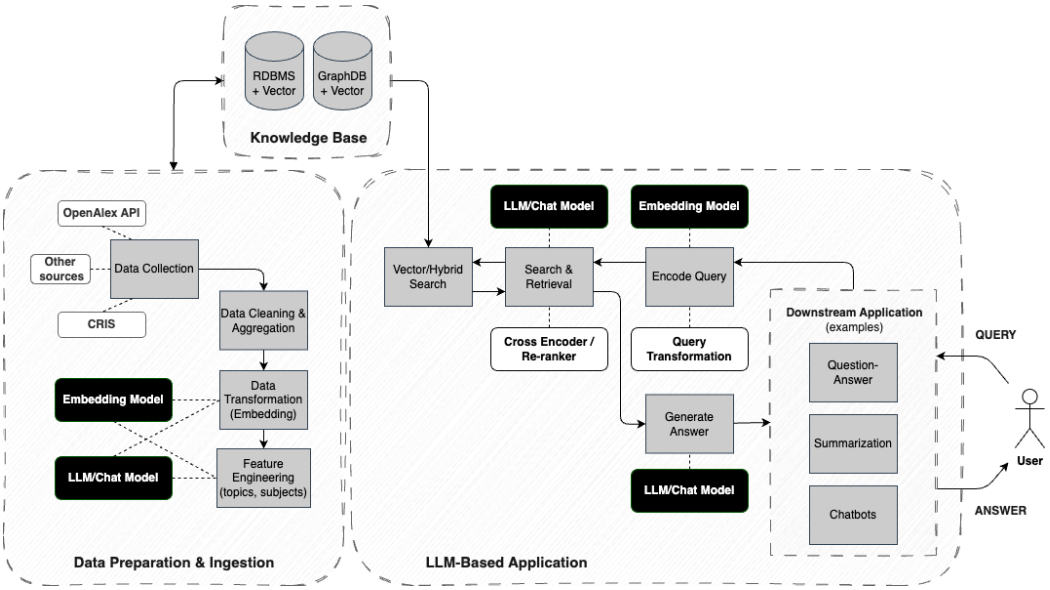


Figure 1: Overview of the processes related to RAG applications in the context of bibliographic analysis.

during the data preparation phase, while a vector database or even a graph database could be employed in the final application.

Afterwards, the LLM-based application queries the knowledge base to retrieve relevant publications and generate appropriate responses to user requests. Use cases may include not only information retrieval through semantic similarity, but also question-answering systems, summarization, and even chatbots.

In order to improve accuracy, the application can employ various strategies, such as rewriting or decomposing user queries (query transformation), leveraging different search indexes (hybrid search), merging results and reordering them by relevance (re-ranking, cross-encoder), and processing retrieved documents to generate the final response (generate answer).

3 Current Experiences, Design Decisions and Applications

This section presents some experiences and design decisions that we have made and explored in the development of RAG applications in the context of information retrieval and the analysis of publications from the University of Münster.

3.1 Knowledge Base

The publication data from the University of Münster was extracted via the OpenAlex API. With approximately 260 million indexed works, OpenAlex is one of the largest public indexers of scientific articles, covering almost all articles indexed by Scopus (100 million) and Web of Knowledge (92 million) [3]. In addition to offering citation and reference coverage comparable

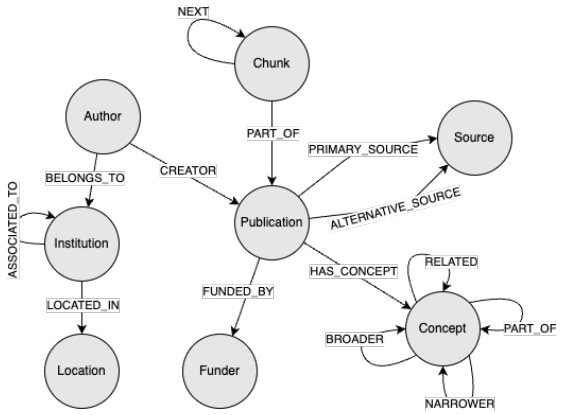


Figure 2: Overview of the main entities and relationships.

to these other databases [3], the quality of its metadata has been continuously improved over the years through the implementation of advanced algorithms [1].

Accordingly, the knowledge base includes all metadata from publications affiliated with the University of Münster (including the University Hospital) that are available in OpenAlex. A graph-based modeling approach (i.e., knowledge graph) was applied, leveraging the entities and metadata extracted from OpenAlex, as illustrated in Figure 2. This approach naturally represents the network structure of publications, authors, institutions, funders, and publication sources, while also facilitating network analyses.

In addition to the metadata retrieved via the OpenAlex API, other information and entities were incorporated into the model (e.g., Chunk and Concept), as explained in Section 3.2. Neo4j was chosen as the database due to its native support for graph representation, its robust query language (Cypher), its support for vector indexing and semantic search, and its seamless integration with frameworks designed for developing LLM- and RAG-based applications (e.g., LangChain).

3.2 Data Preparation and Ingestion

After linking publication data with detailed metadata extracted from OpenAlex (e.g., authors, institutions, source), an initial exploration revealed the need for several data-cleaning procedures. Among them, the removal of irrelevant document types for analysis stood out, such as errata and paratexts (e.g., front/back cover, table of contents, editorial board listings). Publications with generic titles and no abstract were also excluded (e.g., introduction, reply, conclusion, results). Given the identified inconsistencies in language attribution for non-English publications, we employed a procedure to automatically detect the language of those publications, followed by an automated translation of the corresponding titles and abstracts into English.

After the data cleaning process, the knowledge base had approximately 120,000 publications. The titles and abstracts of the publications were then combined, eventually divided into multiple chunks, and later transformed into vector representations. This transformation process is crucial for the development of RAG systems, as it significantly impacts semantic retrieval performance. Overly long text blocks can reduce retrieval accuracy, while excessively short blocks can compromise efficiency [17].

Various strategies have been developed to address this challenge, including static definitions

(e.g., token/character limits, sentence and paragraph segmentation), multi-granularity processing [17], and leveraging LLMs to incorporate contextual metadata into each text block [2]. Moreover, the selection of an appropriate embedding model for the specific use case is critical. It should consider not only its performance on benchmarks [5], but also other factors such as latency, multilingual support, training methodology, context length, among others.

Since the texts to be processed are relatively short (typically fewer than 400 words), we adopted a straightforward chunking strategy based on text length: 512 tokens with an overlap of 25 tokens between consecutive chunks. To convert the text blocks into vector representations, we used the general-purpose model gte-large-en-1.5 [16], as it demonstrated higher accuracy and relevance compared to a model specifically trained for scientific publications (SPECTER [14]) in preliminary experiments.

Each chunk was represented as a node in a graph, linked to its corresponding publication and connected to the subsequent chunk (if multiple chunks were generated for the same publication). Approximately 110,000 publications were associated with a single chunk, 8,000 with two chunks, and around 2,000 with three or more chunks. The occurrence of multiple chunks was mainly due to extended abstracts, short papers, and the inclusion of additional metadata (e.g., reference lists) within the abstract field, highlighting the need for further refinements in future work.

In addition, we used ANNIF to assign keywords (subjects) to each publication, aiming to facilitate thematic categorization, enhance retrieval precision, and enable more comprehensive thematic analyses. ANNIF is an open-source indexing and classification tool developed by the National Library of Finland [15]. This tool enables the construction of configurable NLP-based preprocessing pipelines, supporting various statistical, machine learning, and deep learning methods that can be employed independently or in combination to train a text classifier for a specific corpus and subject vocabulary. Moreover, it enables the reuse of pre-trained models available on the Hugging Face repository.

At this exploratory stage, we employed a pre-trained model based on the subject vocabulary defined by the National Library of Finland, namely the YSO. This vocabulary is structured as a graph of triple statements, incorporating concepts (i.e., terms, keywords, subjects), hierarchies, and relationships between concepts, as well as thesauri and domain-specific groupings. The complete subject vocabulary graph was integrated into our database, allowing traversal through the various relationships and hierarchical structures it defines.

Each publication was assigned the 10 most relevant concepts, determined through a weighted ensemble of the Maui-like Lexical Matching [10], fastText [6], and Bonsai [7] models. The relevance scores were stored as a property of the relationship between each publication and its assigned concepts. Figure 3 provides an overview of the distribution of concepts assigned to publications from the University of Münster, thereby facilitating the identification of the most frequent subject areas and the respective occurrence of terms.

Moreover, the relationships between publications and concepts were leveraged to compute a similarity indicator for the publications. Our hypothesis is that publications sharing multiple common concepts exhibit a certain degree of thematic similarity. Based on this premise, these interconnections were used to calculate the similarity between publication pairs using the Jaccard similarity score. Each publication was linked to its ten most similar counterparts, with the similarity score recorded in the corresponding relationship.

Preliminary tests indicate that this approach effectively captures thematic similarity between publications, as illustrated in Table 1. This result makes the use of similarity-based relationships particularly promising in various applications, including semantic retrieval and the recommendation of related publications based on a given set of selected works for analysis.

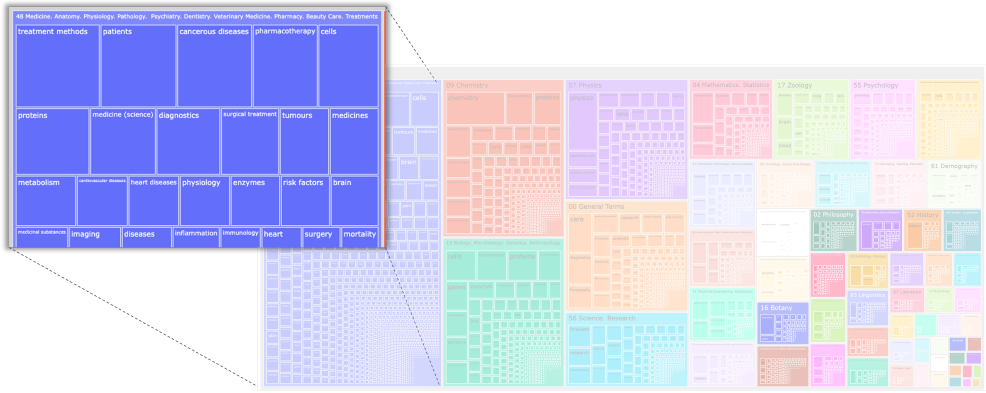


Figure 3: Subject indexing classification for publications of the University of Münster based on the YSO subject vocabulary. Highlighted are some of the most frequent terms for Medicine.

target	1. similar	2. similar	3. similar	4. similar
On the integration of intelligent maintenance and spare parts supply chain management	Application Potentials for an Ontology-based Integration of Intelligent Maintenance Systems and Spare Parts Supply Chain Planning	I2MS2C - intelligent maintenance system architecture proposal	Implementation of constraints satisfaction problems methods for solving the periodic maintenance processes scheduling	Retrograde Terminierung: Ein Verfahren zur Fertigungssteuerung bei diskontinuierlichem Materialfluß oder vernetzter Fertigung
I like, I share, I vote: Mapping the dynamic system of political marketing	The impact of new media on customer relationships	A Hierarchical Marketing Communications Model of Online and Offline Media Synergies	Rezeption von Content Marketing im Spannungsfeld von erwünschten und unerwünschten Medienwirkungen	Picture me in person: Personalization and emotionalization as political campaign strategies on social media in the German federal election period 2021
Scalar curvature rigidity and the higher mapping degree	Positive Scalar Curvature and Homotopy Theory	On homological stability for configuration spaces on closed background manifolds	Twisted spin cobordism and positive scalar curvature	On the Berger conjecture for manifolds all of whose geodesics are closed

Table 1: Examples of the similarity based on the connected concepts. *target* – the target publication; *similar* – the most similar publications ordered by the similarity score.

3.3 LLM-based Application

This section presents some design decisions related to the processes of the LLM-based application currently under development. The applications use the Llama-3.3-70B model via the API provided by UniGPT, a chat platform hosted by the University of Münster [12].

The most basic use case refers to the retrieval of relevant publications based on a user-specified topic description, such as "applications of deep learning for cancer detection." In order to promote higher precision and recall in document retrieval, a hybrid search was adopted that combines the vector and full-text search indexes (built for the title and abstract of the publications). While the full-text search retrieves documents based on the exact occurrence of query terms, the vector search considers the entire context of the user's query, allowing it to retrieve semantically similar documents even if the specified terms do not appear explicitly in the text.

The semantic search is performed directly on the embedding of the topic description provided by the user, while the query used for full-text search is automatically generated by an LLM, which has been carefully instructed with specific guidelines for query transformation, along with illustrative examples. The LLM receives the user's request and converts it into a query following the Apache Lucene syntax (query transformation). Below is the transformed query for the previously mentioned example:

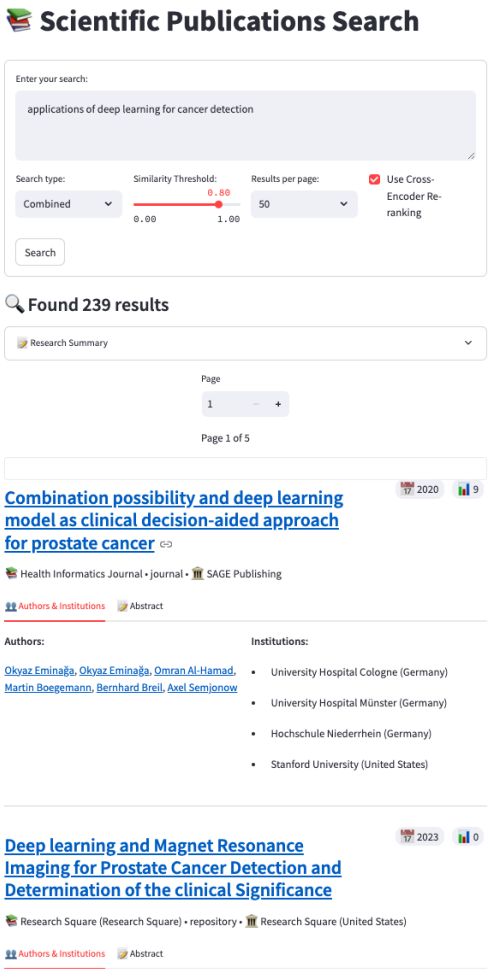
```
(title:("deep learning"^2 OR "machine learning" OR "artificial intelligence") AND
    title:(cancer AND detect*)) OR (abstract:"deep learning"^2 AND
    abstract:(cancer* OR tumor* OR oncology) AND
    abstract:(detect* OR diagnos* OR predict*))
```

To combine the relevance of the documents retrieved by both searches, two approaches were evaluated: Reciprocal Rank Fusion (RRF) and ColBERTv2 as a re-ranker [13]. Although both methods yielded good results, ColBERTv2 demonstrated enhanced consistency and generated a superior relevance ranking in the conducted tests. Nevertheless, it was observed that ColBERTv2 has significantly higher latency compared to RRF, especially when processing a large number of documents (> 100).

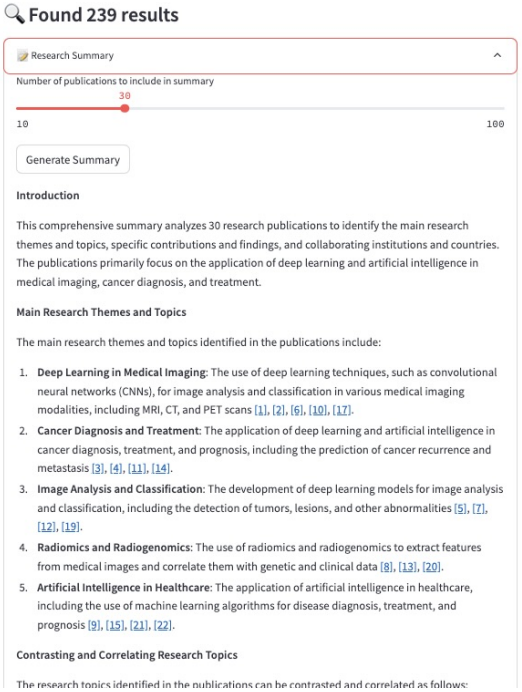
The publications retrieved through this process are then listed for the user, who can leverage them as contextual information for thematic or scientometric analyses. To facilitate quick understanding of the results, an LLM was programmed to generate a structured summary of the most relevant publications, including subtopics, possible correlations among the publications, most prolific authors, key institutional collaborators, and proper citations (see Figure 4). Preliminary tests indicate that the LLM effectively follows instructions and produces summaries that are consistent with the given context.

4 Conclusion and Future Work

This paper presents our experiences and design decisions related to the development of LLM and NLP-based applications for retrieving and supporting the analysis of publications at the University of Münster. We described the main processes involved in implementing a RAG-based application, from data preparation and ingestion to the development of search and retrieval procedures. The knowledge base comprised publication data from OpenAlex. It was modeled as a graph, reflecting the inherent network structure between publications, authors, institutions, and other entities. We additionally discussed the necessary data preparation steps, including data cleaning, language detection and translation, chunking strategies, vector representation,



(a) Overview of the results.



(b) Summary of the most relevant publications (expanded element).

Figure 4: Playground developed in streamlit for prototyping and testing the search, retrieval, and summarization of publications by topic.

and subject indexing using ANNIF. Additionally, we explored the computation of publication similarities based on shared concepts, which demonstrated encouraging results for enhancing semantic retrieval and recommendation.

For the LLM-based application, we implemented a hybrid search approach combining vector and full-text indices and demonstrated how LLMs can be effectively used for query transformation and summarization of the retrieved documents. Despite its higher latency, preliminary tests indicated that the ColBERTv2 re-ranker provided more consistent and relevant results compared to the RRF approach. The developed application successfully generated structured summaries of the most relevant publications related to a user's query, which incorporated information about subtopics, thematic correlations, the most prolific authors, and the most prevalent institutions.

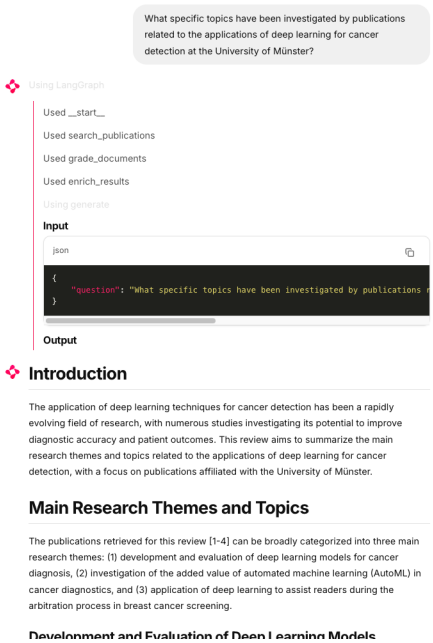


Figure 5: Chatbot for question answering and analysis of the publications (ongoing development). Tracing and displaying the execution / reasoning flow to generate the responses for enabling users to audit the responses' correctness.

Future work should concentrate on key areas for improvement, as well as further developments and integrations. This includes enhancing data quality through (i) a more robust cleaning process that addresses invalid or incorrectly formatted abstracts, (ii) an improved handling of missing and incomplete metadata, and (iii) the exploration of tools and methods for detecting duplicate publications (e.g., ASySD) and disambiguating authors. Furthermore, the aggregation of data from multiple sources can lead to a more comprehensive coverage of the institution's scientific outcomes (e.g., CRIS, Semantic Scholar).

Another promising direction regards the development of a custom subject vocabulary in collaboration with librarians and researchers from the University of Münster, aiming to reflect the institution's specific research areas better and facilitate a more accurate classification of publications. Moreover, the investigation of topic modeling approaches that combine semantic similarity with more traditional methods (e.g., Latent Dirichlet Allocation) could lead to an enhanced identification of research themes.

In order to enhance the retrieval of relevant publications, future work can investigate the integration of concept-based similarity measures into the search process and optimizations of the ColBERTv2 re-ranker to address latency issues. We also plan to leverage LLMs to extract (a priori) structured information from publications that can facilitate fine-grained retrieval and grouping of related work (e.g., research objectives, methods, contributions).

We are currently working on the development of a more capable chatbot for question answering and analysis of the publications, which we plan to integrate into the application (also under development, see Figure 5). In this context, future work should focus on improving the robustness of query understanding and Cypher generation, as well as on the development of an AgenticRAG that processes diverse types of user queries and provides correct responses to

complex requests along with explanations of the processing steps.

These enhancements aim to create a more effective system for institutional bibliometric analysis while leveraging recent advances in LLMs and NLP technologies.

References

[1] Juan Pablo Alperin, Jason Portenoy, Kyle Demes, Vincent Larivière, and Stefanie Haustein. An analysis of the suitability of openalex for bibliometric analyses, 2024.

[2] Anthropic. Introducing contextual retrieval, 2024.

[3] Jack Culbert, Anne Hobert, Najko Jahn, Nick Haupka, Marion Schmidt, Paul Donner, and Philipp Mayr. Reference coverage analysis of openalex compared to web of science and scopus, 2024.

[4] Santo Fortunato, Carl T. Bergstrom, Katy Börner, James A. Evans, Dirk Helbing, Staša Milojević, Alexander M. Petersen, Filippo Radicchi, Roberta Sinatra, Brian Uzzi, Alessandro Vespignani, Ludo Waltman, Dashun Wang, and Albert-László Barabási. Science of science. *Science (New York, N.Y.)*, 359(6379), 2018.

[5] HuggingFace. Mmteb: Massive multilingual text embedding benchmark, 2025.

[6] Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*, 2016.

[7] Sujay Khandagale, Han Xiao, and Rohit Babbar. Bonsai – diverse and shallow trees for extreme multi-label classification.

[8] Zihang Lin, Yian Yin, Lu Liu, and Dashun Wang. Sciscinet: A large-scale open data lake for the science of science research. *Scientific data*, 10(1):315, 2023.

[9] Lu Liu, Benjamin F. Jones, Brian Uzzi, and Dashun Wang. Data, measurement and empirical methods in the science of science. *Nature human behaviour*, 7(7):1046–1058, 2023.

[10] O. Medelyan. *Human-competitive automatic topic indexing*. PhD thesis, The University of Waikato, Hamilton, New Zealand, 2009.

[11] Daniele Pretolesi, Ilaria Stanzani, Stefano Ravera, Andrea Vian, and Annalisa Barla. Artificial intelligence and network science as tools to illustrate academic research evolution in interdisciplinary fields: The case of italian design. *PloS one*, 20(1):e0315216, 2025.

[12] Jonathan Radas, Benjamin Risse, and Raimund Vogl. Building unigpt: A customizable on-premise llm-solution for universities. In Raimund Vogl, Laurence Desnos, Jean-François Desnos, Spiros Bolis, Lazaros Merakos, Gill Ferrell, Effie Tsili, and Manos Roumeliotis, editors, *Proceedings of EUNIS 2024 annual congress in Athens*, volume 105 of *EPiC Series in Computing*, pages 108–116. EasyChair, 2025.

[13] Keshav Santhanam, Omar Khattab, Jon Saad-Falcon, Christopher Potts, and Matei Zaharia. Colbertv2: Effective and efficient retrieval via lightweight late interaction.

[14] Amanpreet Singh, Mike D'Arcy, Arman Cohan, Doug Downey, and Sergey Feldman. Scirepeval: A multi-format benchmark for scientific document representations. In *Conference on Empirical Methods in Natural Language Processing*, 2022.

[15] Osma Suominen, Juho Inkinen, and Mona Lehtinen. Annif and finto ai: Developing and implementing automated subject indexing. *JLIS.it*, 13(1):265–282, Jan. 2022.

[16] Xin Zhang, Yanzhao Zhang, Dingkun Long, Wen Xie, Ziqi Dai, Jialong Tang, Huan Lin, Baosong Yang, Pengjun Xie, Fei Huang, et al. mgte: Generalized long-context text representation and reranking models for multilingual text retrieval, 2024.

[17] Zijie Zhong, Hanwen Liu, Xiaoya Cui, Xiaofan Zhang, and Zengchang Qin. Mix-of-granularity: Optimize the chunking granularity for retrieval-augmented generation, 2024.