DCPAPERS

Generative AI for Bibliographic Description: What Works, What Doesn't

Myung-Ja K. Han[1,*], Greta Heng[2,*,†] and Patricia Lampron[3,†]

[1] University of Illinois Urbana-Champaign, 1408 W. Gregory Dr. Urbana, IL, USA
[2] San Diego State University, 5500 Campanile Dr. San Diego, CA, USA
[3] University of California Irvine, 204 Aldrich Hall, Irvine, CA, USA

Abstract

As interest in applying artificial intelligence (AI) to cataloging and metadata creation grows, there remains a lack of comparative analysis on how current generative AI performs in real-world workflows. This project evaluates the practical capabilities of four freely available generative AI models in creating descriptive metadata, examining which aspects of bibliographic description they handle effectively and where they fall short. By analyzing each model's output and identifying strengths and limitations, the study offers guidance for catalogers seeking to integrate generative AI into their work. The findings aim to support informed decision-making, set realistic expectations, and contribute to ongoing discussions around automation, labor, and quality in metadata creation.

Keywords

AI, cataloging, metadata, ChatGPT, CoPilot, DeepSeek, Gemini

1. Introduction

There has been growing interest in applying artificial intelligence (AI) tools to cataloging and metadata creation. While recent discussions reflect both excitement and caution regarding AI integration, there remains limited comparative understanding of the currently existing generative AI models and how they function in real-world cataloging contexts. What kinds of generative AI models can support cataloging and metadata workflows? Which aspects of bibliographic description do they perform well in, and where do they fall short?

Although several studies have already evaluated ChatGPT in this area [1, 2, 3], this project seeks to assess the practical capabilities of other generative AI models in addition to ChatGPT for descriptive metadata creation. By comparing outputs and identifying strengths, limitations, and areas for improvement, we aim to offer catalogers clearer guidance on how generative AI can be thoughtfully and realistically incorporated into cataloging workflows. The findings are intended to support informed decision-making and contribute to broader conversations about automation, labor, and quality in metadata creation.

* Corresponding author.
† These authors contributed equally.
mhan3@illinois.edu (M. Han); gheng@sdsu.edu (G. Heng); plampron@uci.edu (P. Lampron)
0000-0001-5891-6466 (M. Han); 0000-0002-3606-6357 (G. Heng); 0000-0003-1070-5998 (P. Lampron)

https://doi.org/10.23106/dcmi.952585940

2. Methodology

For this study, we selected seven items representing various formats and subject domains, as these often contain different bibliographic features governed by distinct cataloging practices. Our test set included six English-language resources—a juvenile fiction monograph, an adult fiction monograph, a nonfiction monograph, a fiction monograph with a non-standard font, a monographic series, and a serial—as well as one non-English language monograph in Panjabi. For each item, we prepared a scanned image of the title page. When the title page lacked sufficient information (e.g., date of publication), we supplemented the input with additional scanned pages such as the title verso or table of contents.

We evaluated four widely used AI models, ChatGPT (version 4o), DeepSeek, Gemini, and Copilot. Each tool was tested with two types of prompts: one simple and one detailed, with and without example outputs. In total, we conducted 112 tests by running four versions of prompts with seven resources using four AI models. The generated outputs were evaluated using criteria focused on accuracy, completeness, and consistency.

3. Findings

Overall, the selected generative AI models performed well on relatively straightforward tasks such as generating a table of contents note, identifying language, and capturing publication dates. DeepSeek and Gemini attempted to distinguish between copyright date and publication date. For series information, which can be difficult to identify without understanding the concept, Gemini and Copilot were able to recognize and capture this data.

However, performance declined significantly for more complex tasks, particularly in generating subject headings using controlled vocabularies and constructing coordinated subject strings. All tools were able to generate lists of relevant keywords, but these were not drawn from established vocabularies. For entity URIs from authority sources, all tools provided incorrect URIs rather than acknowledging a lack of match, with the exception of DeepSeek for one specific heading.

Regarding prompt design, our findings show that more detailed prompts did improve performance compared to simpler ones in this experiment. Additionally, including a URL link to a guidance document within the prompt proved ineffective; the AI tools did not treat the link as a usable source. Instead, any essential instructions should be fully embedded within the prompt text.

4. Conclusion

This project evaluated four widely known and freely available generative AI models to assess their current capabilities in creating descriptive metadata. Echoing previous studies, our findings confirm that while the selected AI models perform reasonably well with straightforward bibliographic elements (e.g., language, date, and table of contents notes), it struggles with more complex tasks such as assigning subject headings and retrieving accurate URIs for entities. As other researchers have noted [1, 2, 3], the performance of generative AI in these tasks is inconsistent and often inaccurate. As Moulaison-Sandy (2024) emphasizes the importance of prompt design [4], our results also suggest that prompts must be detailed and

https://doi.org/10.23106/dcmi.952585940

self-contained for better outcomes; embedded links to external guidance are generally ignored by current AI models.

While there is both great optimism and valid concern about AI's role in cataloging and metadata work, our study reinforces the need for further refinement of generative AI models, particularly through training with library catalog and metadata-specific data models, controlled vocabularies, and standards. For generative AI to be a reliable partner and tool in metadata creation, it must be able to work effectively with established standards and authority frameworks. Libraries also need to set clear expectations regarding the level of cataloging quality that AI-generated metadata should meet.

References

[1] Shoichi Taniguchi."Creating and Evaluating MARC 21 Bibliographic Records Using ChatGPT." Cataloging & Classification Quarterly 62:5 (2024): 527–546.

[2] Jenny Bodenhamer. "The Reliability and Usability of ChatGPT for Library Metadata" SHAREOK Repository (2023). https://shareok.org/handle/11244/339626.

[3] Vyacheslav Zavalin and Oksana L Zavalina. "Are We There yet? Evaluation of AI-Generated Metadata for Online Information Resources." Information research 30. iConf (2025): 732–740.

[4] Heather Moulaison-Sandy and Zach Coble. "Leveraging AI in Cataloging: What Works, and Why?" Technical Services Quarterly, 41:4 (2024): 375–383.

https://doi.org/10.23106/dcmi.952585940