DCPAPERS

DCMI 2026

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

ISSN: 1939-1366

Are AI Models Getting Better at Cataloging? - Evidence from a Two-Point Comparative Study

Myung-Ja (MJ) K Han [1]

[1] University of Illinois Urbana-Champaign

Abstract

This study examines how the cataloging performance of four AI models, ChatGPT, Copilot, DeepSeek, and Gemini, evolved over eight months when tasked with extracting bibliographic information from scanned images across seven items of varying publication types and subject domains. Using four prompt variations and a consistent methodology established in an earlier round of testing, the second round revealed meaningful overall improvement in the accuracy and completeness of cataloging records, with models more consistently acknowledging missing information, providing inline justification for decisions, and exhibiting behaviors aligned with Explainable AI (XAI) and Retrieval-Augmented Generation (RAG) principles. Persistent challenges remained in controlled subject headings and URI accuracy, and a new concern emerged around balancing prompt over- and under-specification. These findings support a human-in-the-loop approach to AI-assisted cataloging and highlight the value of continued longitudinal monitoring.

1. Introduction

Research on artificial intelligence (AI) in library cataloging and metadata has grown rapidly since the early 2020s. Engel et al. (2025) identified approximately 14,000 publications on AI applications in library cataloging since ChatGPT's release in November 2022 alone, reflecting the field's explosive growth [1]. A systematic literature review found that subject heading assignment, metadata enhancement for digital repositories, and crosswalk generation between metadata schemas have emerged as the dominant topics at the intersection of AI and cataloging [2]. Despite this substantial body of experimentation and literature, a critical gap remains: longitudinal studies examining how the cataloging performance of widely used large language models (LLMs) changes over time as these models are updated and retrained are exceedingly rare. This study addresses that gap by assessing how four widely adopted AI models, ChatGPT, Copilot, DeepSeek, and Gemini, perform on cataloging tasks at two points in time, approximately eight months apart, to capture how performance shifts as these models evolve. Cataloging is a particularly demanding domain for AI evaluation, as it requires application of complex rules and standards such as RDA, MARC, authority control, and use of subject headings. This two-point comparative study offers insight not only into whether these models improve, but into how they are being trained to interpret and apply specialized professional knowledge. The findings are intended to provide libraries with an empirical foundation for making informed decisions about integrating AI tools into cataloging and metadata management workflows.

Poster

Available Aug 01, 2026

Correspondence to
Myung-Ja (MJ) K Han
mhan3@illinois.edu

2. Methodology

The methodology for this study follows the protocol established in the first test, whose findings were presented at the Internal Conference on Dublin Core and Metadata Applications 2025 [3]. Briefly, seven items representing various publication types and subject domains were selected, including seven items: four English-language monographs (a juvenile fiction, an adult fiction, a nonfiction, and a fiction with a non-standard font), one non-English language (Punjabi) monograph, a monographic series, and a serial. For each item, scanned images of the title page and, where necessary, additional pages such as the title page verso or table of contents, were provided as the basis for cataloging record creation.

The second test evaluated the same four AI models, ChatGPT, DeepSeek, Gemini, and Copilot, using the same four prompt variations (simple and detailed, each with and without an example output) applied to the same seven items. Notably, all four models had been updated within the eight-month interval between the two tests, making version change an inherent variable in this study. Generated outputs were evaluated against the same criteria of accuracy, completeness, and consistency, allowing direct comparison with the first test results.

3. Findings

The second test showed significant overall improvement compared to the first, particularly in the accuracy and completeness of generated cataloging records, with transcribed fields and capitalization showing especially notable gains. Several factors contributed. First, rather than fabricating information when data was unavailable, models generally acknowledged when information was unavailable, indicating "Not Available" or "N/A" for missing information, and in some cases noting the source from which information was drawn. This reflects the integration of Explainable AI (XAI) principles, representing a meaningful shift toward transparency. Second, when uncertain, models often provided inline reasoning rather than asserting values without qualification, for example noting a transcription decision such as ‘(appears on cover as ‘LYND WARD’) ", mirroring professional cataloging practice. Third, all models demonstrated the ability to capture information beyond what was explicitly requested, including physical description, contributor information, and copyright date. In some cases, models retrieved physical description data through web search even when no physical item was provided, consistent with the use of Retrieval-Augmented Generation (RAG).

Despite these improvements, significant challenges remain. All models continued to exhibit difficulty with assigning controlled subject terms correctly, and URI accuracy for author authority records was inconsistent, both of which are critical issues for linked data applications. Compliance with RDA formatting rules was also uneven across models, particularly for place of publication and title. Additionally, all models exhibited a range of errors in MARC data field usage. Finally, while the comprehensive prompt outperformed the simple prompt in the first round of testing, this advantage did not

persist in the second round. In some cases, the simple prompts yielded better results, consistent with evidence that LLMs can infer unspecified requirements from minimal prompts [4], though this inference is fragile. These findings suggest that future prompt design should carefully balance over- and under-specification to achieve more robust and consistent outputs.

4. Conclusion

These findings suggest that AI models are evolving in ways that are relevant to real-world cataloging workflows. The gains in transparency, source attribution, and contextual reasoning achieved within eight months are encouraging, particularly the emergence of XAI-informed behaviors such as explicitly flagging unavailable information and providing inline justification for cataloging decisions. These practices align with the accountability standards that professional cataloging demands. At the same time, persistent difficulties with controlled subject vocabularies and authority record URIs indicate that AI models are not yet ready to operate independently in cataloging workflows. Controlled vocabularies require deep familiarity with hierarchical relationships and evolving usage conventions, and errors in URIs carry significant consequences for metadata interoperability and discoverability in linked data environments.

For libraries considering AI integration, a human-in-the-loop model, where AI handles initial record generation and trained catalogers review and enhance the output, remains the most prudent near-term approach. Given the rapid pace of model development, continued monitoring of AI performance will be essential, and future studies should expand to broader languages, formats, and subject domains.

References

[1] J. Y. Engel, D. T. Do, B. Salem, and T. A. Cunningham, “Artificial intelligence in library cataloging: A review of literature,” *Journal of Library Metadata*, vol. 25, no. 4, pp. 261–276, 2025, doi: 10.1080/19386389.2025.2526913.

[2] D. Oyighan, E. S. Ukubeyinje, B. T. David-West, and B. D. Oladokun, “The role of AI in transforming metadata management: Insights on challenges, opportunities, and emerging trends,” *Asian Journal of Information Science and Technology*, vol. 14, no. 2, pp. 20–26, 2024, doi: 10.70112/ajist-2024.14.2.4277.

[3] M. K. Han, G. Heng, and P. Lampron, *Generative AI for bibliographic description: What works, what doesn't*. in International Conference on Dublin Core and Metadata Applications. 2025. [Online]. Available: https://www.dublincore.org/conferences/2025/sessions/posters/#recunnWsPRj2C8JgC

[4] C. Yang, Y. Shi, Q. Ma, M. X. Liu, C. Kästner, and T. Wu, “What prompts don't say: Understanding and managing underspecification in LLM prompts,” 2025.