

Exploring the Application of Large Language Models in Library Information Organization

Workshop

Lingzi Hong

Jason Thomale

Charlene Chou

Jiangping Chen

University of North Texas, USA | lingzi.hong@unt.edu

University of North Texas Libraries, USA | jason.thomale@unt.edu

New York University Libraries, USA | charlene.chou@nyu.edu

University of Illinois Urbana-Champaign | jpchen@illinois.edu

ABSTRACT

This two-session, 3-hour workshop aims to foster an understanding of how Large Language Models (LLMs) can be applied to help libraries organize and manage information and to promote discussion about their potential applicability and how they might be effectively integrated into library workflows. LLMs are highly capable in natural language processing tasks. Researchers and practitioners are exploring whether this technology can assist in traditional information organization tasks, such as classification and subject analysis in cataloging. This workshop will: 1) highlight trending research efforts in developing LLMs for information organization tasks; and 2) discuss LLM adoption in library workflows, the potential impact, and applicability. This workshop serves as a platform to gather researchers and practitioners from diverse backgrounds, who share similar interests and have done pioneering work in this field, to communicate research findings, build collaborations, discuss implications, and foster new research agendas.

KEYWORDS

LLMs, library, information organization, applicability, impact

PURPOSE AND INTENDED AUDIENCE

LLMs pre-trained on vast and diverse datasets have demonstrated high intelligence in natural language processing tasks, including text classification, summarization, generation (Achiam et al., 2023), and translation (Wang et al., 2023). LLMs' potential impact has attracted the attention of many researchers and practitioners in the Library and Information Science (LIS) field. Information organization researchers in particular are exploring models to help with tasks such as knowledge object classification and subject analysis, aiming to understand how LLMs might assist library catalogers and automate cataloging processes, and how this might impact the role of catalogers in libraries.

Researchers began exploring using artificial intelligence (AI) to assist the organization of knowledge objects in the early 2000s (Wang, 2009) with the development of machine learning models for classification (Golub et al., 2020; Kragelj & Kljajic Borstnar, 2021) and keyword extraction (Wu et al., 2023). However, due to inferior accuracy, little of that research has transferred to real-world use. Most of the work remains in the exploration and testing stages, attempting to identify more accurate models. Nowadays, we are witnessing AI develop at an accelerated rate. LLMs have shown superior capability in many natural language processing tasks compared to traditionally trained machine learning models – even with no additional training. This sparks people's imagination about LLMs' potential real-world applications in assisting people and automating their work. We seem to be standing at a turning point in which traditional library work may be facing a significant transformation (Marchionini, 2024). Our proposed event would bring interested researchers and practitioners together to discuss new LLM-related research in the field of LIS and, critically, the potential application of these techniques to real-world systems.

The iConference is a prestigious venue, attracting LIS professionals from diverse backgrounds with diverse perspectives. This workshop intends to serve as a platform to bridge cutting-edge research and practices, attracting researchers, library practitioners (such as system developers and catalogers), and graduate students in LIS.

Recently, researchers in LIS and Computer Science have started to explore methods for training LLMs to complete cataloging tasks (Martorana et al., 2024; Zhang et al., 2023), but many studies are still in progress. We hope that, through this platform, we can bring together professionals who are in the exploratory phase to share their current progress, showcase the challenges they are encountering, and discuss key research questions for the advancement of this field. Communication is essential for building upon current solutions and working collaboratively to answer challenging questions. The proposed workshop would also give researchers the opportunity to learn from library practitioners about what kinds of real-world tasks might lend themselves to this kind of research and how such research may be applied in new and helpful ways.

Library professionals who may have a general understanding of LLMs' capabilities and an awareness of their potential for assisting in cataloging-related work may lack the technical expertise to implement LLM-based tools that would address their specific needs, workflows, and contexts. By engaging in these conversations with researchers, they can gain a deeper understanding of what LLM-based tools are available and their specific capabilities, including open-source models such as Llama and BERT and proprietary models such as GPT-4. They will also gain insight into LLMs' performance on real cataloging data, allowing them to better assess how these AI technologies can be utilized and the corresponding impact of LLMs.

Library technology and systems developers who are responsible for developing or maintaining library information systems may gain a better understanding of the possibility of integrating AI-driven functions into their existing infrastructure. LIS students – the future of the LIS field – will gain a comprehensive understanding of both the theoretical and practical applications of LLMs in library information management and retrieval, thus preparing them for the challenges AI will pose to LIS in the future.

PROPOSED FORMAT

This two-session (3-hour), in-person workshop is expected to have 20-30 attendees. The workshop is designed to bridge the gap between the theoretical and technological development of LLMs for cataloging and the practical integration of LLMs in library systems and cataloger workflows. It will also serve as a platform where researchers and practitioners who share similar interests in exploring the use of LLMs in information organizations can communicate, collaborate, and network.

The workshop will comprise two parts. The first will focus on the recent research progress, updates on new projects, and community practices in applying LLMs to different types of cataloging tasks, such as classifying knowledge objects based on the Library of Congress Classification Code, extracting subject headings, and generating metadata. We will send a call for Work-in-Progress paper presentations on this topic through channels such as ASIS&T iConnect and the AutoCat and Code4lib lists. We will use the official email from the llm4cat project to accept submissions. We expect to select three to five submissions for presentations at the workshop, assessed by at least three jury members. The workshop organizers can secure at least two presentations based on their research practices and findings if needed. Each author will be given 10-15 minutes to present their work followed by 5 minutes of questions and answers.

The second part of the workshop will focus on the potential applicability of LLMs to cataloging, specifically. First, one of the workshop organizers, Jason Thomale, will present findings from informal interviews with catalogers from the University of North Texas Libraries, including the perceived need for using AI in cataloging tasks, application scenarios, and concerns about the impact of AI. Then attendees will be split into groups. Each group will include 5-8 individuals with diverse backgrounds and 1 workshop organizer for coordination and reporting. There will be two rounds of group discussions. The first will explore questions regarding the factors related to the adoption of LLMs for library cataloging – for example, perceptions about the performance of LLMs for cataloging tasks, the perceived risk of adopting LLMs in library cataloging, and the facilitating conditions for adopting LLMs. The second round of discussion will

seek participants' ideas about the ways LLMs can be integrated into the systems or cataloging workflows. For both discussions, we will provide papers and pens for attendees to draft their ideas and take notes.

| Agenda |
|---|
| Welcome, settle in, present the agenda for the day |
| Part I: Development of LLMs for Cataloging Tasks (1.5hr) <ul style="list-style-type: none">• Research Presentations (3-5 Presentations TBD), Q&A |
| Coffee Break |
| Part II: Applicability of LLMs in Cataloging (1.5hr) <ul style="list-style-type: none">• Presentation on Needs and Application Scenarios (Jason Thomale), Q&A• Group Discussion 1 on Adoption of AI, Report• Group Discussion 2 on Integration of AI into Cataloging Workflows, Report |

The workshop format will be interactive and participatory. The research presentations will bring the recent research development in the LLMs for information organization to a diverse audience, including practitioners. This will foster a deeper and more comprehensive understanding of the techniques, methods, and findings related to the topic. The follow-up discussions will spark ideas from researchers and practitioners with diverse backgrounds and areas of expertise. It will be an opportunity for attendees to gain different perspectives, build connections, and come up with solutions that could help make the application of LLMs in libraries a reality. Afterward, we will conduct a survey and collect feedback from the workshop participants to assess their overall satisfaction with the workshop and determine what new insights they may have gained. We will write a report based on the discussion results and submit it to the Journal of Information Matters.

The workshop will be highly relevant to the iConference, given its focus on information organization, library science, and the role of technology in managing information. The workshop will promote research and discussions on cutting-edge AI techniques in the Information Science field. The iConference emphasizes the importance of bridging the gap between academic research and practical implementations, and this workshop aims to integrate topics in theoretical model development and how these models can be transformed into solutions that enhance library and information systems. The workshop will also encourage discussions and bring awareness about the ethical, social, and policy implications of LLMs use in libraries, which also aligns with the core themes of the iConference.

REFERENCES

- Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., ... & McGrew, B. (2023). Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.
- Golub, K., Hagelbäck, J., & Ardö, A. (2020). Automatic Classification of Swedish Metadata Using Dewey Decimal Classification: A Comparison of Approaches. *J. Data Inf. Sci.*, 5(1), 18-38.
- Kragelj, M., & Kljajić Borštnar, M. (2021). Automatic classification of older electronic texts into the Universal Decimal Classification–UDC. *Journal of Documentation*, 77(3), 755-776.
- Marchionini, G. (2024). Information and library professionals' roles and responsibilities in an AI-augmented world. *Journal of the Association for Information Science and Technology*, 75(8), 865-868.
- Martorana, M., Kuhn, T., Stork, L., & van Ossenbruggen, J. (2024). Text classification of column headers with a controlled vocabulary: leveraging LLMs for metadata enrichment. *arXiv preprint arXiv:2403.00884*.
- Wang, J. (2009). An extensive study on automated Dewey Decimal Classification. *Journal of the American Society for Information Science and Technology*, 60(11), 2269-2286.

- Wang, L., Lyu, C., Ji, T., Zhang, Z., Yu, D., Shi, S., & Tu, Z. (2023, December). Document-Level Machine Translation with Large Language Models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (pp. 16646-16661).
- Wu, M., Brandhorst, H., Marinescu, M. C., Lopez, J. M., Hlava, M., & Busch, J. (2023). Automated metadata annotation: What is and is not possible with machine learning. *Data Intelligence*, 5(1), 122-138.
- Zhang, S., Wu, M., & Zhang, X. (2023). Utilising a Large Language Model to Annotate Subject Metadata: A Case Study in an Australian National Research Data Catalogue. *arXiv preprint arXiv:2310.11318*.