Last Minute at SemEval-2025 Task 5: RAG System for Subject Tagging

Zahra Sarlak, Ebrahim Ansari
Institute for Advanced Studies in Basic Sciences (IASBS)
Iran, Zanjan
z.sarlak@iasbs.ac.ir, ansari@iasbs.ac.ir

Abstract

The LLMs4Subjects shared task focuses on utilising Large Language Models to improve subject classification in technical records from the Leibniz University's Technical Library (TIBKAT). Participants are challenged to recommend appropriate subject headings from the GND taxonomy while processing bibliographic data in both German and English. Our approach combines RAG with contrastive learning to refine the embedding model. To further improve retrieval quality, we implement a re-ranking system. We evaluate our model on a test set of TIBKAT records, measuring its performance through precision, recall, and overall classification effectiveness. These findings contribute to the advancement of automated subject classification methodologies in digital library systems, showcasing the potential of large language models (LLMs) in managing multilingual and domain-specific bibliographic data.

1 Introduction

Subject tagging is essential for organizing and retrieving information in vast collections of technical records. The Leibniz Information Centre for Science and Technology (TIB) manages TIBKAT, an open-access bibliographic database that encompasses a significant variety of scientific and technical metadata. To enhance user accessibility, TIB aims to provide precise subject tagging based on the GND (Gemeinsame Normdatei) taxonomy, which is widely used in German-speaking libraries.

The task of manually tagging records is labor-intensive due to the vast number of different subject tags available, leading to potential inconsistencies and inefficiencies.

The LLMs4Subjects shared task(D'Souza et al., 2025) invites researchers to design models capable of processing bilingual technical documents—specifically, those written in German and English. By accurately tagging these documents, systems can significantly enhance the discoverability of information and improve research workflows.

Our study leverages the recent advances in natural language processing (NLP) made possible by retrieval-augmented generation (RAG). This method integrates retrieval mechanisms with language generation, allowing for a more context-aware approach to tagging. Despite much of the focus in NLP being on English and monolingual tasks, the demand for effective bilingual models remains high, especially in library settings.

In our approach, we utilize RAG to dynamically retrieve relevant subject headings from the GND taxonomy based on the technical record's title and abstract[1]. We also employ contrastive learning to fine-tune embedding model for subject tags and incorporate a re-ranking mechanism to optimize our recommendations. By utilizing Milvus (Guo et al., 2022) for efficient vector storage, we aim to enhance both the accuracy and speed of the subject tagging process.

2 Related Work

The integration of automated tagging and structured vocabularies in digital libraries is becoming increasingly important for improving the discoverability of academic content. Recent research has presented various strategies to enhance metadata management, which closely aligns with our use of Retrieval-Augmented Generation (RAG) for tagging in the TIBKAT database.

A key contribution in this area is by (Venktesh et al., 2021), who focus on using a hierarchical learning taxonomy to automatically tag academic questions. Their method addresses the challenges of understanding the relationships between terms in the taxonomy and the questions being tagged. Simi-

[1] https://github.com/jd-coderepos/llms4subjects

Proceedings of the The 19th International Workshop on Semantic Evaluation (SemEval-2025), pages 2432–2436

larly, (Cheng et al., 2023) propose a data-driven approach for analyzing biodiversity subject metadata, using named entity recognition and word embeddings to group related terms effectively. (Aske and Giardinetti, 2023) also highlight the importance of using Artificial Intelligence tools to improve metadata creation, particularly to fix inconsistencies and outdated descriptions in digital archives.

These advances reflect a trend in natural language processing that emphasizes the need to combine retrieval methods with language models to enhance their performance on tasks that require knowledge. A key study in this area is by (Lewis et al., 2020), who introduced RAG. This approach shows that combining a pre-trained sequence-to-sequence language model with a dense vector index allows for better access to external knowledge, leading to improved performance, especially in open-domain question answering.

(Gao et al., 2024) provided a comprehensive survey of RAG methodologies, identifying the advantages of incorporating external databases into language generation. Their work emphasizes that RAG can reduce issues such as hallucination and outdated knowledge by continuously updating and integrating information from external sources. This approach not only enhances the accuracy of generated outputs but also facilitates a more transparent reasoning process, making it a robust framework for knowledge-intensive applications.

In the framework presented by (Khattab et al., 2022), the DEMONSTRATE–SEARCH–PREDICT (DSP) model outlines a sophisticated method for combining retrieval and language models. By creating a pipeline for passing information between these models, the DSP framework seeks to leverage the strengths of both retrieval and generation techniques, achieving new state-of-the-art results in various knowledge-intensive settings. This work exemplifies the potential for improving the interaction between retrieval and language generation, leading to more reliable and coherent output.

(Ovadia et al., 2023) explored the comparison between retrieval-augmented generation and traditional unsupervised fine-tuning for knowledge injection in large language models. Their findings suggest that while fine-tuning can provide benefits, RAG consistently outperforms it, particularly in integrating new knowledge and enhancing the models' overall capabilities. This study underscores the limitations of conventional methods and highlights the growing importance of retrieval-based approaches in effectively updating language model knowledge.

Lastly, (Ram et al., 2023) examined In-Context Retrieval-Augmented Language Models (RALMs), presenting a simple yet effective method for conditioning language models on relevant documents without altering their architecture. This approach focuses on maintaining the existing model while improving performance through external document incorporation, which can simplify deployment and usability. The findings suggest that leveraging retrieval mechanisms can lead to significant advantages in language modeling without necessitating extensive changes to the underlying model architecture.

These studies shows the impact of retrieval-augmented methods in enhancing the capabilities of language models, particularly for tasks that require accurate knowledge retrieval and integration.

3 Methodology

Our approach to the LLMs4Subjects shared task involved multiple key steps designed to enhance the accuracy and efficiency of subject tagging for technical records from the TIBKAT database. Below, we outline the methodology employed:

1. **Model Fine-Tuning for Embedding:** Since the main retrieval method relies on cosine similarity searching, it is crucial to have an effective embedding model that emphasizes the distinction between subject-related and unrelated content. With this in mind, we began by fine-tuning the stella-en-400M-v5 model (Zhang et al., 2025), recognized as one of the most effective lightweight models in the MTEB (Muennighoff et al., 2023) benchmark for embedding tasks[2]. The model was fine-tuned using the training data with the Multiple Negatives Ranking Loss (Henderson et al., 2017) for one epoch. Using more epochs would increase the risk of overfitting the model. The training was conducted under the following configurations:

- Number of training epochs: 1
- Per-device training batch size: 32
- Learning rate: 1e-5
- Floating point precision: bf16

[2] https://huggingface.co/spaces/mteb/leaderboard

- Gradient accumulation steps: 2

These settings were chosen to strike a balance between effective model training and generalization, ensuring that the model could learn meaningful representations without becoming overly specialized to the training data.

2. **Embedding Subject Tags:** Once the model was fine-tuned, we created embeddings for subject tags using the trained embedding model. For each subject in the GND dataset, we have the following fields: *Name* and *Classification Name*. Additionally, for a portion of the subjects, the fields *Definition* and *Related Subjects* are also available.

The embeddings for each subject tag were generated by concatenating *Name* and *Classification Name* along with *Definition* and *Related Subjects* when available. This comprehensive representation effectively captures the contextual meaning of each tag.

3. **Vector Storage with Milvus:** After embedding the subject tags, they were stored in Milvus vector storage. This enabled us to efficiently retrieve and manage the high-dimensional vectors associated with the subject tags during the later stages of our methodology.

4. **Retrieving Similar Tags:** For each record in the test dataset, the *title* and *abstract* fields were concatenated and embedded using the same embedding model that was used for subject tags. Then, we retrieved a set of 100 similar subject tags (measured by cosine similarity) from the Milvus database for the embedded representations of the test records. This retrieval process enabled us to identify potential tags that could be relevant to the given technical records based on their embeddings.

5. **Re-ranking with a Cross-Encoder Model:** To refine the initial set of retrieved tags, we employed a cross-encoder model from the MTEB benchmark (Muennighoff et al., 2023), focusing solely on its re-ranking capabilities. As of the time of writing this paper, the granite-embedding-278m-multilingual (Granite Embedding Team, 2024) model is among the top 30 low-memory models in the MTEB benchmark (Muennighoff et al., 2023) sorted by re-ranking score. We used this model to re-rank the 100 similar tags. This model evaluated the relevance of each tag in relation to the title and abstract, resulting in a more accurate ranking of the tags.

6. **Final Tag Selection:** From the re-ranked list, we selected the top 70 candidate tags. These were then passed to a large language model (Llama-3.2-1B (Grattafiori et al., 2024)) as part of a prompt to further refine the results. The LLM evaluated the tags based on contextual understanding and relevance to the input data. The prompt we passed to the model:

Given the following abstract of a technical document:

[title + abstract]

And the top retrieved subject tags:

[Retrieved tags]

Please assess the relevance of each tag in relation to the abstract provided. Sort the tags based on their appropriateness, and select the top 5 most relevant subject tags that best represent the content of the abstract.

Finally, we extracted the top 50 tags recommended by the LLM as the final output for each title and abstract. This multi-step methodology, combining embedding, tagging retrieval, re-ranking, and LLM refinement, aimed to enhance the overall accuracy and effectiveness of subject classification for the technical records in the TIBKAT database.

4 Result

The initial results of our tagging approach revealed several challenges in the subject tagging process, despite the potential of Retrieval-Augmented Generation (RAG). These challenges underscore the need for further refinement and optimization. However, they also provide valuable insights of subject classification in a multilingual and domain-specific context. In this section, we present the detailed evaluation results.

The evaluation of our subject tagging system was carried out by the SemEval team through both quantitative and qualitative assessments. Table 3 shows the overall evaluation result for quantitative and qualitative evaluation result, representing the Average precision, Recall and F1 for different evaluation methods.

The combined language and record-levels results is available at appendix A, with the separated evaluation result for different record types and languages. Tables 4-6 each represented result at k=5,10,15 respectively.

Having better recall for more @K is natural, because the more subject retrieved increase the chance of more correct tag selection.

K	Precision	Recall	F1
@5	0.0241	0.0459	0.0316
@10	0.0224	0.0848	0.0354
@15	0.0213	0.1223	0.0363
@20	0.0207	0.1587	0.0366
@25	0.0201	0.1940	0.0365
@30	0.0199	0.2316	0.0367
@35	0.0196	0.2669	0.0366
@40	0.0194	0.2998	0.0364
@45	0.0191	0.3332	0.0361
@50	0.0188	0.3623	0.0357

Table 1: Quantitative result of the proposed approach on tib-core dataset, evaluated by SemEval organizers. Rows of tables shows scores at different values of k (5, 10, and 15), where @k indicates the number of top tags retrieved by our model.

K	Precision	Recall	F1
case1 @5	0.2719	0.1491	0.1926
case1 @10	0.2369	0.2525	0.2445
case1 @15	0.2138	0.3226	0.2572
case1 @20	0.2202	0.4426	0.2941
case2 @5	0.1288	0.1025	0.1142
case2 @10	0.1086	0.1749	0.1340
case2 @15	0.0998	0.2275	0.1387
case2 @20	0.1048	0.3090	0.1565

Table 2: Qualitative result of the proposed approach on tib-core dataset, evaluated by SemEval organizers. Rows of tables shows scores at different values of k (5, 10, and 15), where @k indicates the number of top tags retrieved by our model.

5 Challenges

One of the main reasons for the low scores is that, despite fine-tuning the embedding model, the distance between related and unrelated subject embeddings in the embedding space is still not adequate. Since extracting relevant subject tags heavily relies on cosine similarity, the distance between subject embeddings is crucial; however, achieving this distinction through fine-tuning the embedding model proves to be challenging. It is important to note that a subject may be related to one record while being unrelated to another, yet all of them may have near vectors in the vector space. Fine-tuning the model may improve some similarities, but it can also distort others, further complicating the differentiation process. Therefore, fine-tuning must be implemented with great care and involve iterative reviews of the results.

Additionally, the vast, highly imbalanced, and diverse set of subject tags, along with the need to embed and index this extensive set of tags in vector storage, makes evaluation challenging and limits our options for testing different embedding models.

Evaluation	Average Recall	Average Precision	Average F1
Quantitative	0.2099	-	-
Qualitative case 1	0.2917	0.2357	0.2471
Qualitative case 2	0.2035	0.1105	0.1359

Table 3: Result of the proposed approach on tib-core dataset, evaluated by SemEval organizers

6 Conclusion

We examined the effectiveness of Retrieval-Augmented Generation (RAG) for tagging subjects in technical records from the TIBKAT database. By fine-tuning a lightweight language model and integrating a retrieval mechanism, we aimed to improve the accuracy and efficiency of the tagging process. The challenge of accurately tagging subjects is heightened by the vast number of subject tags—over 240,000—making it difficult to achieve effective results with simple similarity searches. For future work, we propose adopting a structured approach, such as using graph-based representations, to store subject tags and utilize their hierarchical taxonomy for more precise candidate selection. Additionally, since our solution relies heavily on similarity search, evaluating its effectiveness poses challenges due to the large search space. To address this, we can implement heuristics that limit the search set for each abstract sample, ultimately improving both efficiency and accuracy in the tagging process.

References

Katherine Aske and Marina Giardinetti. 2023. (mis)matching metadata: Improving accessibility in digital visual archives through the eycon project. *J. Comput. Cult. Herit.*, 16(4).

Yi-Yun Cheng, Nikolaus Nova Parulian, and Ly Dinh. 2023. A text mining approach to uncover the structure of subject metadata in the biodiversity heritage library. *Proceedings of the Association for Information Science and Technology*, 60(1):926–928.

Jennifer D'Souza, Sameer Sadruddin, Holger Israel, Mathias Begoin, and Diana Slawig. 2025. Semeval-2025 task 5: Llms4subjects - llm-based automated subject tagging for a national technical library's open-access catalog. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 1082–1095, Vienna, Austria. Association for Computational Linguistics.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang,

and Haofen Wang. 2024. Retrieval-augmented generation for large language models: A survey. *Preprint*, arXiv:2312.10997.

IBM Granite Embedding Team. 2024. Granite embedding models.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Rentong Guo, Xiaofan Luan, Long Xiang, Xiao Yan, Xiaomeng Yi, Jigao Luo, Qianya Cheng, Weizhi Xu, Jiarui Luo, Frank Liu, and 1 others. 2022. Manu: a cloud native vector database management system. *Proceedings of the VLDB Endowment*, 15(12):3548–3561.

Matthew Henderson, Rami Al-Rfou, Brian Strope, Yun hsuan Sung, Laszlo Lukacs, Ruiqi Guo, Sanjiv Kumar, Balint Miklos, and Ray Kurzweil. 2017. Efficient natural language response suggestion for smart reply. *Preprint*, arXiv:1705.00652.

O. Khattab, Keshav Santhanam, Xiang Lisa Li, David Leo Wright Hall, Percy Liang, Christopher Potts, and Matei A. Zaharia. 2022. Demonstrate-search-predict: Composing retrieval and language models for knowledge-intensive nlp. *ArXiv*, abs/2212.14024.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc.

Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. 2023. Mteb: Massive text embedding benchmark. *Preprint*, arXiv:2210.07316.

Oded Ovadia, Menachem Brief, Moshik Mishaeli, and Oren Elisha. 2023. Fine-tuning or retrieval? comparing knowledge injection in llms. In *Conference on Empirical Methods in Natural Language Processing*.

Ori Ram, Yoav Levine, Itay Dalmedigos, Dor Muhlgay, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. 2023. In-context retrieval-augmented language models. *Transactions of the Association for Computational Linguistics*, 11:1316–1331.

V. Venktesh, Mukesh K. Mohania, and Vikram Goyal. 2021. Tagrec: Automated tagging of questions with hierarchical learning taxonomy. In *ECML/PKDD*.

Dun Zhang, Jiacheng Li, Ziyang Zeng, and Fulong Wang. 2025. Jasper and stella: distillation of sota embedding models. *Preprint*, arXiv:2412.19048.

A Appendix A: Detailed Result

Record Type	Language	Precision	Recall	F1
Article	en	0.0778	0.0778	0.0778
Book	de	0.0255	0.0483	0.0334
Book	en	0.0237	0.0490	0.0319
Conference	de	0.0269	0.0288	0.0278
Conference	en	0.0243	0.0540	0.0335
Report	de	0.0179	0.0528	0.0267
Report	en	0.0238	0.0437	0.0308
Thesis	de	0.0223	0.0267	0.0243
Thesis	en	0.0188	0.0324	0.0238

Table 4: Detailed Quantitative Result by Record type and language, on tib-core dataset, at k=5, where k indicates the number of top tags retrieved by our model

Record Type	Language	Precision	Recall	F1
Article	en	0.0444	0.0889	0.0593
Book	de	0.0254	0.1003	0.0405
Book	en	0.0214	0.0871	0.0343
Conference	de	0.0288	0.0699	0.0408
Conference	en	0.0236	0.0963	0.0379
Report	de	0.0241	0.1037	0.0391
Report	en	0.0198	0.0700	0.0309
Thesis	de	0.0185	0.0429	0.0258
Thesis	en	0.0154	0.0447	0.0229

Table 5: Detailed Quantitative Result by Record type and language, on tib-core dataset, at k=10, where k indicates the number of top tags retrieved by our model

Record Type	Language	Precision	Recall	F1
Article	en	0.0370	0.1148	0.0560
Book	de	0.0241	0.1434	0.0413
Book	en	0.0206	0.1283	0.0356
Conference	de	0.0257	0.0946	0.0404
Conference	en	0.0216	0.1332	0.0372
Report	de	0.0232	0.1406	0.0399
Report	en	0.0175	0.0946	0.0295
Thesis	de	0.0178	0.0619	0.0277
Thesis	en	0.0155	0.0649	0.0250

Table 6: Detailed Quantitative Result by Record type and language, on tib-core dataset, at k=15, where k indicates the number of top tags retrieved by our model