SemEval-2025 Task 5: LLMs4Subjects - LLM-based Automated Subject Tagging for a National Technical Library's Open-Access Catalog

Jennifer D'Souza, Sameer Sadruddin, Holger Israel, Mathias Begoin, and Diana Slawig
TIB Leibniz Information Centre for Science and Technology, Hannover, Germany
{firstname.lastname}@tib.eu

Abstract

We present SemEval-2025 Task 5: LLMs4Subjects, a shared task on automated subject tagging for scientific and technical records in English and German using the GND taxonomy. Participants developed LLM-based systems to recommend top-k subjects, evaluated through quantitative metrics (precision, recall, F1-score) and qualitative assessments by subject specialists. Results highlight the effectiveness of LLM ensembles, synthetic data generation, and multilingual processing, offering insights into applying LLMs for digital library classification.

1 Introduction

Subject classification within library systems involves organizing books and resources based on their content and subject matter to facilitate easy retrieval and access. Automated methods for classifying scientific texts include Springer Nature's CSO classifier (Salatino et al., 2019), which uses syntactic and semantic analysis to categorize papers based on the Computer Science Ontology (CSO) using abstracts, titles, and keywords. Other approaches have utilized citation metadata for classification (Mahdi and Joorabchi, 2011), while research like New Mexico State University's application of topic modeling on digital news releases (Glowacka-Musial, 2022) demonstrates the diversity of techniques. Methods range from embeddings (Buscaldi et al., 2017; Růžička and Sojka, 2021) and deep learning (Ahanger and Wani, 2022) to citation analysis (Small et al., 2014) and topic modeling (Bolelli et al., 2009; Griffiths and Steyvers, 2004). Thus, while the use of NLP in digital library subject classification is well-established (Gooding et al., 2019), the potential for leveraging Large Language Models (LLMs) for their extensive knowledge representation remains largely unexplored.

Several open-source toolkits integrate machine learning (ML) and NLP for automated subject indexing, notably ANNIF (https://annif.org/), developed by the National Library of Finland, as a DIY automated subject indexing tool supporting the training of multiple traditional machine learning algorithms (Suominen, 2019). ANNIF allows users to train models on a chosen subject taxonomy and metadata to generate subject headings for new documents. It has performed well on scientific papers and books but struggles with older or diverse materials like Q&A pairs or Finnish Wikipedia. European national libraries, including Sweden's National Library, the Leibniz Information Centre for Economics, and the German National Library, have adopted ANNIF. Supporting multiple languages and vocabularies, it offers command-line, web, and REST API interfaces, demonstrating the adaptability required for effective subject classification.

With these insights, as a SemEval 2025 shared task, we organized Task 5 — LLMs4Subjects — to explore the untapped potential of LLMs for subject classification and tagging. The task was defined on the catalog of the TIB – Leibniz Information Centre for Science and Technology, Germany's national library for science and technology. Its catalog, TIBKAT, holds 5.7 million records (as of March 2025), including bibliographic data and metadata from freely available electronic collections. A subset of around 100,000 records is available as open access, and the task focused on this subset. The collection includes various record types such as technical reports, publications, and books, primarily in English and German. Regardless of full-text language, records are consistently annotated using subject terms from the Gemeinsame Normdatei (GND), the integrated authority file and subject taxonomy used in the German library system. LLMs offer promising opportunities for subject classification through their ability to process natural language at scale and capture the nuances of complex, interdisciplinary topics. This can significantly improve the accuracy and efficiency of organizing

Proceedings of the The 19th International Workshop on Semantic Evaluation (SemEval-2025), pages 2570–2583

large collections, enhancing the accessibility and discoverability of information. The solutions developed in this task serve as a benchmark for applying LLMs in digital library systems, fostering innovation and setting new standards in the field. Moreover, the task aligns with the goals of the SemEval series by evaluating a novel application of computational semantics essential for effective information organization and retrieval.

While the shared task focused on practical systems development, it was also driven by the following four research questions. **RQ1** Multilingual vs. Monolingual Models in Subject Tagging: How do multilingual pre-trained models compare to monolingual models in bilingual subject tagging tasks?, **RQ2** Effect of Training Data Size and Diversity: How does the size and diversity of training data affect LLM performance in subject tagging?, **RQ3** Role of Augmented Generation: How impactful are RAG approaches versus finetuning?, and **RQ4** Efficiency of Language Models: How effective are small versus large LMs in performing the task?

A key observation from the systems submitted to this shared task is that the advantages of LLMs over traditional machine learning algorithms for subject indexing remain debatable (Kluge and Kähler, 2024). While the first iteration of this shared task brings this question to light, to further explore the possibilities offered by LLMs, we will reorganize the LLMs4Subjects shared task a second time and this time with a theme to build solutions based on *energy- and compute-efficient LLMs*.

2 Background

About TIB. TIB – the Leibniz Information Centre for Science and Technology and University Library – promotes free access to knowledge, information sharing, and open scientific publications and data. As Germany's national library for science and technology, including Architecture, Chemistry, Computer Science, Mathematics, and Physics, it maintains a globally unique collection, including audiovisual media and research data.

About the TIBKAT Open-access Subset. A subset of TIB's collection—including bibliographic data in science and technology from its library catalog (TIBKAT Data), metadata from freely available electronic collections, and metadata with thumbnails from the TIB AV-Portal—is made available under the CC0 1.0 Universal Public Domain Dedication, allowing unrestricted use. More here.

The GND Taxonomy. The Gemeinsame Normdatei[1] (GND, German for "integrated authority file") is an international authority file used primarily by German-speaking library systems to catalog and link information on topics, organizations, people, and works. It is publicly available for download[2] in various formats under a CC0 license. The GND's records specifically cover entities such as persons, corporate bodies, conferences, geographic locations, subject headings, and works, relevant to cultural and scientific collections.

For the LLMs4Subjects shared task, only the GND subject heading (Sachbegriff) records are of interest. Since accessing the GND for the first time for new users can be overwhelming, for the convenience of our participants, we have created a how-to guide to download the latest GND file.

3 Source Dataset

We queried the TIBKAT service to restrict its metadata to records containing abstracts and GND subject indexing. The query is fully reproducible via this persistent search link. It returned 189,665 records at the time of dataset creation. The TIB open-access catalog spans nine media types: Book (136,434), Thesis (31,859), Conference (12,212), Report (6,711), Article (2,080), Collection (188), AudioVisualDocument (167), Periodical (57), and Chapter (11), detailed in Appendix A. Using the `langdetect`[3] Python library, we identified 48 languages. The top five were German (108,637), English (76,735), French (1,741), Indonesian (945), and Spanish (311).[4] For the official shared task corpus, we retained only records in German and English and excluded the four least represented media types, resulting in a dataset of 123,589 records. The excluded data is available as supplementary data. The final shared task dataset is available at: `https://github.com/jd-coderepos/llms4subjects/tree/main/shared-task-datasets`.

3.1 How are subject annotations obtained?

Subject annotations in the TIB catalog are continuously created by a dedicated team of

[1] `https://www.dnb.de/EN/gnd`

[2] The GND is available for download at `https://www.dnb.de/EN/Professionell/Metadatendienste/Datenbezug/Gesamtabzuege/gesamtabzuege_node.html`.

[3] `https://pypi.org/project/langdetect/`

[4] Reflecting the real-world nature of the corpus, many records contain mixed-language content and are not reliably classifiable under a single language.

statistics	lang	Article	Book	Conference	Report	Thesis
num. records	en	1,042/253	26,966/17,669	3,619/2,840	1,275/896	3,452/2,506
	de	6/5	33,401/12,528	2,210/717	1,507/761	8,459/3,727
num. subjects (avg, max)	en	(3/4, 7/6)	(3/3, 39/26)	(3/3, 14/16)	(3/3, 12/13)	(4/4, 20/19)
	de	(3/3, 8/7)	(3/3, 27/25)	(3/4, 17/16)	(3/3, 15/15)	(4/4, 20/19)

Table 1: Train dataset statistics (all-subjects/tib-core collections) for the LLMs4Subjects shared task.

17 expert subject specialists covering 28 disciplines—including Architecture, Chemistry, Electrical Engineering, Mathematics, Traffic Engineering, and others—ensuring broad and expert-driven subject classification.

In libraries, content is typically described using controlled vocabularies. In Germany, the GND is used for cataloging literature. In addition to descriptive cataloging (e.g., author, title, year, publisher), subject cataloging is performed by subject librarians. Based on the title, abstract, and full text, librarians assign appropriate GND keywords to describe the content as precisely as possible. This collaborative work is carried out across various libraries and national library networks.

With TIB adding around 15,000 new titles each month, subject cataloging is a labor-intensive task. Integrating AI-driven solutions—especially LLMs—can significantly boost efficiency, partially automate workflows, and improve usability, all while maintaining cataloging quality. Such innovations are key to modernizing information management and supporting research at scale.

4 Shared Task Description and Dataset

The LLMs4Subjects shared task challenged participants to develop LLM-based systems for recommending relevant subjects from the GND taxonomy to annotate TIB technical records. Given a record's title and abstract as input, systems were expected to generate a customizable top-k ranked list of relevant GND subjects. Since the dataset included records in both English and German, systems were required to support bilingual semantic processing. The task was defined over two dataset collections.

4.1 Dataset Collections

all-subjects.[5] This dataset comprises the full TIBKAT open-source collection, with predefined splits: 81,937 records for training and 13,666 for development. A detailed dataset overview is available in the shared task repository. Participants also received the accompanying GND subject taxonomy,[6] which included 204,739 subjects, with coverage and distribution frequencies published online.

Due to the large dataset size (>100,000 records), participants could opt for a smaller subset focused on TIB's core subject classification.

tib-core.[7] This subset includes only records annotated with at least one GND subject from the so-called TIB core domains. It contains 41,902 training and 6,980 development records across 14 domains: Architecture (arc), Civil Engineering (bau), Mining (ber), Chemistry (che), Chemical Engineering (cet), Electrical Engineering (elt), Materials Science (fer), Information Technology (inf), Mathematics (mat), Mechanical Engineering (mas), Medical Technology (med), Physics (phy), Engineering (tec), and Traffic Engineering (ver). A refined GND subject taxonomy[8] with 79,427 subjects accompanied this dataset, along with subject coverage and frequency distributions.

Participants could choose between the all-subjects dataset for comprehensive indexing, the tib-core dataset for a more focused classification task, or even attempt both.

4.2 Dataset Format and Statistics

Both datasets were released in JSON-LD format and include metadata such as title, type, abstract, and authors. The key attribute for the task, `dcterms:subject`, holds the GND subject headings assigned to each record. Example records are available in English here and in German here. Table 1 provides a detailed dataset breakdown. Books were the most common record type in both language collections, with each record annotated with an average of 3 to 7 subjects.

[5] https://github.com/jd-coderepos/llms4subjects/tree/main/shared-task-datasets/TIBKAT/all-subjects

[6] https://github.com/jd-coderepos/llms4subjects/blob/main/shared-task-datasets/GND/dataset/GND-Subjects-all.json

[7] https://github.com/jd-coderepos/llms4subjects/tree/main/shared-task-datasets/TIBKAT/tib-core-subjects

[8] https://github.com/jd-coderepos/llms4subjects/blob/main/shared-task-datasets/GND/dataset/GND-Subjects-tib-core.json

5 Task Setup

The shared task offered multiple communication channels: a dedicated website,[9] a Google Group for FAQs, and direct email support (llms4subjects@gmail.com). The organizing team included two Computer Scientists and three TIB subject specialists who supported participants throughout. The task timeline spanned four months, from October 2024 to January 2025. A Declaration of Interest for Participation (DIP) survey initially recorded **33 teams**; of these, **14 teams** submitted system outputs, and **12** also contributed system description papers to the SemEval workshop. The next section summarizes their approaches.

6 Shared Task Participant Systems

The participant systems are described with a focus on their key methodological contributions. Teams are listed in alphabetical order of their names. An overview of the systems is provided in Table 2.

1. Annif (Suominen et al., 2025) The key ideas in this system's subject tagging approach were: 1) **Traditional extreme multi-label text classification (XMTC)** implemented in the Annif toolkit (2022) – Using Omikuji Bonsai (Khandagale et al., 2020), a tree-based machine learning approach; MLLM (Maui-like (Medelyan, 2009) Lexical Matching), a lexical matching algorithm; and XTransformer, a transformer-based classification model (Yu et al., 2022) for XMTC. 2) **Ensemble Models** – Combining individual classifiers into simple averaging and neural ensembles to improve predictions. 3) **LLM-Assisted Translation** – Using the Llama-3.1-8B-Instruct LLM (2024) to translate bibliographic records and subject vocabularies into English and German. 4) **Synthetic Data Generation** – Expanding training data with the same LLM by generating new records with modified subject labels. And 5) **Multilingual Merging** – Combining monolingual predictions to form a multilingual ensemble, improving overall performance.

2. DNB-AI-Project (Kluge and Kähler, 2025) This was an LLM-driven ensemble approach which achieved top qualitative scores without fine-tuning and few-shot prompting. Key steps included: 1) **LLM Ensemble for Keyword Generation** – Multiple off-the-shelf LLMs use few-shot prompting with 8-12 examples to improve recall and precision. 2) **Map** – A BGE-M3 embedding model (Chen et al., 2024) maps LLM-generated free keywords to controlled GND subject terms via nearest neighbor search. 3) **Summarize** – Predictions from the LLM ensemble are aggregated, with similarity scores summed and normalized into a confidence-based score. 4) **Rank** – A new LLM, Llama-3.1-8B-Instruct (2024) assesses relevance of each predicted term on a 0-10 scale, refining rankings beyond frequency-based measures. And 5) **Combine** – Ensemble and relevance scores are weighted and combined to optimize subject ranking.

3. DUTIR831 (Tian et al., 2025b) The key steps of this system are: 1) **Data Synthesis and Filtering** – The Qwen2.5-72B-Instruct LLM is used to generate synthetic data to expand training sets by selecting related subject terms and creating titles and abstracts. The LLM is then applied to filter low-quality samples based on coherence and relevance. 2) **GND Knowledge Distillation** – The LLM is then finetuned on GND subject collections improves its understanding of subject hierarchies and relationships. 3) **Supervised Fine-Tuning and Preference Optimization** – LoRA-based (Hu et al., 2022) fine-tuning on TIBKAT data is combined with Direct Preference Optimization (DPO) (Rafailov et al., 2023) to align model outputs with human-like subject assignments. 4) **Subject Term Generation** – A multi-sampling ranking strategy improves diversity, LLM-based keyword extraction selects high-confidence terms, and BGE-M3 (2024) embedding-based vector retrieval adds missing terms to ensure 50 subject labels per record. And 5) **Re-Ranking for Final Selection** – Subject terms from multiple sources are re-ranked using LLMs to prioritize the most relevant terms, improving recall and ranking consistency.

4. Homa (Bayrami Asl Tekanlou et al., 2025) Subject tagging is tackled using retrieval-augmented generation (RAG) (Lewis et al., 2020) to match TIBKAT records with GND subjects leveraging the OntoAligner toolkit (Giglou et al., 2025). The key methods steps are: 1) **Multi-Level Data Representation** – Records are represented at three levels: title-based, contextual (including metadata), and hierarchical (parent-level relationships) to improve subject mapping. 2) **Retrieval with Embeddings** – Nomic-AI embeddings (Nussbaum et al., 2024) are used to retrieve the top-k relevant subjects by computing cosine similarity between records and subject embeddings. 3) **LLM-Assisted Subject Selection** – Qwen2.5-0.5B-Instruct LLM (Yang et al., 2024) assesses retrieved subjects, verifying rele-

[9] https://sites.google.com/view/llms4subjects/

Team	Method	LLMs Used	Ranking
Annif	LLM-based synthetic data generation and XMTC traditional classifier models ensemble	Llama-3.1-8B-Instruct	1st all-subjects, 2nd tib-core, 4th qualitative
DNB-AI	Few-shot prompting to an LLM ensemble	Llama-3.2-3B-Instruct, Llama-3.1-70B-Instruct, Mistral-7B-v0.1, Mixtral-8x7B-Instruct-v0.1, OpenHermes-2.5-Mistral-7B, Teuken-7B-instruct-research-v0.4, LLama-3.1-8B-Instruct	4th all-subjects, N/A tib-core, 1st qualitative
DUTIR831	Synthetic data generation, GND knowledge distillation, and supervised finetuning	Qwen2.5-72B-Instruct	2nd all-subjects, 4th tib-core, 2nd qualitative
Homa	RAG-based ranking and retriever fine-tuning	Qwen2.5-0.5B	N/A all-subjects, 10th tib-core, 10th qualitative
Jim	BERT model ensemble	German BERT, Multilingual cased/uncased, ModernBERT base	7th all-subjects, N/A tib-core, 5th qualitative
LA²I²F	Transfer of concepts from similar documents to target and concept similarity to target document	Llama-3.1-8B as baseline, all-mpnet-base-v2 Sentence Transformer	6th all-subjects, 3rd tib-core, 8th qualitative
last_minute	Subjects are ranked, then re-ranked with embeddings, and refined with an LLM	stella-en-400M-v5, granite-embedding-125m-english, Llama-3.2-1B	N/A all-subjects, 9th tib-core, 11th qualitative
NBF	Finetuned embeddings using Burst Attention and multi-layer perceptron	all-mpnet-base-v2, german-roberta	9th all-subjects, N/A tib-core, 9th qualitative
RUC Team	Retrieves pre-indexed similar records and uses their subject tags as candidates	Arctic-Embed 2.0, Llama-8B, Chat GLM 4 (130B)	3rd all-subjects, 1st tib-core, 3rd qualitative
silp_nlp	Multilingual sentence transformer-based embedding similarity	jina-embeddings-v3-559M, distiluse-base-multilingual-cased-v2	11th all-subjects, 6th tib-core, N/A qualitative
TartuNLP	Bi-encoder candidate subject retrieval, and finetuned cross-encoder re-ranking model	multilingual-e5-large-instruct, mdeberta-v3-base	8th all-subjects, 7th tib-core, 7th qualitative
YNU-HPCC	Combines Sentence-BERT with contrastive learning	distilroberta, minilm, mpnet	10th all-subjects, 8th tib-core, 12th qualitative

Table 2: Overview of teams, methods, LLMs used, and rankings from quantitative scores over the two dataset collections and from the qualitative evaluations. The full leaderboard is released on our shared task website.

vance in a RAG-based ranking framework. And 4) **Fine-Tuning with Contrastive Learning** – The retriever is fine-tuned using contrastive learning, training on positive and negative record-subject pairs to improve distinction between relevant and irrelevant subjects.

5. Jim (Hahn, 2025) The key method steps of this system are: 1) **Multilingual BERT Ensemble** – The system uses an ensemble of four BERT models (two multilingual m1 & m2, one German-only, one English-only). 2) **Fine-Tuning on TIBKAT and GND Data** – The models are finetuned on TIBKAT records paired with GND subject labels, leveraging the AutoTrain framework (Thakur, 2024) for efficient optimization. 3) **Ensemble-Based Inference** – Subject predictions are ranked by summing confidence scores across models.

6. LA^{2}I^{2}F (Salfinger et al., 2025) The system retrieves subjects based on document similarity (analogical reasoning) and semantic similarity with ontology concepts (ontological reasoning), combining both for optimal subject tagging. The key steps are: 1) **Embedding-Based Retrieval** – MPNet sentence embeddings (Reimers and Gurevych, 2019) are used to represent documents and GND subjects in a shared vector space, enabling similarity-based matching. 2) **Analogical Reasoning for Subject Transfer** – The system identifies semantically similar training documents and transfers their human-assigned subject labels to the target document. 3) **Ontology-Based Subject Matching** – GND subjects are embedded and matched to documents based on semantic closeness, retrieving the most conceptually relevant subjects. And 4) **Final Fusion and Re-Ranking** – Predictions from both methods are merged and ranked by similarity, ensuring complementary information is integrated.

7. last_minute (Sarlak and Ansari, 2025) The system followed a rank, rerank, and refine approach using contextual vector embeddings stored in the

Milvus vector database for efficient retrieval. The key steps are: 1) **Finetuned Embedding Model** – The stella-en-400M-v5 model (Zhang et al., 2024) is finetuned on the training data with Multiple Negatives Ranking Loss (Henderson et al., 2017). 2) **Re-Ranking with a Cross-Encoder Model** – A granite-embedding-125m model (Granite Embedding Team, 2024) re-ranks the top 100 retrieved subject tags from the prior step, refining predictions before LLM processing. And 3) **LLM Refinement** – The Llama-3.2-1B LLM (2024) evaluates and selects the top 50 most relevant subject tags from the re-ranked list using prompt-based filtering.

8. NBF (Islam et al., 2025) The system introduces the use of Burst Attention (Sun et al., 2024), a lightweight self-attention mechanism that treats each embedding dimension as a token, capturing inter-dimensional dependencies to enhance subject retrieval. The methodology consists of four key steps. 1) **Sentence Transformer Embeddings** (2019): Articles and GND subjects are embedded using all-mpnet-base-v2 for en and german-roberta-sentence-transformer-v2 for de, aligning them in a shared space. 2) **Margin-Based Retrieval with BurstAttention**: The model is trained with a margin-based ranking loss, leveraging BurstAttention to refine embeddings by bringing relevant subjects closer and pushing irrelevant ones away. 3) **Feed-Forward MLP for Refinement**: A multilayer perceptron (MLP) further refines embeddings to improve subject retrieval accuracy. 4) **Top-k Search**: At inference, cosine similarity between article and subject embeddings is computed, selecting the top-k closest subjects as final predictions.

9. RUC Team (Tian et al., 2025a) This team used a retrieval-based method, prioritizing accuracy, speed, and scalability over heavy LLM inference. Key steps are: 1) **Cross-Lingual Embeddings for Retrieval** – Uses Arctic-Embed 2.0 (Yu et al., 2024), a multilingual embedding model, to match documents across English and German in a shared semantic space. 2) **Vector-Based Nearest Neighbor Search** – Computes inner product similarities between document embeddings using Faiss indexing to efficiently retrieve the most relevant records. And 3) **Ranking Relevant Subject Terms** – Merges and re-ranks candidate subjects based on document similarity, term position, and term occurrence in the title/abstract.

10. silp_nlp (Singh et al., 2025) This system utilizes **sentence transformer embeddings** (2019) for titles and abstracts, retrieving subjects based on cosine similarity. It employs JinaAi/jina-embeddings-v3 (Sturua et al., 2024), a **novel multilingual model supporting 89 languages**, to process both en and de text. Performance was compared against distiluse sentence transformers, with JinaAi embeddings achieving superior results.

11. TartuNLP (Dorkin and Sirts, 2025) The system first retrieves a coarse set of candidate subjects using a **bi-encoder**, viz. multilingual-e5-large-instruct (Wang et al., 2024), then refines the selection with a **cross-encoder re-ranking model**, viz. mdeberta-v3-base model (He et al.), finetuned on the task dataset. The key insight being the two-stage approach nearly doubles recall compared to using the bi-encoder alone, confirming the cross-encoder re-ranking's impact on performance.

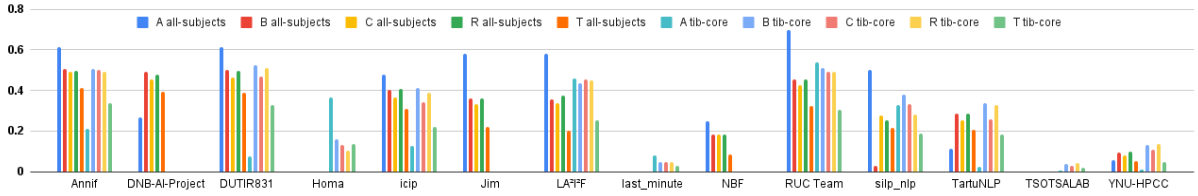
12. YNU-HPCC (Mao et al., 2025) The system fine-tuned multilingual sentence-BERT models, such as paraphrase-multilingual-MiniLM, -mpnet, and mpnet-base, using contrastive loss, to improve semantic alignment between the records and subjects. Key was their **Balanced Positive-Negative Sampling** – Two strategies were tested: multi-label sampling, aggregating all true labels per document, and single-label sampling, constructing 1:1 positive-negative pairs to improve classification.

The shared task attracted a diverse range of systems. A review of the 12 submissions revealed unique methodological contributions, including Burst Attention (Islam et al., 2025), analogical/ontological reasoning (Salfinger et al., 2025), the use of toolkits like Annif (2022) (Suominen et al., 2025) and OntoAligner (2025) (Bayrami Asl Tekanlou et al., 2025), and multi-stage prompt engineering. Some teams (Tian et al., 2025a; Singh et al., 2025) also evaluated newer embedding models such as Arctic-Embed and JinAI. As shown in Table 2, top-performing systems went beyond standard sentence transformer embeddings (2019). Key strategies among the leading teams—Annif (Suominen et al., 2025), DNB-AI (Kluge and Kähler, 2025), DUTIR831 (Tian et al., 2025b), RUC Team (Tian et al., 2025a), and Jim (Hahn, 2025)—included: (1) model ensembles, (2) synthetic training data generation, and (3) multilingual language models, the latter being common across submissions. Notably, DUTIR831 and RUC Team deployed very large LLMs—Qwen2.5-72B-Instruct and ChatGLM 4 (130B)—while others used models with 8B or fewer parameters.

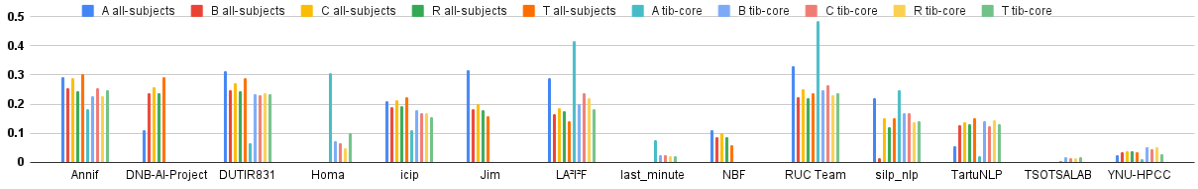
In the next section, we present the leaderboard results for the shared task.

all-subjects						tib-core					
Team Name	P@5	R@5	P@10	R@10	Ov. R@k	Team Name	P@5	R@5	P@10	R@10	Ov. R@k
Annif	0.26	0.49	0.16	0.57	0.63	RUC Team	0.25	0.48	0.16	0.57	0.66
DUTIR831	0.26	0.48	0.15	0.56	0.6	Annif	0.23	0.48	0.14	0.54	0.59
RUC Team	0.23	0.44	0.14	0.52	0.59	LA^{2}I^{2}F	0.2	0.41	0.13	0.49	0.58
DNB-AI	0.25	0.47	0.15	0.54	0.56	DUTIR831	0.23	0.49	0.13	0.54	0.56
icip	0.2	0.39	0.12	0.46	0.53	icip	0.17	0.37	0.1	0.44	0.5
LA^{2}I^{2}F	0.17	0.34	0.11	0.41	0.48	silp_nlp	0.16	0.34	0.11	0.42	0.49
Jim	0.18	0.34	0.11	0.41	0.47	TartuNLP	0.14	0.3	0.09	0.36	0.4
TartuNLP	0.13	0.27	0.08	0.33	0.38	YNU-HPCC	0.05	0.12	0.03	0.16	0.23
NBF	0.08	0.17	0.06	0.23	0.32	last_minute	0.02	0.05	0.02	0.08	0.21
YNU-HPCC	0.04	0.09	0.03	0.12	0.17	Homa	0.08	0.15	0.05	0.18	0.2
silp_nlp	0.05	0.08	0.03	0.11	0.13	TSOTSALAB	0.02	0.04	0.01	0.05	0.07
Homa	-	-	-	-	-	NBF	-	-	-	-	-
TSOTSALAB	-	-	-	-	-	DNB-AI	-	-	-	-	-
last_minute	-	-	-	-	-	Jim	-	-	-	-	-

Table 3: Quantitative performance comparisons for all-subjects and tib-core datasets. The 'Ov. R@k' column represents the average recall across $k = 5, 10, ..., 45, 50$.

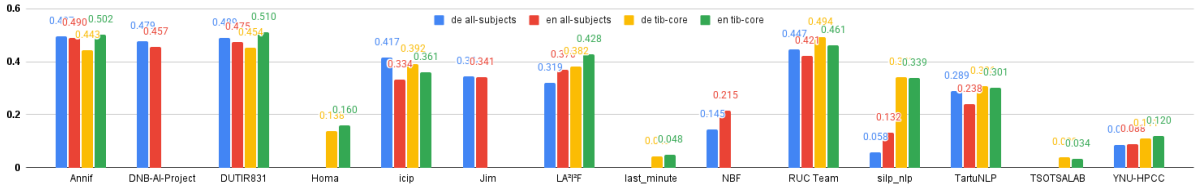


(a) Recall@5 results for all-subjects and tib-core datasets based on the record-type ablation.

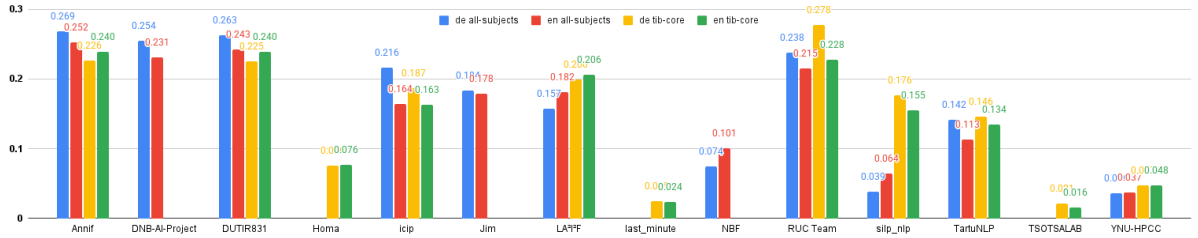


(b) Precision@5 results for all-subjects and tib-core datasets based on the record-type ablation.

Figure 1: K@5 Results on the record type ablation. A - article, B - book, C - conference, R - report, and T - thesis. On the x-axis, teams are listed in alphabetical order of names.



(a) Recall@5 results for all-subjects and tib-core datasets based on the language ablation.



(b) Precision@5 results for all-subjects and tib-core datasets based on the language ablation.

Figure 2: K@5 Results on the language ablation. On the x-axis, teams are listed in alphabetical order of names.

7 Shared Task Leaderboard Results

In this shared task, we provided two leaderboards: 1) quantitative results and 2) qualitative results.

Quantitative Metrics. System performance was evaluated using average precision@k, recall@k, and F1-score@k at multiple cutoffs (k = 5, 10, 15, ..., 50). These metrics were chosen as subject tagging was treated as a bag-of-words among applicable subjects, making precision, recall, and F1-score more suitable. Given the dataset structure of the LLMs4Subjects shared task, evaluation scores were released at varying levels of granularity: (1) language-level, separately for en and de, (2) record-level, across five types of technical records, and (3) combined language and record-levels, offering a comprehensive performance breakdown. This approach provided deeper insights into system performance and facilitated detailed discussions in the task overview and system description papers. To ensure transparency, the shared task evaluation script was publicly released.

Qualitative Metrics. To assess system-generated results in real-world scenarios, a qualitative evaluation was conducted over three weeks. TIB subject specialists manually reviewed 122 test records common to both all-subjects and tib-core, sampling 10 records from each of 14 subject classifications. The top 20 GND codes from teams' submissions were extracted, and subject librarians labeled them as Y (correct), I (irrelevant but technically correct), or N/Blank (incorrect). Two evaluation criteria were used: case 1 - treating both Y and I as correct, and case 2 - considering only Y. Results were summarized using average P@20, R@20, and F1@20.

Detailed results leaderboards are released on the shared task website.

7.1 Quantitative Evaluations

The primary evaluation metric was recall, with the overall leaderboard ranking based on average recall scores across k values from 5 to 50. For practical use by subject specialists, systems should predict relevant subjects at lower k values, ideally between 5 and 10, with a maximum of 20. Table 3 presents the results for both collections: all-subjects and tib-core. The top teams consistently predicted over half of the subject annotations in both collections. A caveat here is that our precision score would never amount to one and heavily penalizes actual system performance.[10] Despite this, they were included to provide a comparison. The top three teams based on average recall for all-subjects were Annif, DUTIR831, and RUC Team, while for tib-core, the top three were RUC Team, Annif, and LA^{2}I^{2}F. Notably, the top-performing teams on all-subjects maintained strong rankings on tib-core, with DUTIR831 placing fourth. LA^{2}I^{2}F, ranking third on tib-core, appeared more effective on the smaller subjects taxonomy of tib-core.

At the record-type level (Figure 1a), most systems achieved high recall@5 for articles in both collections, while books, conference papers, and reports showed similar performance. The weakest results were observed for theses. For the top teams on all-subjects (Annif, DUTIR831, and RUC Team), precision scores across record types were similar (Figure 1b). RUC Team demonstrated consistently high precision on articles in both collections. On tib-core, LA^{2}I^{2}F's boost to third place stemmed from its high precision on articles—second only to RUC—despite comparable recall scores to other teams. At the language level, results from both recall (Figure 2a) and precision (Figure 2b) showed no significant difference between processing de or en records across all teams. The only consistent variation was that RUC Team performed slightly better on de records, while LA^{2}I^{2}F for en records.

7.2 Qualitative Evaluations

We now turn to the qualitative manual evaluations from the shared task's evaluation phase.

Table 4 shows results for both qualitative evaluation cases. Based on subject librarian assessments, both Y (correct) and I (irrelevant but technically correct) labels were counted as correct in case 1, while only Y was considered correct in case 2. The top 4 teams ranked consistently across both cases, with minor changes among the remaining teams. Case 1 accounts for situations where models predicted multiple semantically similar subjects as top-ranked, leading to generally higher precision scores. However, in practice, it is preferred that models predict semantically distinct and relevant subjects as top-ranked. Therefore, the remainder of this section focuses on case 2, where only Y labels are treated as correct.

[10] Precision@k is the number of correct predictions among the top-k, divided by k. Since TIBKAT records average around 5 true GND subjects, precision is inherently limited at higher k. Even a perfect system would achieve at most 0.5 precision at $k = 10$, as only 5 of 10 predictions can be correct.

qualitative eval. case 1						qualitative eval. case 2					
Team Name	P@5	R@5	P@10	R@10	Ov. R@k	Team Name	P@5	R@5	P@10	R@10	Ov. R@k
DNB-AI	0.74	0.33	0.65	0.54	0.57	DNB-AI	0.53	0.34	0.41	0.5	0.51
DUTIR831	0.7	0.31	0.61	0.49	0.53	DUTIR831	0.49	0.32	0.39	0.46	0.49
RUC Team	0.71	0.28	0.6	0.46	0.52	RUC Team	0.48	0.29	0.38	0.43	0.47
Annif	0.66	0.28	0.56	0.46	0.5	Annif	0.46	0.3	0.33	0.42	0.45
TartuNLP	0.63	0.26	0.55	0.44	0.49	Jim	0.4	0.29	0.29	0.39	0.43
Jim	0.62	0.29	0.5	0.44	0.49	icip	0.39	0.28	0.3	0.4	0.42
icip	0.57	0.27	0.48	0.43	0.48	TartuNLP	0.4	0.26	0.31	0.38	0.41
LA^2I^2F	0.52	0.25	0.43	0.39	0.46	LA^2I^2F	0.34	0.25	0.25	0.35	0.4
NBF	0.44	0.21	0.41	0.37	0.43	NBF	0.23	0.17	0.2	0.28	0.32
last_minute	0.27	0.15	0.24	0.25	0.29	Homa	0.19	0.15	0.15	0.21	0.22
Homa	0.3	0.16	0.25	0.26	0.27	last_minute	0.13	0.1	0.11	0.18	0.2
YNU-HPCC	0.21	0.14	0.18	0.22	0.26	YNU-HPCC	0.12	0.1	0.1	0.16	0.17

Table 4: Qualitative performance comparison across teams for two cases: case 1 — treating both Y and I as correct, and case 2 — treating only Y as correct. Metrics reported are precision and recall at k. The 'Ov. R@k' columns represent the average recall across $k = 5, 10, 15, 20$.

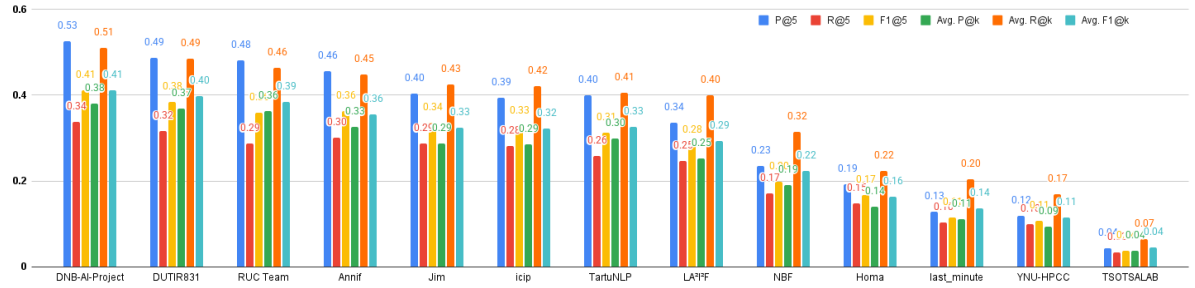
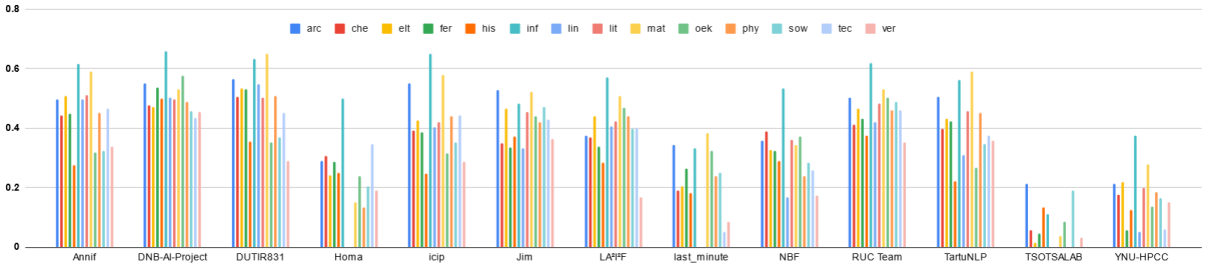
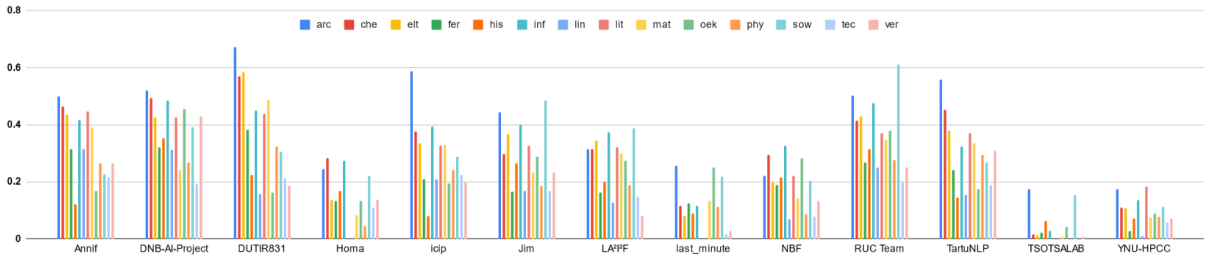


Figure 3: Overall qualitative evaluation results w.r.t. metric@5 and averages per metric@k where k = 5, 10, 15, and 20. On the x-axis, teams are listed in ranked order of performance based on average recall@k.



(a) Average recall@k scores per domain over k = 5, 10, 15, and 20.



(b) Average precision@k scores per domain over k = 5, 10, 15, and 20.

Figure 4: Qualitative results per 14 distinct domains. Acronyms used: Architecture (arc), Chemistry (che), Electrical Engineering (elt), Material Science (fer), History (his), Computer Science (inf), Linguistics (lin), Literature Studies (lit), Mathematics (mat), Economics (oek), Physics (phy), Social Sciences (sow), Engineering (tec), and Traffic Engineering (ver). On the x-axis, teams are listed in alphabetical order of names.

The overall results, shown in Figure 3, report six metrics: P@5, R@5, F1@5, Avg. P@k, Avg. R@k, and Avg. F1@k, with k ranging from 5 to 20 in the qualitative setting. These results do not distinguish between all-subjects and tib-core since the 122 evaluated records were shared across both collections.[11] The top teams from the quantitative leaderboard (DUTIR831, RUC Team, and Annif) remained among the top four, with the DNB-AI-Project emerging as the best-performing system in qualitative evaluations. Here, precision reflected true system performance, measuring the proportion of predicted subjects marked correct by subject librarians. Recall was adjusted to account for any newly identified correct subjects not present in the gold standard. DNB-AI-Project stood out for its high precision among the top 20 recalled subjects, employing a purely LLM-based approach with an ensemble of LLMs and few-shot prompting, requiring no fine-tuning. This supports the premise of the shared task—assessing whether LLMs can generalize effectively compared to traditional machine learning approaches that rely on extensive fine-tuning. Among the 14 evaluated domains, Computer Science (inf) consistently had the highest average recall (Figure 4a), while Linguistics (lin) and Literature Studies (lit) showed no predictions from Homa and last_minute. Most teams struggled with Engineering (tec) and Traffic Engineering (ver) records, and Annif and DUTIR831 also exhibited low recall for History (his) and Economics (oek). In contrast, the DNB-AI-Project and RUC Team demonstrated consistent performance across all domains. Finally, as shown in Figure 4b, precision did not always align with recall rankings; the Architecture (arc) domain, however, exhibited the top 2 highest precision among all domains.

8 Discussion

To establish a reference point for the LLMs4Subjects shared task, we developed a baseline system using OpenAI's GPT-4o via the Assistant API.[12] Two assistants were configured—one for all-subjects and another for tib-core—each equipped with the respective GND subject taxonomies stored in OpenAI's vector stores. For each TIBKAT record, the assistants embedded the title and abstract and queried the vector store to retrieve 50 GND subjects.

Both assistants followed an identical prompt that defined their role as a subject matter expert in a technical library using the GND taxonomy. The prompt instructed them to select exactly 50 valid, semantically relevant subject tags based on the input title and abstract, and return them in a strict JSON format. It also enforced constraints such as avoiding duplicates and non-matching entries, and supported bilingual input in German and English.

Despite using a GPT LLM and structured prompt, the Assistant API showed reliability issues: some outputs broke schema constraints or included malformed GND codes, requiring multiple runs. Still in beta, its stability at scale is uncertain. This baseline ranked below all participant submissions on the leaderboard, underscoring that while the API enables quick prototyping, effective subject tagging demands more specialized, robust systems like those of the top teams.

9 Conclusion

SemEval-2025 Task 5: LLMs4Subjects presented the first evaluation of LLMs for GND-based subject indexing in bilingual (German/English) technical library records, combining quantitative metrics with expert assessments. The task revealed four key takeaways: multilingual models and training data outperformed monolingual ones (RQ1, RQ2); synthetic data and retrieval-augmented pipelines improved performance, underscoring the value of data diversity and system design (RQ3); and smaller, well-engineered systems often rivaled large instruction-tuned LLMs, highlighting trade-offs between scale and specialization (RQ4).

All data, code, and evaluation resources are openly available at `https://github.com/jd-coderepos/llms4subjects`. A second edition[13] of the task is planned, with an emphasis on *energy- and compute-efficient* LLM systems.

Acknowledgments

The SemEval Task-5 LLMs4Subjects shared task was jointly supported by the TIB Leibniz Information Centre for Science and Technology, the SCINEXT project (BMBF, Grant 01IS22070), and the NFDI4DataScience initiative (DFG, Grant 460234259).

[11] The silp_nlp team output was not evaluated qualitatively since it was submitted after the deadline.

[12] `https://platform.openai.com/docs/api-reference/assistants`

[13] `https://sites.google.com/view/llms4subjects-germeval/`

References

Mohammad Munzir Ahanger and M Arif Wani. 2022. Novel deep learning approach for scientific literature classification. In *2022 9th International Conference on Computing for Sustainable Global Development (INDIACom)*, pages 249–254. IEEE.

Hadi Bayrami Asl Tekanlou, Jafar Razmara, Mahsa Sanaei, Mostafa Rahgouy, and Hamed Babaei Giglou. 2025. Homa at semeval-2025 task 5: Aligning librarian records with ontoaligner for subject tagging. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 2371–2377, Vienna, Austria. Association for Computational Linguistics.

Levent Bolelli, Şeyda Ertekin, and C Lee Giles. 2009. Topic and trend detection in text collections using latent dirichlet allocation. In *European conference on information retrieval*, pages 776–780. Springer.

Davide Buscaldi, Simon David Hernandez, and Thierry Charnois. 2017. Classification of keyphrases from scientific publications using wordnet and word embeddings. In *VADOR (Valorisation et Analyse des Données de la Recherche) 2017*.

Jianlyu Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. M3-embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 2318–2335.

Aleksei Dorkin and Kairit Sirts. 2025. Tartunlp at semeval-2025 task 5: Subject tagging as two-stage information retrieval. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 2414–2419, Vienna, Austria. Association for Computational Linguistics.

Hamed Babaei Giglou, Jennifer D'Souza, Oliver Karras, and Sören Auer. 2025. Ontoaligner: A comprehensive modular and robust python toolkit for ontology alignment.

Monika Glowacka-Musial. 2022. Applying topic modeling for automated creation of descriptive metadata for digital collections. *Information Technology and Libraries*, 41(2).

Paul Gooding, Melissa Terras, Linda Berube, Mike Bennett, and Richard Hadden. 2019. Subjectifying library users to the macroscope using automatic classification matching. In *Digital Humanities 2019: Annual conference of ADHO (Assoc of Digital Humanities Organisations)*.

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, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Thomas L Griffiths and Mark Steyvers. 2004. Finding scientific topics. *Proceedings of the National academy of Sciences*, 101(suppl_1):5228–5235.

Jim Hahn. 2025. Jim at semeval-2025 task 5: Multilingual bert ensemble. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 2378–2383, Vienna, Austria. Association for Computational Linguistics.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing. In *The Eleventh International Conference on Learning Representations*.

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

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Baharul Islam, Nasim Ahmad, Ferdous Barbhuiya, and Kuntal Dey. 2025. Nbf at semeval-2025 task 5: Light-burst attention enhanced system for multilingual subject recommendation. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 956–961, Vienna, Austria. Association for Computational Linguistics.

Sujay Khandagale, Han Xiao, and Rohit Babbar. 2020. Bonsai: diverse and shallow trees for extreme multi-label classification. *Machine Learning*, 109(11):2099–2119.

Lisa Kluge and Maximilian Kähler. 2024. Few-shot prompting for subject indexing of German medical book titles. In *Proceedings of the 20th Conference on Natural Language Processing (KONVENS 2024)*, pages 141–148, Vienna, Austria. Association for Computational Linguistics.

Lisa Kluge and Maximilian Kähler. 2025. Dnb-ai-project at semeval-2025 task 5: An llm-ensemble approach for automated subject indexing. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 1114–1124, Vienna, Austria. Association for Computational Linguistics.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.

Abdulhussain E Mahdi and Arash Joorabchi. 2011. Automatic subject classification of scientific literature using citation metadata. In *International Conference on Digital Enterprise and Information Systems*, pages 545–559. Springer.

Yuheng Mao, Jin Wang, and Xuejie Zhang. 2025. Ynuhpcc at semeval-2025 task 7: Multilingual and cross-lingual fact-checked claim retrieval. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 238–244, Vienna, Austria. Association for Computational Linguistics.

Olena Medelyan. 2009. *Human-competitive automatic topic indexing*. Ph.D. thesis, The University of Waikato.

Zach Nussbaum, John X Morris, Brandon Duderstadt, and Andriy Mulyar. 2024. Nomic embed: Training a reproducible long context text embedder. *arXiv preprint arXiv:2402.01613*.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36:53728–53741.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992.

Michal Růžička and Petr Sojka. 2021. Towards math-aware automated classification and similarity search of scientific publications: Methods of mathematical content representations. *arXiv preprint arXiv:2110.04040*.

Angelo A Salatino, Francesco Osborne, Thiviyan Thanapalasingam, and Enrico Motta. 2019. The cso classifier: Ontology-driven detection of research topics in scholarly articles. In *Digital Libraries for Open Knowledge: 23rd International Conference on Theory and Practice of Digital Libraries, TPDL 2019, Oslo, Norway, September 9-12, 2019, Proceedings 23*, pages 296–311. Springer.

Andrea Salfinger, Luca Zaccagna, Francesca Incitti, Gianluca G. M. De Nardi, Lorenzo Dal Fabbro, and Lauro Snidaro. 2025. La²i²f at semeval-2025 task 5: Reasoning in embedding space – fusing analogical and ontology-based reasoning for document subject tagging. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 2384–2394, Vienna, Austria. Association for Computational Linguistics.

Zahra Sarlak and Ebrahim Ansari. 2025. Last minute at semeval-2025 task 5: Rag system for subject tagging. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 2403–2407, Vienna, Austria. Association for Computational Linguistics.

Sumit Singh, Pankaj Kumar Goyal, and Uma Shanker Tiwary. 2025. silp_nlp at semeval-2025 task 5: Subject recommendation with sentence transformer. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 2420–2425, Vienna, Austria. Association for Computational Linguistics.

Henry Small, Kevin W Boyack, and Richard Klavans. 2014. Identifying emerging topics in science and technology. *Research policy*, 43(8):1450–1467.

Saba Sturua, Isabelle Mohr, Mohammad Kalim Akram, Michael Günther, Bo Wang, Markus Krimmel, Feng Wang, Georgios Mastrapas, Andreas Koukounas, Nan Wang, et al. 2024. jina-embeddings-v3: Multilingual embeddings with task lora. *arXiv preprint arXiv:2409.10173*.

Ao Sun, Weilin Zhao, Xu Han, Cheng Yang, Zhiyuan Liu, Chuan Shi, and Maosong Sun. 2024. Burstattention: An efficient distributed attention framework for extremely long sequences. *arXiv preprint arXiv:2403.09347*.

Osma Suominen. 2019. Annif: Diy automated subject indexing using multiple algorithms. *LIBER Quarterly: The Journal of the Association of European Research Libraries*, 29(1):1–25.

Osma Suominen, Juho Inkinen, and Mona Lehtinen. 2022. Annif and finto ai: developing and implementing automated subject indexing. *Bibliographic control in the digital ecosystem.-(Biblioteche & bibliotecari, 2612-7709; 5)*, pages 265–282.

Osma Suominen, Juho Inkinen, and Mona Lehtinen. 2025. Annif at semeval-2025 task 5: Traditional xmtc augmented by llms. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 2395–2402, Vienna, Austria. Association for Computational Linguistics.

Abhishek Thakur. 2024. Autotrain: No-code training for state-of-the-art models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 419–423.

Xia Tian, Yang Bruce Xin, Wu Wen Jing, Xiu Yue Heng, Xin Zhang, Li Jin Yu, Tong Gao, Tan Zhuo Xi, Hu Run Dong, Chen Tao, and Jia Jun Zhi. 2025a. RUC Team at SemEval-2025 Task 5: Fast Automated Subject Indexing: A Method Based on Similar Records Matching and Related Subject Ranking. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, Vienna, Austria. Association for Computational Linguistics.

Yicen Tian, Erchen Yu, Yanan Wang, Dailin Li, jiaqi yao, Hongfei LIN, Linlin Zong, and Bo Xu. 2025b. Dutir831 at semeval-2025 task 5: A multi-stage llm approach to gnd subject assignment for tibkat records. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 367–376, Vienna, Austria. Association for Computational Linguistics.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. Multilingual e5 text embeddings: A technical report. *arXiv preprint arXiv:2402.05672*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

Hsiang-Fu Yu, Kai Zhong, Jiong Zhang, Wei-Cheng Chang, and Inderjit S Dhillon. 2022. Pecos: Prediction for enormous and correlated output spaces. *Journal of Machine Learning Research*, 23(98):1–32.

Puxuan Yu, Luke Merrick, Gaurav Nuti, and Daniel Campos. 2024. Arctic-embed 2.0: Multilingual retrieval without compromise. *arXiv preprint arXiv:2412.04506*.

Dun Zhang, Jiacheng Li, Ziyang Zeng, and Fulong Wang. 2024. Jasper and stella: distillation of sota embedding models. *arXiv preprint arXiv:2412.19048*.

A TIBKAT media types

Book: A comprehensive written work published as a volume or a series of volumes.
Thesis: A document submitted in support of candidature for an academic degree or professional qualification, presenting the author's research and findings.
Conference: Proceedings or collections of papers presented at academic conferences or symposiums.
Report: Detailed and systematic accounts of research findings, often prepared for a specific audience or purpose.
Article: A written composition on a specific topic, usually intended for publication in a journal or magazine. Collection: A curated assembly of documents or works, typically related by theme, subject, or author.
AudioVisualDocument: Media content that combines both sound and visual components, such as videos or films.
Periodical: Publications issued at regular intervals, such as journals, magazines, or newspapers.
Chapter: A specific section or segment of a book, usually focusing on a particular topic within the larger work.

B TIBKAT Language Distribution

There are 48 different languages in the TIBKAT with abstracts. Their distributions are listed below.

de (German): 108637; en (English): 76735; fr (French): 1741; id (Indonesian): 945; es (Spanish): 311; it (Italian): 294; nl (Dutch): 167; da (Danish): 129; sq (Albanian): 100; ro (Romanian): 93; ca (Catalan): 86; fi (Finnish): 80; so (Somali): 67; sv (Swedish): 65; no (Norwegian): 50; unknown (Unknown): 41; tl (Tagalog): 31; et (Estonian): 24; pt (Portuguese): 16; sw (Swahili (macrolanguage)): 15; pl (Polish): 15; hr (Croatian): 12; lt (Lithuanian): 10; hu (Hungarian): 10; af (Afrikaans): 10; tr (Turkish): 8; sk (Slovak): 7; sl (Slovenian): 4; cy (Welsh): 4; cs (Czech): 3; lv (Latvian): 3; vi (Vietnamese): 2; he (Hebrew): 2; ko (Korean): 1; ru (Russian): 1.

C Additional details about the GND

At TIB, the GND is used as follows. The subject specialists are usually using online services like GND-Explorer (https://explore.gnd.network/) or OGND (https://swb.bsz-bw.de/) for searching the GND. There you can also restrict to Sachbegriff/Topical term (saz), which is the term class we are using for subject indexing (there are some additional term classes used, like geographical terms (swg), but to start topical terms (saz) should be sufficient). New terms are added in a cooperative process and are searchable as soon as they passed a review process. For more general details you can also have a closer look here. There is also a way to get complete sets of the GND that are updated on a regular basis. Details can be found here. A centralized resource pool and a guide for accessing the GND are provided in the LLMs4Subjects GitHub repository https://github.com/jd-coderepos/llms4subjects/tree/main/gnd-how-to.

Note, all terms in GND usually are in German and every TIB record regardless of its language is described by it. But there are some cases, where a term is e.g. English, if the preferred naming is English. In this case a German naming can be listed under synonyms. The synonyms are also important for the subject classification, as many relevant terms are listed under synonyms. These are not always synonyms in a strong sense, as the GND is a growing database and meanings of terms once created do change or shift and larger corrections are rarely realized, if terms are not wrong in sense of content. The GND's purpose extends to enhancing the discoverability of literature in bibliographic systems, where synonyms are also utilized, emphasizing their importance in indexing.

D Pilot Task

At TIB, ANNIF has been used in production code since the start of 2024 for the use case of assigning items from the TIB portal discovery system to one or several subject facets. The TIB portal employs a multi-stage algorithm to attribute a record to one of the 28 TIB's different subjects, viz. Architecture, Civil Engineering, Biochemistry, Biology, Chemistry, Chemical Engineering, Electrical Engineering, Energy Technology, Educational Science, Earth Sciences, History, Information Technology, Literary Studies and Linguistics, Mechanical Engineering, Mathematics, Medical Technology, Plant Sciences, Philosophy, Physics, Law, Study of Religions, Social Sciences, Sports Sciences, Theology, Environmental Engineering, Traffic Engineering, Materials Science, and Economics, the last of which is the so-called automatic stage. If no more salient information is available, machine learning methods are used to assign the subject(s). Note, the subjects reference here can be seen directly akin to fields of study or scientific disciplines, whereas LLMs4Subjects includes a much broader scope for its subjects. Previously utilizing a commercial algorithm, TIB switched to ANNIF for its customization potential and community-driven improvements. The training data of ANNIF algorithms consists of document metadata from the TIB catalog, partially overlapping with the training dataset for LLMs4Subjects. Since the documents to be indexed by ANNIF include many cases where abstracts or fulltexts cannot be accessed programmatically, we only consider the the titles and publishers. ANNIF has shown good overall results in assigning the 14 subjects is has so far been incrementally trained on, with an overall F1 score of ≈ 0.65 for several algorithms. Both English and German-language documents were considered, with little difference in performance when training both languages combined or separately. Leveraging the capabilities of LLMs as a complementary approach to ANNIF marks a logical next step in the automation of subject indexing.