DCPAPERS

DCMI 2026

Proceedings of the International Conference on Dublin Core and Metadata Applications 2026

ISSN: 1939-1366

Automated Classification of Chinese Books: A Large Language Model Approach to Knowledge Transfer and Domain Adaptation

Xin Yang[1], **Junzhi Jia**[1], and **Ying-Hsang Liu**[2]

[1]Renmin University of China, No. 59 Zhongguancun Street, 100872 Beijing, China, [2]Professorship of Predictive Analytics, Chemnitz University of Technology, 09126 Chemnitz, Germany

Abstract

Automated subject indexing remains a critical challenge for digital libraries and knowledge organization systems. To address this issue, this study develops a knowledge-augmented domain adaptation framework that aligns general-purpose language models with the hierarchical logic of the Chinese Library Classification (CLC). A supervised fine-tuning (SFT) strategy is proposed to resolve domain knowledge drift and deep-category recognition bottlenecks in large language model (LLM)-based Chinese book indexing, using dual bibliographic and category data. Three evaluation experiments were conducted to assess the effectiveness of the proposed techniques for quantifying ontological contributions: hyperparameter sensitivity analysis for baseline establishment, backbone model comparison for architectural fitness, and knowledge injection ablation for. Results demonstrate that dual-data fine-tuning significantly enhances precision for long-tail and fine-grained categories. While ensuring high-quality output, the solution features low computational thresholds, robust local deployment, and high scalability, effectively internalizing knowledge organization systems within LLMs. By bridging the gap between classical theory and generative AI, this work provides a high-accuracy, institutionally autonomous solution for automated indexing, offering substantial theoretical and practical significance for the intelligent transformation of digital libraries.

1. Introduction

Since late 2022, the library community has actively explored integrating generative AI (GenAI) into technical service operations by leveraging its capacity for semantic generalization and deep contextual understanding. Cataloging practices, particularly subject indexing, is a natural focus of this inquiry. As an intellectually demanding task [1] requiring mastery of standards such as the DDC, LCSH, CLC, and MeSH, subject indexing involves assigning controlled notations through content analysis and thematic interpretation. Automated subject indexing has long been recognized as a challenging task in cataloging automation [2] partly due to the deep semantic knowledge required for the intellectual task.

From a technical perspective, both classification and subject indexing tasks constitute extreme multi-label classification (XMLC) problems [3]. Given label spaces spanning tens of thousands of categories and severe long-tail distributions, conventional machine learning and neural network approaches face significant hurdles in accuracy and scalability. LLMs are characterized by massive parameter counts, powerful training algorithms, and large-scale computation, through which they have internalized broad encyclopedic knowledge during pre-training. Existing research has shown that directly applying LLMs to the highly regulated task of indexing remains infeasible, with persistent issues of fabrication, classification errors, omissions, and hallucinations [4].

Full Paper

Available Aug 01, 2026

Correspondence to
Junzhi Jia
junzhij@163.com

However, these limitations can be addressed for the effective application of LLMs to subject indexing tasks. First, LLMs demonstrate strong in-context learning capabilities in natural language processing (NLP) tasks, enabling generalization and inference from limited examples. Second, their large parameter capacity and deep semantic understanding can provide a strong foundation for effective domain adaptation.

The library and knowledge organization (KO) community has increasingly recognized the need of domain adaptation for subject indexing tasks, among which supervised fine-tuning (SFT) has emerged as a promising and widely adopted approach [5]. However, it remains unclear to what extent SFT can facilitate the internalization of classification knowledge and hierarchical logic embedded in knowledge organization systems (KOS), such as the CLC. Previous studies have largely focused on indexing performance and feasibility, while providing limited evidence regarding how domain knowledge contributes to hierarchical alignment and deep-category recognition [4]. This study therefore proposes a knowledge-augmented SFT framework for CLC-Based indexing and evaluates its effectiveness in supporting hierarchical alignment and fine-grained category recognition under the Chinese Library Classification (CLC) system.

2. Related Work

Researchers have explored domain adaptation through three main directions: SFT, retrieval-augmented generation (RAG), and hybrid knowledge-constrained architectures.

Regarding SFT and its optimization, Hu et al. applied Low-Rank Adaptation (LoRA) to enhance deep-level CLC recognition and error correction [5], while Tian et al. implemented multi-stage adaptation involving knowledge distillation and preference optimization to internalize hierarchical logic [6]. For RAG and semantic mapping, Xia et al. developed the MaRSI framework to simulate expert indexing via similarity matching [7], and Luo and Liu utilized a dual pipeline of fine-tuning and controlled vocabulary retrieval to mitigate hallucinations. In terms of hybrid architectures and collaboration [8], Liu et al. integrated regression-based constraints with generative models to improve term alignment [9] Usta combined synthetic data generation with architecture ensembling to balance accuracy and efficiency [10] Kluge and Kähler leveraged multi-model voting and semantic mapping to improve recall [11]. Additionally, Golub et al. [12] and Suominen et al. [13] demonstrated that algorithmic ensembling and LLM-based translation significantly bolster system robustness and cross-lingual performance.

Although these approaches have improved indexing performance, it is challenging to enable LLMs to internalize the hierarchical logic of KOS like CLC. This study therefore investigates a knowledge-augmented SFT framework for domain adaptation in indexing and hierarchical alignment.

3. Methodology

3.1. Theoretical Basis and Research Question

From an information science perspective, book subject indexing is fundamentally a mapping from natural language semantic space onto a structured KOS [14]: it assigns a document, based on its explicit attributes and latent semantic content, to the corresponding category within that system. This process requires catalogers or automated systems to possess strong semantic recognition capabilities and reliable disciplinary judgment. While general-purpose LLMs offer broad semantic understanding, a substantial distributional gap exists between general-domain corpora and the rigorously hierarchical structure of CLC. Moreover, the evaluation result has demonstrated that zero-shot or few-shot indexing with general LLMs is not viable [15], with characteristic failures including difficulty distinguishing deep-level category boundaries, coarse-grained assignments, and category hallucination.

The central methodological challenge, therefore, is how to leverage the semantic competence and broad world knowledge of LLMs to improve their performance on indexing tasks. This study adopts transfer learning as its theoretical foundation. Transfer learning works by reusing general features learned in a source domain to support task completion in a target domain, thereby mitigating limitations arising from annotation scarcity or task complexity. This domain adaptation process utilizes specific fine-tuning and expert-annotated data to adjust model parameters. By calibrating internal attention mechanisms, it enables precise mapping from document semantics to the vast KOS, such as CLC. Accordingly, this study addresses the following research question: to what extent can domain knowledge-enhanced supervised fine-tuning improve hierarchical alignment and fine-grained category recognition in CLC-based indexing?

3.2. Knowledge-Augmented SFT Framework

Based on the transfer learning perspective outlined above, this study proposes a Knowledge-Augmented SFT Framework for CLC-Based Subject Indexing. The framework consists of three components: (1) bibliographic semantic data, (2) CLC hierarchical knowledge, and (3) LoRA-based supervised fine-tuning. By integrating domain knowledge into the adaptation process, the framework aims to improve hierarchical alignment and fine-grained category recognition.

This study adopts SFT as the primary domain adaptation pathway, enabling parameter-level alignment between general semantic representations and the disciplinary logic embedded in the CLC. Given the availability of expert-annotated bibliographic records and the need for efficient local deployment, a lightweight LoRA-based adaptation strategy is employed to balance indexing accuracy and computational cost.

4. Data Preparation and Experimental Design

CLC is a hierarchical tree-based classification system featuring 22 letter-coded main classes followed by numerical sub-levels. The indexing task addressed in this study

Table 1. *Structured instances for CLC indexing SFT*

Instance type	Structured instance
# Structured instances for bibliographic metadata：	{"messages": [{"role": "user", "content": "Title: Research on China's Urban-Rural Correlation and Symbiotic Development under the Evolution of Urban-Rural Relations \nSeries: Economic and Management Library \nAbstract: From the perspective of urban-rural interaction, this book incorporates institutional and technical variables into a symbiotic development framework. Utilizing New Economic Geography methods, it explores interaction mechanisms and evolutionary conditions to optimize urban-rural relations and promote integrated development, providing practical references for China's current urbanization phase."}, {"role": "assistant", "content": "F299.21"}]}
# Knowledge-augmented instances from CLC category definitions：	"messages": [{"role": "user", "content": "Define the following CLC code: A"}, {"role": "assistant", "content": "Marxism, Leninism, Mao Zedong Thought, and Deng Xiaoping Theory"}]

is formulated as a hierarchical classification problem, and the CLC structure directly informs data construction and experimental design.

4.1. Data Collection and Preprocessing

(1) Data Collection. Two data sources were used. The first comprises 32,837 CNMARC bibliographic records, from which titles, abstracts, and series information were extracted as semantic inputs. The second is the complete CLC (5th edition) dataset, yielding 53,326 class notation–category name pairs. These datasets jointly form the training and evaluation corpus for this study. For consistency, all bibliographic notations were validated against the CLC 5th edition, and the classification task was restricted to the standardized CLC category space.

(2) Dataset Construction. Bibliographic and category data were converted into structured fine-tuning instances (see Table 1). This dual-data approach maps document semantics to notations, while using category definitions to internalize the CLC's hierarchical logic. To optimize throughput, simplified task instructions were used to reduce per-sample token length, as empirical observation confirmed this does not degrade indexing performance.

Note: Original records are in Chinese; English translations are provided for illustrative purposes.

(3) Dataset Partitioning. Stratified sampling was applied to maintain consistent disciplinary distribution. To address long-tail distribution, categories with a frequency below five were assigned exclusively to the training set, while remaining categories followed a 9:1 split. For multi-label records, training data was decomposed into independent single-label instances to refine semantic correspondence, while original multi-label annotations were retained for testing to ensure realistic evaluation. The final dataset comprises 36,786 training instances and 2,019 validation instances.

4.2. Experimental Design

To identify the optimal trade-off between computational cost and indexing accuracy, this study develops a progressive technical framework consisting of three interrelated controlled experiments. These experiments form a hierarchical path where each stage establishes foundational parameters for the subsequent phase.

(1) Hyperparameter Sensitivity Experiment. This stage evaluates core SFT and LoRA hyperparameters, including rank (r), scaling factor (α), learning rate, batch size, and gradient accumulation steps, to derive a stable baseline configuration. By establishing

this unified control scheme, the study ensures rigorous variable control for subsequent model selection and knowledge injection.

(2) Backbone Model Comparative Experiment. Using the fixed baseline from Experiment 1, this study assesses the impact of model scale and architecture on domain adaptation. By comparing representative open-source LLMs under identical configurations, we quantify knowledge transfer efficiency across CLC long-tail distributions to guide model selection for resource-constrained library environments.

(3) Knowledge Injection Ablation Experiment. Built upon the optimal configuration and backbone from previous stages, this experiment isolates the contribution of dataset construction. We compare fine-tuning on bibliographic metadata alone versus a dual-data approach incorporating CLC notation-category name pairs. The aim is to quantify the role of CLC ontological knowledge in mitigating hallucinations and improving precision for deep-level category recognition.

4.3. Experimental Configuration

Experiments were conducted on the AutoDL platform (Ubuntu 22.04) using Python 3.12, PyTorch 2.8.0, and CUDA 12.8. Fine-tuning utilized LLaMA-Factory (v0.9.5.dev0) on an RTX 5090 (32 GB) GPU, incorporating LoRA for efficiency, 8-bit BitsAndBytes for quantization, and PyTorch SDPA for long-text acceleration.

To ensure deep Chinese semantic understanding and professional knowledge representation, we selected the Qwen model family, specifically Qwen 2.5-1.5B-Instruct, Qwen 2.5-7B-Instruct, and Qwen 3-8B. These instruction-tuned variants were chosen for their robust alignment and instruction-following capabilities, which eliminate the need for large-scale alignment from scratch while enhancing training efficiency. This selection spans multiple parameter scales, enabling a systematic investigation of the relationship between model size and indexing performance.

4.4. Evaluation Metrics

Evaluating CLC indexing requires addressing both classification accuracy and hierarchical coherence. This study utilizes precision, recall, and F1 scores through a layer-by-layer evaluation framework extending to Level 10. Instead of a global exact-match criterion, this hierarchical strategy quantifies performance degradation as depth increases, capturing the inherent complexity of XMLC tasks. This approach provides an objective assessment of logic internalization by isolating stable upper-level performance from specific errors at deeper subcategory levels.

5. Experimental Results

5.1. Hyperparameter Sensitivity Results

To balance experimental rigor with training efficiency, this experiment was conducted as a pilot study. Smaller-scale models exhibit sufficient sensitivity gradients in hyperparameter response to serve as proxies for full-scale experiments. Qwen 2.5-1.5B-Instruct

Table 2. *Hyperparameter sensitivity results for qwen 2.5-1.5b fine-tuning*

	1.5b-Exp-01(13_1615)			1.5b-Exp-02(13_2058)			1.5b-Exp-03(29_1930)		
Level	Precise	Recall	F1	Precise	Recall	F1	Precise	Recall	F1
1	73.05%	100.00%	84.43%	83.61%	100.00%	91.08%	83.47%	100.00%	90.99%
2	62.57%	100.00%	76.98%	78.84%	98.35%	87.52%	78.42%	98.40%	87.28%
3	53.89%	99.80%	69.98%	74.39%	96.96%	84.19%	74.13%	96.80%	83.96%
4	46.66%	97.97%	63.21%	70.10%	91.74%	79.47%	70.40%	91.50%	79.58%
5	40.03%	97.09%	56.69%	65.95%	77.22%	71.14%	66.26%	77.59%	71.48%
6	33.87%	34.18%	34.03%	59.32%	55.37%	57.28%	60.38%	57.92%	59.13%
7	32.54%	30.34%	31.40%	49.16%	35.47%	41.21%	48.79%	38.11%	42.79%
8	29.71%	14.59%	19.57%	36.88%	19.88%	25.83%	36.49%	22.36%	27.73%
9	24.82%	0.00%	0.00%	27.62%	11.43%	16.17%	27.09%	14.29%	18.71%
10	23.06%	0.00%	0.00%	26.89%	13.51%	17.99%	25.95%	13.51%	17.77%

was selected as the test backbone to rapidly identify optimal configurations. Table 2 reports results from three representative runs out of ten, with corresponding configurations detailed in Table 3.

Based on the hyperparameter sensitivity analysis, the configuration from Experiment 02 was adopted as the unified baseline for all subsequent experiments, employing a learning rate of 5×10^{-5}, an effective batch size of 8, max samples >50,000, reasoning enabled = ture, and LoRA parameters $r = 8$ and $\alpha = 16$.

(1) Impact of sample scale on deep-level recognition. Expanding training samples from 10,000 to 100,000 (Exp 01 vs. Exp 02) yielded a 10-15% F1 gain at Levels 1-5. Critically, larger data volume activated recognition of long-tail categories, raising Level 10 F1 from 0% to 17.99%. This confirms that absolute data volume is a prerequisite for extracting differentiating semantic features in high-density XMLC tasks. This result suggests that large-scale hierarchical classification systems such as the CLC require sufficient representational coverage of bibliographic evidence to support fine-grained conceptual differentiation. In knowledge organization terms, deeper category boundaries are not defined by surface lexical cues alone, but by cumulative distinctions among disciplinary domains, subfields, and cataloging conventions. Therefore, a diverse and adequately sized set of training instances is necessary for the model to approximate the classificatory judgment required to distinguish semantically adjacent categories across the CLC hierarchy.

(2) Influence of learning dynamics. Learning efficiency was found highly sensitive to effective batch size and update frequency. An excessively large effective batch size, such as 192 led to premature convergence toward coarse-grained patterns, resulting in degraded classification performance. From a knowledge organization perspective, such learning dynamics favor superficial regularities while failing to support the gradual formation of stable conceptual boundaries. For large-scale hierarchical systems like CLC, effective domain adaptation depends on fine-grained, incremental parameter

Table 3. *CLC indexing fine-tuning results across model variants*

	Qwen 2.5-1.5b-Instruct			Qwen 2.5-7b-Instruct			Qwen 3-8b		
level	Precise	Recall	F1	Precise	Recall	F1	Precise	Recall	F1
1	83.61%	100.00%	91.08%	86.43%	100.00%	92.72%	86.38%	100.00%	92.69%
2	78.84%	98.35%	87.52%	81.64%	99.95%	89.87%	82.03%	100.00%	90.13%
3	74.39%	96.96%	84.19%	76.80%	98.79%	86.42%	77.79%	99.24%	87.22%
4	70.10%	91.74%	79.47%	72.25%	92.30%	81.05%	72.84%	96.04%	82.85%
5	65.95%	77.22%	71.14%	67.52%	81.16%	73.72%	68.60%	88.58%	77.32%
6	59.32%	55.37%	57.28%	61.77%	60.76%	61.26%	62.53%	62.11%	62.32%
7	49.16%	35.47%	41.21%	53.96%	32.61%	40.65%	56.05%	35.61%	43.55%
8	36.88%	19.88%	25.83%	46.48%	18.84%	26.81%	49.10%	25.58%	33.64%
9	27.62%	11.43%	16.17%	39.84%	5.83%	10.17%	40.93%	5.21%	9.25%
10	26.89%	13.51%	17.99%	36.88%	4.21%	7.56%	37.30%	1.75%	3.35%

updates that allow the model to progressively differentiate closely related categories across levels, rather than relying on shallow pattern memorization.

5.2. Backbone Model Performance Comparison Results

With the fine-tuning strategy and training corpus held constant, this experiment evaluates model adaptability across parameter scales for CLC hierarchical classification (see Table 3).

Experimental results exhibit a clear scale dependency in model capacity. Qwen 3-8B demonstrated superior precision across all levels, with F1 scores exceeding 90% at Levels 1 and 2, suggesting that larger parameter spaces better encode latent disciplinary signals. Conversely, the 1.5B model showed lower precision despite high shallow-level recall, reflecting insufficient capacity for fine-grained discrimination at semantically ambiguous boundaries.

Performance degrades nonlinearly with increasing depth, driven by a progressive dilution of semantic distinctiveness. Beyond Level 6, the narrowing of intra-class semantic distance and the intensification of inter-class overlap exponentially increase discriminative difficulty in the extreme label space. While Qwen 2.5-1.5B precision falls below 30% at Levels 7–10, Qwen 3-8B maintains precision above 37% at Level 10, underscoring the necessity of parameter capacity for capturing fine-grained professional context.

Although Qwen 3-8B achieves a superior balance with an F1 of 77.32% at Level 5, diminishing marginal returns relative to the 7B variant are notable. Given the exacting standards of library cataloging, Qwen 2.5-7B offers the most balanced trade-off between computational cost and indexing accuracy for practical deployment.

5.3. Knowledge Injection Ablation Results

Qwen 3-8B was selected for the ablation study due to its superior performance in Experiment 2. This strategy aims to transition the model from statistical association

Table 4. *Ablation results of knowledge injection on Qwen-3-8B*

	Qwen3-8B (Base)			Qwen3-8B (Enhanced)			Performance Gains		
level	Precise	Recall	F1	Precise	Recall	F1	ΔP	ΔR	ΔF1
1	86.38%	100.00%	92.69%	88.11%	100.00%	93.68%	1.73%	0.00%	0.99%
2	82.03%	100.00%	90.13%	84.22%	100.00%	91.43%	2.19%	0.00%	1.30%
3	77.79%	99.24%	87.22%	80.36%	99.65%	88.97%	2.57%	0.41%	1.75%
4	72.84%	96.04%	82.85%	76.59%	96.85%	85.54%	3.75%	0.81%	2.69%
5	68.60%	88.58%	77.32%	72.69%	90.31%	80.55%	4.09%	1.73%	3.23%
6	62.53%	62.11%	62.32%	67.73%	72.29%	69.94%	5.20%	10.18%	7.62%
7	56.05%	35.61%	43.55%	61.59%	44.09%	51.39%	5.54%	8.48%	7.84%
8	49.10%	25.58%	33.64%	53.08%	26.51%	35.36%	3.98%	0.93%	1.72%
9	40.93%	5.21%	9.25%	44.99%	7.06%	12.20%	4.06%	1.85%	2.95%
10	37.30%	1.75%	3.35%	41.51%	3.16%	5.87%	4.21%	1.41%	2.52%

to logically guided classification by utilizing CLC's category definitions as training constraints.

Instead of uniformly injecting the full CLC dataset, this study employs a sampling strategy focused on deep-level categories with extended notation paths. This approach leverages the hierarchical nesting of CLC; by processing deep-level notations, the model repeatedly activates the semantic prefixes representing their corresponding upper-level classes. This repetitive activation reinforces the model's internal representation of the overall hierarchical structure through the inherent inclusion relationships within the notation strings.

The enhanced model achieved consistent performance gains across all classification levels (Table 4). Improvements at Levels 1-3 were modest, indicating that broad disciplinary distinctions are largely internalized during large-scale pre-training. The most substantial and structurally meaningful gains were observed at levels 4–6, where F1 scores increased steadily by 2.69% to 7.62%. These levels correspond to the core strata of the CLC hierarchy, where broad disciplinary classes are systematically differentiated into more specialized subject domains. The introduction of CLC category knowledge therefore helped reinforce the model's sensitivity to hierarchical relations and conceptual boundaries within the classification system. However, at Levels 8–10, the gains became increasingly constrained by semantic sparsity and conceptual fragmentation. This indicates that parameter-efficient fine-tuning alone cannot fully reproduce the deepest classificatory decision logic of the CLC, where expert cataloging judgment and highly specific domain knowledge remain essential.

5.4. Qualitative Error Analysis

To complement the quantitative evaluation, we conducted a brief qualitative analysis of representative successful and erroneous classifications produced by the best-performing models. The cases reveal distinct patterns across the CLC hierarchy. Successful classifications typically occur when bibliographic resources contain clear disciplinary signals,

enabling accurate alignment with both core classes and auxiliary notations. In contrast, shallow errors often arise in interdisciplinary works, where models prioritize dominant semantic themes over cataloging conventions. Most errors occur at intermediate levels of the hierarchy, where models correctly identify the general subject area but fail to follow the precise classificatory path or assign auxiliary subdivisions. Deep-level errors are concentrated in highly specialized categories with sparse training examples, indicating limitations in fine-grained hierarchical reasoning. Overall, the results suggest that knowledge-augmented fine-tuning substantially improves CLC adaptation, while intermediate and deep-level classification remains challenging for automated subject indexing.

6. Discussion and Conclusion

6.1. Discussion

This study explores an efficient pathway to automated Chinese book classification through LoRA-based SFT of open-source LLMs. Compared with direct API-based use of large proprietary models, the local fine-tuning approach offers substantial advantages in professional specificity and data privacy protection; after targeted adaptation, models can deeply internalize the classification logic of CLC. The best-performing model, Qwen 3-8B, achieves a peak global F1 score of 93.68% at Level 1 and maintains 80.55% at Level 5. While the current state of the technique is not yet sufficient to replace human review entirely, it holds considerable promise as an assistive tool for substantially reducing the repetitive workload of cataloging staff.

Across the three controlled experiments, several consistent patterns emerge. Larger parameter models provide richer semantic representation space, enabling more accurate capture of latent disciplinary signals in bibliographic abstracts, though the relationship between model scale and indexing performance is nonlinear. The hyperparameter sensitivity experiment reveals that XMLC tasks are strongly dependent on data volume, with sample scale being decisive for activating recognition of long-tail categories. The knowledge injection experiment demonstrates that CLC ontological content plays a key role in supporting fine-grained semantic discrimination [16], particularly at deeper classification levels where the introduction of CLC category data substantially improves differentiation between semantically adjacent disciplinary boundaries. Taken together, these findings suggest that improving automated indexing depends not only on scaling model size but critically on deep alignment between high-quality domain knowledge and general semantic representations.

6.2. Conclusion

In summary, LoRA-based SFT provides a viable approach to automated Chinese book subject indexing that balances accuracy with cost. Dual-data fine-tuning combining bibliographic semantic data with CLC category knowledge is an effective method for improving LLM-based indexing performance. By incorporating a domain knowledge dataset encoding category definitions and hierarchical relationships, the SFT approach

demonstrates strong knowledge internalization when processing CLC's complex hierarchical nesting. The fine-tuning procedure developed here exhibits robust knowledge transfer in deep classification tasks and clearly captures the characteristic performance degradation as semantic distinctiveness decreases across levels, reflecting the representational bottlenecks that remain for LLMs at extremely fine-grained classification depths. Low computational requirements make this approach a viable indexing support tool for small to medium-sized libraries and community documentation centers.

Several limitations merit acknowledgment. Compared with RAG approaches that enable retrieval-based reasoning, the SFT pathway operates as a black box, and its performance is highly sensitive to precise hyperparameter tuning, introducing a degree of interpretability limitation. Dataset construction and computational resource management during fine-tuning also retain a degree of technical complexity that could impede large-scale deployment. Finally, the generalizability of this approach to international indexing schemes with differing structures, including DDC and LCSH, remains to be investigated, as the current study focus is limited to the CLC system.

Acknowledgements

This study was supported by the Chinese Library Society Key Research Project "Research on Intelligent Classification Indexing of Chinese Books Using Generative AI" (No. 2026LSCZZZD0003) and the Deutsche Forschungsgemeinschaft (DFG) project "Intentional Forgetting and Changes in Work Processes: A Process-Conditional Approach in the Administrative and IT Context" (No. 427257555).

References

[1] E. Svenonius, *The intellectual foundation of information organization*. MIT Press, 2000.

[2] K. Golub, "Automated subject indexing: An overview," *Cataloging & Classification Quarterly*, vol. 59, pp. 702–719, 2021, doi: 10.1080/01639374.2021.2017386.

[3] K. Bhatia, K. Dahiya, H. Jain, Y. Prabhu, and M. Varma, "Extreme multi-label classification: A survey," 2016.

[4] B. Dobreski and C. Hastings, "AI chatbots and subject cataloging: A performance test," *Library Resources & Technical Services*, vol. 69, 2025.

[5] Z. Hu, D. Shui, and J. Wu, "Generative and hierarchical classification of literature based on fine-tuned large language models," *Journal of the China Society for Scientific and Technical Information*, vol. 44, pp. 425–437, 2025.

[6] Y. Tian *et al.*, *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). 2025, pp. 363–372.

[7] T. Xia *et al.*, *RUC Team at SemEval-2025 Task 5: Fast automated subject indexing via similar records matching and related subject ranking*. in Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025). 2025, pp. 2437–2442.

[8] P. Luo and Z. Liu, "Automatic subject indexing with large language model fine-tuning and controlled vocabulary retrieval," *Aslib Journal of Information Management*, 2025, doi: 10.1108/AJIM-05-2025-0253.

[9] J. Liu, X. Song, D. Zhang, J. Thomale, D. He, and L. Hong, "A hybrid framework for subject analysis: Integrating embedding-based regression models with large language models," University of North Texas, 2025.

[10] G. Usta, "Transformer and statistical models for LCSH assignment: A comparative study in digital libraries," *The Electronic Library*, vol. 43, pp. 695–714, 2025.

[11] L. Kluge and M. Kähler, *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). 2025, pp. 1118–1128.

[12] K. Golub, O. Suominen, A. T. Mohammed, H. Aagaard, and O. Osterman, "Automated Dewey Decimal Classification of Swedish library metadata using Annif software," *Journal of Documentation*, vol. 80, pp. 1057–1079, 2024.

[13] O. Suominen, J. Inkinen, and M. Lehtinen, *Annif at SemEval-2025 Task 5: Traditional XMTC augmented by LLMs*. in Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025). 2025, pp. 2424–2431.

[14] B. Hjørland, "Domain analysis in information science: Eleven approaches—traditional as well as innovative," *Journal of Documentation*, vol. 58, pp. 422–462, 2002, doi: 10.1108/00220410210431136.

[15] S. Deng, "AI, cataloging & metadata," University of Central Florida, 2023.

[16] G. Hodge, "Systems of knowledge organization for digital libraries: Beyond traditional authority files," CLIR, 2000.