DUTIR831 at SemEval-2025 Task 5: A Multi-Stage LLM Approach to GND Subject Assignment for TIBKAT Records

Yicen Tian[1], **Erchen Yu**[1], **Yanan Wang**[1], **Dailin Li**[1],
Jiaqi Yao[1], **Hongfei Lin**[1], **Linlin Zong**[2], **Bo Xu** [1*],
[1]School of Computer Science and Technology, Dalian University of Technology, China
[2]School of Software, Dalian University of Technology, China
{yicentian, yuerchen0809, wangyanan, ldlbest, 1741917591}@mail.dlut.edu.cn
{hflin, llzong, xubo}@dlut.edu.cn

Abstract

This paper introduces DUTIR831's approach to SemEval-2025 Task 5, which focuses on generating relevant subjects from the Integrated Authority File (GND) for tagging multilingual technical records in the TIBKAT database. To address challenges in understanding the hierarchical GND taxonomy and automating subject assignment, a three-stage approach is proposed: (1) a data synthesis stage that utilizes LLM to generate and selectively filter high-quality data, (2) a model training module that leverages LLMs and various training strategies to acquire GND knowledge and refine TIBKAT preferences, and (3) a subject terms completion mechanism consisting of multi-sampling ranking, subject terms extraction using a LLM, vector-based model retrieval, and various re-ranking strategies. The quantitative evaluation results show that our system is ranked **2nd** in the all-subject datasets and **4th** in the tib-core-subjects datasets. And the qualitative evaluation results show that the system is ranked **2nd** in the tib-core-subjects datasets.

1 Introduction

In the era of information explosion, the efficient organization and retrieval of knowledge have become paramount. Libraries, as custodians of vast repositories of information, play a critical role in this endeavor. The process of cataloging and tagging library records with relevant subject headings is not merely a clerical task but a foundational activity that enhances the discoverability and accessibility of resources. The advent of Large Language Models (LLMs) has opened new avenues for automating and refining this process, thereby addressing the challenges posed by the sheer volume and complexity of modern bibliographic data (Devlin et al., 2019).

The LLMs4Subjects task focuses on developing LLM-based systems that generate the most relevant subjects from the Integrated Authority File (GND) subject collection to tag a given TIBKAT technical record in either German or English (D'Souza et al., 2025). This task involves five types of records: articles, books, conference papers, reports, and thesis. To achieve this, two key challenges are addressed:

1. Understand the taxonomy of the GND subject. The GND is an international authority file that is primarily used by German libraries to catalog and link subjects, organizations, and works. The objective is to enable LLMs to learn and apply this taxonomy effectively, capturing its hierarchical structure and semantic nuances.

2. Automatically assign subjects to TIBKAT records. The system should be able to recommend relevant GND subjects by analyzing the semantic relationships between subjects and the titles and abstracts of technical records. To train and evaluate our system, two datasets are utilized: tib-core-subjects and all-subjects.

Our approach utilizes a three-stage architecture:

(1) A data synthesis stage that utilizes LLM and prompt design to generate and selectively filter high-quality data.

(2) A model training module that leverages LLMs, knowledge distillation, supervised fine-tuning, and preference alignment to acquire GND knowledge and align with TIBKAT preferences.

(3) A subject term completion mechanism that includes frequency-position ranking based on multi-sampling for term generation, subject term extraction using LLMs, vector-based retrieval for title-driven term generation, and various re-ranking strategies.

The quantitative evaluation results show that our system is ranked **2nd** in the all-subject datasets and **4th** in the tib-core-subjects datasets. And the qualitative evaluation results show that the system

*Corresponding author.

Proceedings of the The 19th International Workshop on Semantic Evaluation (SemEval-2025), pages 363–372

is ranked **2nd** in the tib-core-subjects datasets.

2 Related Work

Recent advancements in LLMs revolutionized text classification. Models such as Qwen (Yang et al., 2024) and GPT (Brown et al., 2020) demonstrated exceptional performance in capturing semantic nuances. Fine-tuning pre-trained models on domain-specific data, such as technical records, yielded promising results (Brzustowicz, 2023). Regarding training strategies, the Low-Rank Adaptation (LoRA) method reduced the number of fine-tuning parameters and computational costs while preserving performance (Hu et al., 2021). Additionally, optimization methods such as Direct Preference Optimization (DPO) were introduced to align model outputs with target preferences (Rafailov et al., 2023).

Retrieval models, such as DPR (Karpukhin et al., 2020) and BGE-M3 (Chen et al., 2023), were widely utilized for semantic search. Reranking strategies, including the use of the Deepseek (DeepSeek-AI et al., 2024), Qwen, and ChatGLM (GLM et al., 2024) APIs, further enhanced the accuracy of retrieval systems. These techniques played a crucial role in improving subject classification performance in our task.

3 System Overview

Our system consists three parts: data synthesis, training strategies, and integration of subject terms. The framework of the system is shown in Figure 1.

3.1 Data Synthesis

3.1.1 Incremental Data Generation

Wei and Zou (2019) demonstrates that incremental data can enhance model learning quality. For the two datasets provided by the organizers, tib-core-subjects and all-subjects, we first extract the GND code list from the records in the labeled training set. Using the GND-subjects-all and GND-subjects-tib-core collections provided by the organizers, we then map the extracted GND codes to their corresponding subject terms. To further expand the subject term set, we iteratively search for related terms within the same classification, randomly selecting up to three unused terms in each round until the set is sufficiently enriched.

We employ Qwen2.5-72B-Instruct for incremental data generation. Given a set of subject terms, the model selects relevant terms and generates corresponding titles and abstracts for various record types, including books, articles, theses, reports, and conference papers. The selected terms are then ranked based on their relevance to the generated titles and abstracts. The final output consists of the generated title, abstract, and the selected subject terms. The specific prompt template used for incremental data generation is detailed in Appendix B.1.

3.1.2 Data filtering for Incremental Data

Although Qwen2.5-72B-Instruct demonstrates exceptional performance in text comprehension and generation, some generated texts may exhibit suboptimal quality. To mitigate this issue, we implement a filtering procedure using Qwen2.5-72B-Instruct to evaluate and refine the generated records. Specifically, the model evaluates the logical coherence of the generated titles and abstracts, as well as their relevance to the selected subject terms.

The detailed prompt template used for data filtering is provided in Appendix B.2. Due to time constraints, this filtering process is applied only to the incremental data within the all-subjects dataset. Ultimately, 10% of the incremental data is excluded, retaining only those records that achieve full scores for both logical coherence and subject term relevance.

3.2 Training Strategies

3.2.1 GND knowledge Distillation

The organizers provide two GND subject collections: GND-subjects-all and GND-subjects-tib-core. These datasets contain essential information, including GND Code, Classification Number, Classification Name, Name, Alternative Name, Related Subject, and Definition. Fine-tuning an LLM enables it to internalize domain-specific knowledge, thereby improving its ability to interpret subject terms and their complex semantics.

To enhance the models comprehension and effective utilization of GND knowledge, we fine-tune Qwen2.5-7.5B-Instruct by incorporating GND subject information. The input consists of subject term names, while the output includes their corresponding properties formatted in JSON. The specific prompt template used for GND knowledge distillation is detailed in Appendix B.3.

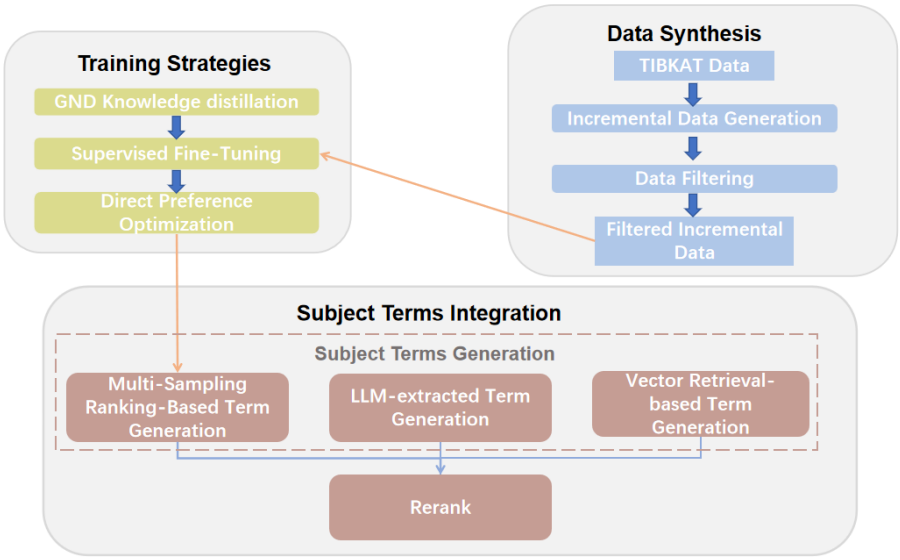

Figure 1: System Overview.

3.2.2 TIBKAT: Supervised Fine-Tuning and Direct Preference Optimization

We utilize LoRA for fine-tuning, where the input consists of titles and abstracts, and the output is a list of predicted GND subject terms. The specific prompt template used for Supervised Fine-Tuning (SFT) is provided in Appendix B.4.

Building upon the model trained in Section 3.2.1, we develop three versions of the SFT model for the two datasets. $M_{sft(raw)}$, trained on the original dataset, serves as the baseline. $M_{sft(all)}$, trained on the original dataset combined with the filtered incremental dataset, serves as the base model for subsequent DPO training. Meanwhile, M_{reject} is trained on half of the original data to generate negative examples for the DPO process. We utilize M_{reject} to construct negative examples for DPO because an SFT model trained on only half of the original dataset has acquired limited knowledge. As a result, its predictions may deviate from the true labels, making them suitable for use as negative examples in the preference optimization process.

3.3 Integration of Subject Terms

The integration of subject terms comprises two key components: Subject Terms Generation and Reranking.

3.3.1 Subject Terms Generation

Subject terms generation consists of three approaches, with the terms generated by these methods arranged sequentially to fulfill the organizer's requirement of producing 50 subject terms for each technical record.

Step 1: Multi-Sampling Ranking-Based Term Generation

Temperature influences the diversity and consistency of the model's generated outputs, while the random seed determines the starting point of the sampling process. Different seeds result in distinct text sequences. By performing multiple samplings with different seeds, we can explore a wider range of possibilities within the model, thereby enhancing the diversity of the results and reducing the likelihood of the model getting trapped in local optima (Holtzman et al., 2020). As a result, we propose a frequency-position sorting method based on multiple samplings.

The frequency-position sorting method first ranks the subject terms based on frequency. For terms with identical frequencies, it then sorts them in ascending order according to their average position. Frequency-based sorting ensures that core subject terms are prioritized. When the frequencies of multiple subject terms are similar, position-based sorting provides an additional criterion for differentiation, thereby improving the rationality of the sorting results.

Step 2: LLM-extracted Term Generation

Wang et al. (2023) has revealed that LLMs like GPT-3, which excel in generation tasks, can also be effectively applied to keyword extraction tasks by employing appropriate prompting techniques.

We verify that the Qwen-Plus model incorporates knowledge of the GND classification system.

By designing a one-shot prompt, the LLM analyzes the title and abstract to extract relevant GND subject terms. Subsequently, the BGE-M3 vector retrieval model calculates the similarity between the extracted terms and the official subject term repository. Terms that surpass a predefined similarity threshold are selected from the repository. The detailed prompt design is provided in Appendix B.5.

Step 3: Vector Retrieval-based Term Generation

The competition organizer stipulates that each record should be accompanied by 50 subject terms. However, the subject terms generated through multi-sampling ranking and LLM extraction are insufficient to reach the required 50 terms. Therefore, we employ the BGE-M3 vector retrieval model to retrieval additional subject terms. We propose three retrieving contents, as follows: (1) Title. (2) Abstract. (3) Title and Abstract.

3.3.2 Rerank

The subject terms extracted using the Qwen-Plus API, along with those supplemented by BGE-M3, include the correct answers. In the context of quantitative evaluation metrics, which focus on the average recall@k (where k ranges from 5 to 50) across all records, it is crucial to prioritize the more relevant subject terms at the forefront of the predicted subject term list. To achieve this, three re-ranking strategies are devised.

Assume that the current subject term list is denoted as $S = [s_1, s_2, s_3]$, with s_1 representing the subject term list predicted by model M_{dpo}, s_2 being the subject term list extracted by Qwen-Plus, and s_3 being the subject term list retrieved by BGE-M3. The following are the three strategies:

Strategy 1: comprehensively re-rank all of s_1, s_2, and s_3.

Strategy 2: let s_1 remain in its original position and re-rank s_2 and s_3.

Strategy 3: keep s_1 and s_2 in their original positions and re-rank s_3.

Qwen-Plus and a one-shot prompting approach are utilized for re-ranking. The design of the prompt is presented in Appendix B.6.

4 Experimental Setup

4.1 Data description and Evaluation

The datasets are provided by Semeval-2025 Task 5. TIBKAT has two main datasets: all-subjects and tib-core-subjects. Only the labeled training dataset was used for training. Furthermore, the data created in 3.1 was also used in the training phase. The distribution of the datasets is shown in Appendix Table 2.

The quantitative evaluation focuses on precision, recall, and F1 scores at various thresholds (k = 5 to 50) for two datasets. The official quantitative evaluation is conducted based on the average recall scores across the specified thresholds, emphasizing the importance of retrieving relevant subjects.

4.2 Implementation

In Sections 3.2.1 and 3.2.2, LoRA fine-tuning was applied. In Section 3.2.1, Qwen2.5-7B-Instruct served as the base model, resulting in M_{gnd}, which was further fine-tuned in the SFT stage in Section 3.2.2 to obtain $M_{sft(raw)}$ and $M_{sft(all)}$. Subsequently, $M_{sft(all)}$ was refined in the DPO stage to produce M_{dpo}. Hyper-parameter settings are detailed in Appendix Table 3. All experiments were conducted on 8 RTX 4090 GPUs with 24GB of memory each.

In Section 3.3.1, Step 1: Multi-Sampling Ranking-Based Term Generation, the temperature was set to 0.5, and different random seeds were utilized. Due to time constraints, sampling experiments were conducted exclusively on the all-subjects dataset, with the number of samples set to 50, 70, and 100.

In Section 3.1.1, experimental results indicate that the category distribution proportions between the two datasets do not exhibit significant differences. Consequently, the proportion of synthesized data was determined based on the category distribution in the training set of the all-subjects dataset, ensuring a one-to-one correspondence between English and German. The specific distribution ratios are as follows: "Book" (0.74), "Thesis" (0.15), "Conference" (0.07), "Report" (0.03), and "Article" (0.01).

5 Result

The overview statistics of six different models on the all-subjects dataset in the validation set are presented in Table 1. Among these models, $M_{samplings(100)}$ achieves the highest performance in 7 out of 10 categories and also demonstrates the best overall recall across all categories. Similarly, the summary statistics for three different models

Record Type	Language	$M_{\text{sft(raw)}}$	$M_{\text{sft(all)}}$	M_{dpo}	$M_{\text{sampling(50)}}$	$M_{\text{sampling(70)}}$	$M_{\text{sampling(100)}}$
Metric: AVG Recall							
Article	de	0.0000	0.0000	**1.0000**	0.9000	0.9000	0.9000
Article	en	0.5667	0.5477	0.5906	**0.7458**	0.7452	0.7440
Book	de	0.6075	0.6076	0.6158	0.7164	0.7199	**0.7219**
Book	en	0.5414	0.5443	0.5505	0.6700	0.6753	**0.6789**
Conference	de	0.5344	0.5360	0.5495	0.6686	0.6747	**0.6798**
Conference	en	0.5412	0.5340	0.5559	0.6843	0.6893	**0.6921**
Report	de	0.6096	0.6310	0.6279	0.7234	**0.7306**	0.7298
Report	en	0.4686	0.4738	0.5121	0.6314	0.6395	**0.6425**
Thesis	de	0.4507	0.4612	0.4610	0.5986	0.6036	**0.6066**
Thesis	en	0.3864	0.3958	0.4069	0.5510	0.5556	**0.5571**
All	All	0.4707	0.4731	0.5870	0.6890	0.6934	**0.6953**

Table 1: Performance comparison for six models on all-subjects validation set (measured by average recall).

on the tib-core-subjects dataset in the validation set are provided in Appendix Table 4, where M_{dpo} attains the highest average recall. However, due to time constraints, we did not conduct additional experiments to optimize the sampling size for the tib-core-subjects dataset. Instead, we applied the optimal sampling size of 100, as determined from the all-subjects dataset, for test set predictions in both the all-subjects and tib-core-subjects datasets.

For the BGE-M3 subject terms generation strategies discussed in Section 3.3.1 Step 2, experimental results in Appendix Table 5 indicate that evaluating similarity exclusively between the title and subject terms leads to superior performance on the all-subjects validation set. Consequently, title-based vector retrieval was adopted for test set predictions in both the all-subjects and tib-core-subjects datasets.

The official quantitative rankings, as presented in Appendix Table 6 and Appendix Table 7, indicate that our system achieved **2nd** place in the all-subjects dataset and **4th** place in the tib-core-subjects dataset. These results highlight the system's strong capability in predicting a top-k list of relevant GND subject terms based on the titles and abstracts of technical records. Regarding the re-ranking strategies discussed in Section 3.3.2, the corresponding results are provided in Appendix Table 8. Experimental findings show that re-ranking strategy 2 yields the best performance in the all-subjects dataset, whereas strategy 1 performs optimally in the tib-core-subjects dataset.

The performance of our system varies across the two datasets, with a higher ranking in the all-subjects dataset compared to the tib-core-subjects dataset. Due to time constraints, we did not investigate the optimal sample size for the tib-core-subjects dataset and instead applied the optimal sample size determined from the all-subjects dataset. Additionally, our approach exhibits suboptimal performance in the article category of the tib-core-subjects dataset. The primary reason for this is the limited availability of training data for articles. Even though data synthesis was conducted based on the category distribution in the training set, the quantity of article-related data remained insufficient. This data scarcity hindered the model's ability to effectively learn the relevant knowledge, ultimately resulting in weaker performance.

6 Conclusion

This paper presents our system for SemEval-2025 Task 5, which integrates GND classification knowledge into LLMs via data synthesis, LoRA fine-tuning, preference optimization, and frequency-position ranking over multiple samplings. Combined with LLM-based extraction, vector retrieval, and re-ranking, our approach addresses multilingual and domain-specific challenges in TIBKAT subject assignment. The system ranks **2nd** on the all-subjects dataset and **4th** on tib-core-subjects quantitatively, and **2nd** on tib-core-subjects qualitatively. Performance is strong overall, though limited training data for articles in tib-core-subjects affects accuracy. Future work may improve sampling and re-ranking under data-scarce settings.

Acknowledgments

We thank our anonymous reviewers for their helpful comments. This work was supported by grant from the Fundamental Research Funds for the Cen-

tral Universities (DUT24MS003), and the Liaoning Provincial Natural Science Foundation Joint Fund Program(2023-MSBA-003).

References

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Richard Brzustowicz. 2023. From chatgpt to catgpt: the implications of artificial intelligence on library cataloging. *Information Technology and Libraries*, 42(3).

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2023. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *Preprint*, arXiv:2309.07597.

DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Haowei Zhang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Li, Hui Qu, J. L. Cai, Jian Liang, Jianzhong Guo, Jiaqi Ni, Jiashi Li, Jiawei Wang, Jin Chen, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, Junxiao Song, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Lei Xu, Leyi Xia, Liang Zhao, Litong Wang, Liyue Zhang, Meng Li, Miaojun Wang, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingming Li, Ning Tian, Panpan Huang, Peiyi Wang, Peng Zhang, Qiancheng Wang, Qihao Zhu, Qinyu Chen, Qiushi Du, R. J. Chen, R. L. Jin, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, Runxin Xu, Ruoyu Zhang, Ruyi Chen, S. S. Li, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shaoqing Wu, Shengfeng Ye, Shengfeng Ye, Shirong Ma, Shiyu Wang, Shuang Zhou, Shuiping Yu, Shunfeng Zhou, Shuting Pan, T. Wang, Tao Yun, Tian Pei, Tianyu Sun, W. L. Xiao, Wangding Zeng, Wanjia Zhao, Wei An, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, X. Q. Li, Xiangyue Jin, Xianzu Wang, Xiao Bi, Xiaodong Liu, Xiaohan Wang, Xiaojin Shen, Xiaokang Chen, Xiaokang Zhang, Xiaosha Chen, Xiaotao Nie, Xiaowen Sun, Xiaoxiang Wang, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xingkai Yu, Xinnan Song, Xinxia Shan, Xinyi Zhou, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, Y. K. Li, Y. Q. Wang, Y. X. Wei, Y. X. Zhu, Yang Zhang, Yanhong Xu, Yanhong Xu, Yanping Huang, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Li, Yaohui Wang, Yi Yu, Yi Zheng, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Ying Tang, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yu Wu, Yuan Ou, Yuchen Zhu, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yukun Zha, Yunfan Xiong, Yunxian Ma, Yuting Yan, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Z. F. Wu, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhen Huang, Zhen Zhang, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhibin Gou, Zhicheng Ma, Zhigang Yan, Zhihong Shao, Zhipeng Xu, Zhiyu Wu, Zhongyu Zhang, Zhuoshu Li, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Ziyi Gao, and Zizheng Pan. 2024. Deepseek-v3 technical report. *Preprint*, arXiv:2412.19437.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

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.

Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Diego Rojas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, Hao Yu, Hongning Wang, Jiadai Sun, Jiajie Zhang, Jiale Cheng, Jiayi Gui, Jie Tang, Jing Zhang, Juanzi Li, Lei Zhao, Lindong Wu, Lucen Zhong, Mingdao Liu, Minlie Huang, Peng Zhang, Qinkai Zheng, Rui Lu, Shuaiqi Duan, Shudan Zhang, Shulin Cao, Shuxun Yang, Weng Lam Tam, Wenyi Zhao, Xiao Liu, Xiao Xia, Xiaohan Zhang, Xiaotao Gu, Xin Lv, Xinghan Liu, Xinyi Liu, Xinyue Yang, Xixuan Song, Xunkai Zhang, Yifan An, Yifan Xu, Yilin Niu, Yuantao Yang, Yueyan Li, Yushi Bai, Yuxiao Dong, Zehan Qi, Zhaoyu Wang, Zhen Yang, Zhengxiao Du, Zhenyu Hou, and Zihan Wang. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *Preprint*, arXiv:2406.12793.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *International Conference on Learning Representations*.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *CoRR*, abs/2106.09685.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.

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. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Yuqi Wang, Wei Wang, Qi Chen, Kaizhu Huang, Anh Nguyen, and Suparna De. 2023. Prompt-based zero-shot text classification with conceptual knowledge. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 30–38, Toronto, Canada. Association for Computational Linguistics.

Jason Wei and Kai Zou. 2019. Eda: Easy data augmentation techniques for boosting performance on text classification tasks. 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 6382–6388, Hong Kong, China. Association for Computational Linguistics.

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*.

A Table

Dataset	all	tib-core
train	81937	41902
labled train	60965	32043
CD	243856	109242
FD	218728	109242
dev	13666	6980
test	27986	6174

Table 2: Dataset distribution, CD refers to constructed data in section 3.1.1, FD refers to filterd data in section 3.1.2.

Setting	**T1**	**T2**	**T3**
Epochs	3	3	1
Max Sequence Length	256	512	600
Batch Size	1	1	1
Optimizer	Adam	Adam	Adam
Learning Rate	3e-5	3e-5	2e-6
Lora Rank	16	16	16
Lora Alpha	32	32	64
Gradient Accumulation Step	8	8	4

Table 3: Hyper-parameter settings of the experiment(T1, T2, and T3 represent the following stages: GND SFT, TIBKAT SFT, and TIBKAT DPO, respectively).

Type	**Lang.**	$M_{\text{sft(raw)}}$	$M_{\text{sft(all)}}$	M_{dpo}
Metric: AVG Recall				
Article	de	0.0000	**0.5000**	**0.5000**
Article	en	0.1394	0.1175	**0.1405**
Book	de	0.5023	0.5050	**0.5080**
Book	en	0.4945	0.5060	**0.5094**
Conference	de	0.3564	**0.3709**	**0.3709**
Conference	en	0.4447	0.4537	**0.4628**
Report	de	0.5464	**0.5769**	0.5645
Report	en	0.4131	0.3954	**0.4447**
Thesis	de	0.2951	0.2961	**0.2969**
Thesis	en	0.2992	**0.3294**	0.3268
All	All	0.3491	0.4055	**0.4124**

Table 4: Performance comparison for three models on the *tib-core-subjects* validation set. **Lang.** is short for language.

Metric	**Title**	**Abstract**	**Title + Abstract**
AVG Recall	**0.4707**	0.4236	0.4465

Table 5: Comparison of three search strategies using $M_{sft(raw)}$ on the all-subjects dev set.

Rank	**Team**	**Average Recall**
1	Annif	0.6295
2	**DUTIR831**	**0.6045**
3	RUC-Team	0.5856
4	dnb-ai-project	0.5631
5	icip	0.5302

Table 6: Results of top 5 teams and average recall on leaderboard for the all-subjects test set.

Rank	Team	Average Recall
1	RUC-Team	0.6568
2	Annif	0.5899
3	LA2I2F	0.5794
4	**DUTIR831**	**0.5599**
5	icip	0.4976

Table 7: Results of top 5 teams and average recall on leaderboard for the tib-core-subjects test set.

Dataset	s1	s2	s3
all	0.6016	**0.6151**	0.6121
tib-core	**0.4774**	0.4706	0.4699

Table 8: Results of 3 re-rank strategies and average recall on leaderboard for test set.

B Prompt

B.1 Incremental Data Generation Prompt Template

[instruction]
You are a subject expert in library sciences. From the provided candidate terms sourced from an internationally recognized authority file widely used by German-speaking libraries, your task is to select a subset of terms that reflect a logical and focused theme of a real research paper.
Caution: The terms you select should be fewer but precise. The terms should be ordered from highest to lowest relevance to the generated title and abstract.
Then, based on the selected terms, generate the appropriate title (under 20 words) and abstract (under 150 words) for the research paper. Ensure that the generated title and abstract are logical, relevant, and well-aligned with the chosen subject terms.
[input]
Candidate terms: ["Digital Libraries", "Metadata Management", "User Experience", "Library Automation", "Data Security"]
[output]
{"{"Selected terms": ["Digital Libraries", "Metadata Management", "User Experience"], "Title": "Enhancing User Experience in Digital Libraries", "Abstract": "This research paper focuses on how ..." }}

B.2 Data Filtering Prompt Template

[instruction]
You are tasked with evaluating the quality of titles and abstracts based on two key criteria: logical coherence and relevance to the selected subject terms. You will be given a title and an abstract of a `doc_type`. Your task is to assess the quality of the title and abstract based on how logically consistent and well-structured they are, as well as how accurately they reflect the core ideas of the selected subject terms. The subject terms are arranged in descending order with respect to relevance, with the most relevant terms appearing first.
Evaluation Criteria for Titles and Abstracts: Logical Coherence: Assess whether the title and abstract reflect the content of a real-world `doc_type`, ensuring alignment with real-world logic. 1: Very PoorThe title or abstract is completely illogical or lacks structure (e.g., random or unrelated phrases). 2: PoorThe title or abstract has noticeable logical flaws or is hard to understand. 3: AverageSomewhat logical, but there is room for improvement. 4: GoodMostly logical and clear, with minor room for improvement. 5: Excellent-Perfectly logical, clear, and well-structured.
Relevance to Selected Subject Terms: Assess whether the title and abstract fully reflect all the selected subject terms and whether the relevance order is correct. 1: Very PoorCompletely irrelevant to the subject terms. 2: PoorPartially related but missing the main focus. 3: AverageReflects most subject terms but lacks completeness. 4: GoodClearly reflects all subject terms, with minor misalignment. 5: Excellent-Fully reflects all subject terms in the correct order. Please provide your evaluation results in the following JSON format:
{"{"Final_Score":{"Logical_Coherence_Score": <score>,
"Relevance_to_Selected_Subject_Terms_Score":<score>}}"}

PLEASE JUDGE WITH STRICT STANDARDS AND DO NOT OUTPUT ANY EXPLANATION.
[input]
selected_subjects
title_and_abstract
[output]
{"{"Final_Score":{"Logical_Coherence_Score": <score>,
"Relevance_to_Selected_Subject_Terms_Score": <score>"}}"}

B.3 GND Knowledge Learning Prompt Template

[instruction]
You will be given a term from the Integrated Authority File Sachbegriffe, an international authority file widely used by German-speaking libraries for cataloging and linking information about works. Your task is to enhance understanding and usability of the term by generating the following attributes: Classification Name, Classification Number, Alternate Name, Related Subjects, and Definition (in German).
[input]
The term to process is: Abhängigkeit.
[output]
{"{"Classification Number": "1", "Classification Name": "Allgemeines, Interdisziplinäre Allgemeinwörter", "Alternate Name": ["Dependenz", "Unselbständigkeit", "Unselbstständigkeit"], "Related Subjects": ["Abhängiger", "Selbstständigkeit", "Interdependenz"] }"}

B.4 Supervised Fine-Tuning Prompt Template

[instruction]
Given the title and abstract of a doc_type in English or German, generate a ranked list of the most relevant subject terms from the Integrated Authority File Sachbegriffe, an international authority file widely used by German-speaking libraries for cataloging and linking information about works. These terms should accurately reflect the key themes or topics described in the title and abstract, with the ranking indicating their relative relevance.
[input]
Title: ...
Abstract: ...
[output]
{json_format_subject_terms}

B.5 GND Subject Terms Extraction Prompt Template

[instruction]
You are a subject term expert. Your task is to extract authoritative subject terms in German from the title and abstract of a given doc_type written in English or German. The subject terms must be selected exclusively from the Integrated Authority File (GND), an international authority file widely used by German-speaking libraries for cataloging and linking information about works. These terms should represent standardized academic concepts or disciplines. Do not provide any analysis or commentary.
<example>
Title: Nico Bloembergen : Master of Light
Abstract: This biography is a personal portrait of one of the ...
List of subject terms: {["Laser","Maser"]}
</example>
[input]
Title: ...
Abstract: ...
[output]
{json_format_subject_terms}

B.6 GND Subject Terms Rerank Prompt Template

[instruction]
You are a subject term expert. Firstly, you need to analyze the title and abstract of a given doc_type in English or German. Then you will also be provided with a list of subject terms from the Integrated Authority File Sachbegriffe, an international authority file widely used by German-speaking libraries for cataloging

and linking information about works.
Your role is to evaluate and rank the subject terms based on their relevance to the title and abstract, from highest to lowest relevance. Ensure that your rankings reflect a clear and logical analysis of how well each term aligns with the main themes and concepts presented in the `doc_type`. DO NOT OUTPUT ANY ANALYSIS!"""
<example>
Title: ...
Abstract: ...
List of subject terms: ...
Rerank list subject terms: ...
</example>
[input]
Title: ...
Abstract: ...
List of subject terms: ...
[output]
Rerank list subject terms: ...