FEATURE

AI Chatbots and Subject Cataloging

A Performance Test

Brian Dobreski and Christopher Hastings

Libraries show an increasing interest in incorporating AI tools into their workflows, particularly easily accessible and free-to-use chatbots. However, empirical evidence is limited regarding the effectiveness of these tools to perform traditionally time-consuming subject cataloging tasks. In this study, researchers sought to assess the performance of AI tools in performing basic subject heading and classification number assignment. Using a well-established instructional cataloging text as a basis, researchers developed and administered a test designed to evaluate the effectiveness of three chatbots (ChatGPT, Gemini, Copilot) in assigning Dewey Decimal Classification, Library of Congress Classification, and Library of Congress Subject Heading terms and numbers. The quantity and quality of errors in chatbot responses were analyzed. Overall performance of these tools was poor, particularly for assigning classification numbers. Frequent sources of error included assigning overly broad numbers or numbers for incorrect topics. Although subject heading assignment was also poor, ChatGPT showed more promise here, backing up previous observations that chatbots may hold more immediate potential for this task. Although AI chatbots do not show promise in reducing time and effort associated with subject cataloging at this time, this may change in the future. For now, findings from this study offer caveats for catalogers already working with these tools and underscore the continuing importance of human expertise and oversight in cataloging.

As with many areas of practice, the cultural heritage domain has shown increasing interest in the use of AI in recent years, particularly in libraries. Gupta and Gupta noted this rise in interest, as well as the potential for libraries to experiment with their use in existing workflows for a variety of areas, including reference, collection management, and reader's advisory.[1] Practitioners in library cataloging spaces are also now demonstrating an interest in leveraging AI in their work. AI-based open source and vendor-backed tools aimed at catalogers and their workflows are beginning to emerge.[2] For now, though, currently available large language model (LLM)–based chatbots such as ChatGPT are appealing here because of their accessibility, their low barriers to use, and their capability to process and generate text information. A recent survey of academic libraries found that more than 50 percent of respondents reported using an AI chatbot in their cataloging work.[3] As Inamdar observed, AI tools hold great potential for metadata tasks, but significant concerns about quality and reliability of their output remain.[4] This has led to the emergence of some exploratory testing of AI tools' ability to perform library cataloging.[5]

One of the most challenging and time-consuming parts of producing library metadata may be subject cataloging: the analysis of a resource's "aboutness" and the assignment of corresponding subject

Brian Dobreski (bdobresk@utk.edu) is an Assistant Professor at the School of Information Sciences, University of Tennessee, Knoxville, https://orcid.org/0000-0002-2448-3495.
Christopher Hastings (chastin8@vols.utk.edu) is a Research Assistant at the School of Information Sciences, University of Tennessee, Knoxville, https://orcid.org/0009-0003-5829-497X.

https://doi.org/10.5860/lrts.69n1.8440

headings and classification numbers.[6] This task not only requires a cataloger to quickly comprehend an often complex resource, but also fluency in the formal and intricate systems used to represent subject and genre in library data. Systems such as Library of Congress Subject Headings (LCSH), Library of Congress Classification (LCC), and Dewey Decimal Classification (DDC) are widely used throughout libraries in the United States and elsewhere but come with high learning curves and typically require specialized education, training, and practice to achieve proficiency.[7] If effective in performing and supporting this kind of work, freely available AI tools such as chatbots hold the potential to reduce the high time and effort costs associated with subject cataloging. The potentials for AI subject cataloging remain relatively untested and underexplored, though the present study is aimed to address this gap.

In this article, researchers present the results of a performance test of three free AI chatbots' capabilities to conduct subject cataloging tasks. Specifically, using a well-established instructional cataloging text as a basis, researchers developed and administered a series of exercises to gauge the ability of OpenAI's ChatGPT, Google's Gemini, and Microsoft's Copilot to produce appropriate LCSH, LCC, and DDC headings and numbers. The goal of this study was to capture the state of AI subject cataloging at this moment and explore the current potentials for chatbots to complete common library subject cataloging work. The findings presented here add further empirical evidence into discussions concerning the quality and reliability of AI-performed metadata work. In addition, the authors present a documented and replicable test that can be used again to assess AI subject cataloging with future versions of these and other AI tools as this technology continues to develop.

Literature Review

The public premiere of AI chatbots such as ChatGPT piqued the interest of many throughout the international library community. Research into the applications and implications for libraries is just beginning to emerge, although it is likely to grow as libraries and their stakeholders are now showing greater awareness of AI and its possible roles in library work.[8] A review of the available literature shows much discussion of the potentials for AI, for example, as in Inamdar's exploration of the possible impacts of AI tools on library workflows, or Chhetri's SWOT (strengths, weaknesses, opportunities, and threats) analysis on AI and libraries.[9] The impact of AI on libraries' information literacy work is also prominent.[10] Actual case studies of AI implementation are, however, somewhat less prevalent. Rodriguez and Mune offer an interesting example with their overview of the development and deployment of an AI chatbot for reference services at San Jose State University.[11] Other recent cases of AI implementation in libraries show the range of work this technology is beginning to touch on. This includes search functions that recommend books based on statements rather than traditional searches, AI personalities imitating real-life figures to teach students, and translation of archival manuscripts.[12] That these examples vary so widely clearly demonstrates librarians' interest in adopting AI to facilitate all manner of their work.

Regarding library cataloging work, no well-documented case studies of integrating AI into active, existing workflows were available, although the community's desire to explore and experiment with this

practical application is clear.[13] A recent survey published by Primary Research Group sought to discover how prevalent AI use was in the workflow of catalogers at twenty-six universities. Of the universities that were contacted, two reported using Google's Bard AI (now Gemini), four reported using AI-enabled Bing (Microsoft Copilot), and thirteen reported using ChatGPT in their workflow.[14] It should be noted, however, that many respondents felt that these tools did not actually increase their productivity. Even so, catalogers' experimentation with AI in their daily work is likely to continue, despite warnings on the dangers of the unreliability of AI tools for such tasks.[15]

Such unreliability is apparent in the handful of documented tests of AI for cataloging work. Breeding prompted ChatGPT for MARC and BIBFRAME records for a specific book, and while the results looked convincing, closer inspection revealed significant fictitious or inaccurate information.[16] This is not surprising given the well-known tendency for AI chatbots to "hallucinate," that is, invent incorrect information.[17] Even so, Breeding felt that such tools, with the correct prompting and oversight, could still be of some use to catalogers.[18] Brzustowicz also put ChatGPT to the test in creating MARC records, finding the results more promising but also recommending that ChatGPT be used only in conjunction with cataloging professionals who could recognize and correct the mistakes.[19] It should be noted, however, that both the methodology and validity of this study has faced criticism from members of the cataloging community.[20]

Testing focused specifically on subject cataloging tasks is less well-documented. Of note is a 2023 study by researchers at Oklahoma State University looking at the reliability and usability of ChatGPT to harvest keywords, assign classification numbers, and choose LCSH terms.[21] This study was relatively small in scale, asking ChatGPT to create three DDC numbers and three LCSHs for a book about trade in ancient Rome and then asking ChatGPT the same questions three months later. The results were underwhelming: of six DDC numbers generated, only two were usable, with three incorrect and one that did not even exist. For subject heading work, however, ChatGPT proved more reliable, being able to generate valid LCSHs for all prompts.[22] In perhaps the most extensive subject cataloging experiment to date, Chow, Kao, and Li tested the ability of ChatGPT to provide subject headings in response to structured prompts containing titles and abstracts for thirty dissertations and theses; the authors noted the promise of these tools for reducing cataloging time but found validity issues that indicate the need for continued cataloger oversight.[23] These results begin to shed light on the potentials and pitfalls of AI chatbots for subject cataloging, although more robust testing and examination is required.

Methodology

In contrast to previous works, the researchers sought here to test multiple tools for multiple subject cataloging tasks, including subject heading and classification number assignment, through a replicable and well-documented methodology. To do so, the test designed for this study was derived from the second edition of Broughton's *Essential Classification*.[24] This monograph was designed as a beginner's text on subject cataloging, suited for graduate students studying library and information science. Earlier chapters of the book focus on the basics of subject analysis and representation, with subsequent

chapters focusing on the application of popular subject and classification systems. Specifically, chapters 12 and 13 cover the use and assignment of LCSH, chapters 15 and 16 cover the construction of LCC numbers, and chapters 17 and 18 cover the construction of DDC numbers. Within the text of these chapters, the reader is periodically presented with exercises to test their ability to construct and assign basic subject headings and classification numbers. These exercises are designed such that, with minimal prompting, a beginner-level student can assign appropriate headings and numbers to books bearing very descriptive titles, based on title, author, and publication information alone. In using these simple prompts as the basis for the current test, the researchers sought to emulate the basic, easily replicable questions a subject cataloger might face; the lack of further prompt engineering stands in contrast to previous work by Chow, Kao, and Li, the implications of which will be addressed further in the "Discussion" section.[25]

Researchers elected to focus the test solely on LCSH, LCC, and DDC due to the prominence of these particular systems in library cataloging. As such, they selected a sample of exercises across the six chapters identified above. In constructing this sample, researchers looked for exercises designed to yield complete subject headings or classification numbers, attempted to balance the number of questions on subject headings with those on classification, and avoided exercises on overly narrow or specific tasks (e.g., Cuttering names, using specific tables of limited applicability). Table 1 lists the exercises included in the test.

Table 1. Test exercises from Broughton's Essential Classification.

System	Exercises Included	No. of Questions
DDC	17.1, 17.2, 18.6	23
LCC	15.1, 15.4, 16.3	25
LCSH	12.1, 13.1, 13.3, 13.4, 13.5	50

All questions from each of these exercises were adapted with minor alterations into prompts and given to three LLM-based AI chatbots: ChatGPT, Gemini, and Copilot. These three tools were chosen for their overall prominence and the mention of their use by library catalogers in current literature.[26] Although premium versions of some of these tools are available, researchers only used the freely available version of each as of May 2024 (ChatGPT 3.5, Gemini 1.0, and Copilot build 2024.5), feeling this would better represent the tools available to all libraries regardless of budget considerations. During the test, each individual question from each exercise was presented as its own prompt, with as little modification as possible. For example, question 4 from exercise 13.4 was given as: "Construct a Library of Congress Subject Heading for the following title: *Chimpanzee: A Topical Bibliography*."[27] The resulting prompts were thus simple but easily replicable. In total, the same set of ninety-eight questions were asked of each of the three tools.

The test was conducted during May 2024. The entire text of each tool's response to each prompt was saved, totaling 294 responses. Responses were reviewed and compared with the answer key given in the Broughton text. If at least one subject heading or classification number provided by the tool matched the text's answer for a given question, this response was marked as correct (e.g., Chimpanzees – Bibliography). This was meant to reflect the text's requirement of only one heading in response to

each prompt, although this approach presents a limitation that will be addressed further below. In a limited number of cases, researchers judged that a nonmatching response was an acceptable alternative or close enough to count; the details of these situations are specific to each system and are described in the Results section. If no subject heading or classification number provided by the tool was found to be a match or otherwise acceptable, the response was marked as incorrect (e.g., Chimpanzees – Bibliography – Topical – Research). Finally, in some cases the tool returned a response stating it could not answer the prompt, leading researchers to mark the response as a refusal. Regardless of whether a subject heading or classification number was the expected answer or not, researchers attempted to validate it against the corresponding system. This allowed researchers to gauge if returned headings and numbers were, if not correct, at least valid in the sense that they existed and meant what the tool said they did. Checking of headings and numbers was performed using WebDewey, Classification Web, *The Classification and Shelflisting Manual*, *The Subject Headings Manual*, and OCLC's WorldShare Record Manager tool. The qualifications of the researchers to assess the results include previous professional experience as a cataloger, as well as more than ten years of experience teaching graduate cataloging and classification courses.

Results

Dewey Decimal Classification

The tools were given three DDC exercises totaling twenty-three questions. General performance across all three tools was poor, with the majority of responses deemed incorrect. Table 2 summarizes the results of the DDC exercises. To calculate the final grade, researchers included all correct and acceptable answers. ChatGPT was slightly more successful than the other two tools, but still only achieved a final score of 26 percent.

Table 2. AI performance on DDC exercises.

	ChatGPT	Copilot	Gemini
Correct	5	1	4
Acceptable alternative	1	1	0
Incorrect	17	21	19
Refusal	0	0	0
Final grade	26%	9%	17%
No. of valid DDCs	16	14	16
Percentage valid	70%	61%	70%

Although the majority of DDC numbers provided were not appropriate for the specified title, many of the provided numbers were at least valid DDC numbers (i.e., the number exists and means what the tool described it to mean). As shown in table 2, the tools ranged from 61 percent to 70 percent success in this regard. Table 3 provides further details on each tool's incorrect responses, including DDC numbers that were valid but still incorrect.

For all three tools, the most common error was providing a valid DDC number that was too broad, for example, assigning 720 to a book on cathedrals when 726.6 was the expected number. On the other hand, assigning a number that was too specific occurred much less frequently. Another common error was assigning a number for an incorrect topic altogether, for example, assigning a number on legal

offenses against the person (345.025) to a bibliography on capital punishment (016.36466). On several occasions, ChatGPT and Copilot returned numbers that do not exist and cannot be built using DDC tables. Finally, an error specific to DDC construction was the failure to follow number order guidance, including "first of two" order or preference order, which are to be followed when multiple numbers are possible.[28]

Library of Congress Classification

The three LCC exercises comprised a total of twenty-five questions. Again, overall performance across all three tools was poor, and especially so for Gemini, which only provided a correct answer to one question. Table 4 shows the tools' performances on the LCC exercises. To calculate the final grade for this set of questions, researchers included all correct, acceptable, and close answers. Here, close answers were considered any response where the classification number itself was correct while the author Cutter was incorrect. Overall, ChatGPT and Copilot performed slightly better than Gemini. It should also be noted that in two instances, Gemini refused to provide a response, claiming it did not have enough information to assist with the request.

Table 3. Nature of errors on DDC exercises.

	ChatGPT (*n* = 17)	**Copilot (*n* = 21)**	**Gemini (*n* = 19)**
Incorrect topic	5	6	8
DDC number does not exist	2	2	0
DDC number too general	7	12	8
DDC number too specific	1	0	1
Did not follow order rules	2	1	2

Table 4. AI performance on LCC exercises.

	ChatGPT	**Copilot**	**Gemini**
Correct	2	5	1
Close	2	0	0
Acceptable alternative	1	1	0
Incorrect	20	19	22
Refusal	0	0	2
Final grade	20%	24%	4%
Number of valid LCCs	13	13	3
Percentage valid	52%	52%	13%

Table 5. Nature of errors on LCC exercises.

	ChatGPT (*n* = 20)	**Copilot (*n* = 19)**	**Gemini (*n* = 22)**
Incorrect topic	6	12	11
LCC number does not exist	2	0	8
LCC number too general	6	5	1
LCC number too specific	2	0	0
Provided main class only	3	2	2
Provided number range only	1	0	0

In comparison with the results of the DDC exercises, the three tools were less successful in providing valid LCC numbers (see table 4). Whereas ChatGPT and Copilot provided a valid LCC 52 percent of the time, Gemini was only able to do so 13 percent of the time. Gemini was also far more likely to hallucinate nonexistent LCC numbers. This and other kinds of errors observed in the results are detailed in table 5.

Gemini provided nonexistent numbers in response to eight questions, whereas ChatGPT did so for two questions. The most common error across all three tools, however, was assigning a number for

an incorrect topic. For instance, when asked to classify a book on jobs in ancient Rome (HD4844), Gemini's response suggested it be classed with books on the history of Egypt at DT57. As with the DDC exercises, the tools were more likely to assign an existing LCC number that was too general or broad rather than one that was too specific. Assigning only the main class was another error encountered in all three tools' responses, for example, assigning simply "BV" as the classification number. In one instance, ChatGPT provided a range of numbers for a single book (TX724-TX727).

Library of Congress Subject Headings

Unlike in the classification number tests, the three tools tended to respond to LCSH questions with multiple possible subject headings. To address this, researchers chose one heading from each response to consider when calculating the test results. If any heading within a response matched the anticipated answer, this heading was chosen. When none of the possible headings matched, researchers chose the closest possible suggested heading, preferring the slightest variations in terminology or subdivision order. When a single closest match was not apparent, researchers chose the first or most prominently recommended heading from the response. The results of the LCSH test, summarized in table 6, are based on these best possible matches.

Table 6. AI performance on LCSH exercises.

	ChatGPT	Copilot	Gemini
Correct	21	4	11
Close	5	1	1
Acceptable alternative	1	0	1
Incorrect	23	45	33
Refusal	0	0	4
Final grade	54%	10%	26%
Number of valid LCSHs	38	18	24
Percentage valid	76%	36%	48%

Final grades on the LCSH test were calculated using all correct, close, and acceptable answers. For the LCSH exercises, an answer was considered close if cataloging software would correct the proposed heading in the course of normal authority control, for example, if a variant term was given rather than the preferred term. Researchers used OCLC WorldShare Record Manager to manually verify these headings were close enough to control to the correct heading automatically. Four refusals were noted, all occurring when using Gemini. Refusal responses simply stated the tool was "not programmed to assist with that" and gave no further information.

Overall, although Gemini and Copilot performance remained poor, ChatGPT showed more promise. Its final score of 54 percent was significantly higher than those of the other tools and was in fact the highest score observed by any tool on any of the tests. As shown in table 6, for 76 percent of the questions, ChatGPT was at least able to provide a valid LCSH heading (i.e., the terms existed, were combined correctly if applicable, and meant what they were described to mean). Gemini and Copilot performed worse here, with neither tool able to provide a valid LCSH even 50 percent of the time. While Gemini was most likely to hallucinate LCC numbers, Copilot was the most likely to hallucinate nonexistent LCSHs. Table 7 details these and other errors in the tools' incorrect responses.

Gemini hallucinated nonexistent LCSHs about as frequently as Copilot, whereas ChatGPT did so far less often. Frequent examples of LCSH hallucinations were fabricating subdivisions (Deserts – China – Periodicals – In Chinese) or using a book's title as part of the subject heading (The greatest weddings of all time [with illustrations] – Weddings – Pictorial works). ChatGPT's most common source of error was the omission of necessary subdivisions. For example, for a book with expected heading "Butterflies – Nomenclature," it instead simply provided "Butterflies" as the heading. This occurred in Copilot's responses as well, although it was not an issue for Gemini, which tended to add more, albeit incorrect, subdivisions to headings. Separate from this error was prescribing an overly broad LCSH (e.g., "Biogeochemistry" rather than "Sea-water – Iron content"). Interestingly, ChatGPT and Gemini were more likely to provide overly narrow headings rather than overly broad ones. For instance, for the above-mentioned book on butterflies, Gemini suggested "Butterflies – Nomenclature – History," a level of specificity that was not warranted from the title. Specific to the LCSH tests was an error concerning geographic subdivisions, where a response did not correctly invert a place name used as a subdivision. Finally, suggested LCSHs with completely incorrect topics were relatively rare.

Table 7. Nature of errors on LCSH exercises.

	ChatGPT (*n* = 23)	**Copilot (*n* = 45)**	**Gemini (*n* = 33)**
Incorrect topic	1	1	2
LCSH does not exist	5	26	19
LCSH too broad	2	2	3
LCSH too narrow	3	1	8
LCSH lacking subdivision	11	15	0
Incorrect geographic subdivision	1	0	1

Given the fact that most responses contained multiple possible subject headings for a given book, researchers broadened their examination to include consideration of whether all of these headings were at least valid LCSHs. As shown in table 8, for any given prompt, ChatGPT responded with an average of six possible headings. Copilot and Gemini typically responded with comparatively less, with averages of 2.5 and 2.9, respectively.

Table 8. Overview of all LCSH provided in responses.

	ChatGPT	**Copilot**	**Gemini**
Total LCSHs provided	298	125	133
Total valid LCSHs provided	189	65	64
Percentage valid	63%	52%	48%
Average number of LCSHs provided per book	6.0	2.5	2.9
Average number of valid LCSHs provided per book	3.8	1.3	1.4

Across 50 LCSH questions, ChatGPT provided a total of 298 headings, Copilot 125 headings, and Gemini 133 headings. Reviewing all of these headings, researchers determined that 63 percent of all headings provided by ChatGPT were valid LCSHs. Copilot performed slightly worse here, with 52 percent of suggested headings valid, as did Gemini with 48 percent. Thus, although ChatGPT provided an average of six LCSHs per book, only 3.8 were valid. Gemini provided an average of 2.9 headings per book with 1.4 being valid, and Copilot provided an average of 2.5 headings per book with 1.3 being valid.

Discussion

All three tools performed inadequately on the classification number tests. In their responses, the suggested DDC numbers tended to be overly broad or general, or for different topics entirely. In the latter case, some responses included reasoned, although perhaps not compelling, explanations for the chosen number. For a title on the topic of bush walking, for example, Copilot assigned a DDC number for hazardous materials, explaining that the terrain and possible presence of camping fuels would make this activity dangerous. Such explanations, particularly for less obvious cases, could be persuasive and would certainly require the user to doubt and double-check the response to determine that it was incorrect. In LCC assignment, the tools were prone to the same kinds of errors, with many overly broad numbers or numbers better suited for other topics. In several instances, the tools seemed to "learn" an LCC number and attempt to reuse it. The most noticeable case occurred with Copilot, which used HE355 (traffic engineering) for three separate titles on railroads, masonry, and investment planning, respectively. Further consideration of prompt feedback within a session is addressed below. Hallucinated nonexistent classification numbers were a larger concern in the LCC exercises, particularly for Gemini. The DDC responses contained fewer such cases, although due to the differing structures of the two systems, most simple number combinations are likely to mean something in DDC, unlike in LCC, where many gaps and unassigned ranges exist.

Although performance in assigning subject headings was similarly disappointing, ChatGPT was noticeably better at this task than was observed in any other tool/task combination during the test. ChatGPT was able to suggest LCSH headings that were valid 63 percent of the time, far more impressive than Gemini or Copilot, which both stood closer to 50 percent. It should also be noted that ChatGPT provided on average more possible headings for each prompt as well, something users might find helpful because, unlike with classification numbers, most resources will receive multiple headings during subject cataloging. Even so, at a final score of 54 percent, ChatGPT was unable to muster a passing grade on subject heading assignment. Suggested headings from ChatGPT were often too general, and although additional prompting may have yielded a narrower, more subdivided version, follow-up prompting was not included in the present test. Many of the headings provided by Copilot exhibited the same kind of error, which tended to assign overly broad headings without any subdivision. Interestingly, in some instances Copilot took a more faceted approach in its subject heading construction, providing separate terms for concept, place, and form, but refraining from combining them. This suggests Copilot might be more successful in assigning terms from a post-coordinate system, such as Faceted Application of Subject Terminology (FAST), instead. In contrast, Gemini provided too many subdivisions, many of which were hallucinated and placed within square brackets like a qualifier (e.g., "Choctaw legends – [Thunderstorms]"). As such, while Gemini performed slightly better than Copilot overall, it was more likely to provide invalid headings in its responses.

The overall findings of the tests back up Bodenhamer's previous observation that ChatGPT showed more immediate potential for assigning LCSH than it did for classification numbers.[29] Indeed, test results here show that none of the tools performed well on DDC or LCC assignment, and while Gemini and Copilot struggled with LCSH as well, ChatGPT showed some promise. Even so, the free version

of ChatGPT tested here could not be relied upon to provide adequate subject headings for a resource without some cataloger oversight. Although test results do not indicate the use of either of the other tools for subject cataloging at this time, it should be noted that Copilot provided much more additional information in its responses, including websites and other resources to help the user in choosing numbers of headings. This kind of assistance could be useful to catalogers but was not analyzed in the present study and should be followed up on separately. It must, of course, be noted that these tools are likely to continue to develop and improve in coming years and may be more useful in subject cataloging tasks in the future. There is already some evidence to suggest that paid, premium versions of AI tools are capable of performing better metadata work, although this remains to be evaluated.[30]

Based on the present findings, AI chatbots are more likely to be of use in subject heading assignment before other areas of subject cataloging, and it is likely that, in time, AI chatbots will continue to perform better on all the tests run in this study. As such, the broader takeaway going forward may lie less in the quantity of errors observed here than in the quality. As shown above, simple subject cataloging prompts to these tools often return classification numbers and subject headings that are either too broad or are for the wrong topic. Catalogers currently working with these tools would do well to check any suggested number or heading to see: (1) that it exists, (2) that it means what the tool claims it does, and (3) that nothing narrower (i.e., more specific number, more subdivided heading) is more appropriate. AI chatbots may even provide a good starting point for subject cataloging, but their human users currently need a firm understanding of DDC, LCC, or LCSH to be able to effectively assess and adjust any provided suggestions using these systems. This makes it doubtful that these tools can do much to reduce the time and training needed for subject cataloging, at least for now.

As these tools continue to develop, so too does user understanding of how to employ them more effectively. Although this study used only simple, single prompts, other works have explored the use of more robust prompt engineering.[31] In focusing on simplicity and replicability, the present study is limited in its omission of these more sophisticated prompts. Similarly, some immediately obvious errors in chatbot performance such as including the title as a subject heading or repeating a previously used classification number could have been quickly addressed with a second feedback prompt. Although catalogers using chatbots will likely find more effective ways of interacting with them in the coming years, it should be stated that if users would need to spend significant time prompting, re-prompting, and verifying results, it would likely be more efficient in many cases to just perform subject cataloging themselves. Still, these tools may hold potential to assist in subject cataloging, even if they cannot independently complete the tasks. Going forward, the knowledge needed to understand and assess chatbot cataloging should be included in emerging conversations about the kind of AI literacy needed by information professionals.

This study was not without a number of other limitations that must also be noted. The Broughton text served as a wonderful resource, providing a number of questions designed to be answered by a human cataloger with very little contextual information.[32] Even so, it is likely these exercises were meant as more formative assessments rather than summative, with students learning and improving throughout

rather than being graded for overall performance. Also, this test was not designed to measure intra-indexer consistency, that is, how consistent a tool is at providing the same number or heading to the same resource, although this may be worth examination in future study. It was also unlikely to reflect how real users would be interacting with a chatbot, which would likely feature more back-and-forth dialogue. Adjusting or changing a prompt on the fly may have led the tool to a more accurate response in any given instance, particularly those with very obvious errors (e.g., title used as a subject heading). Follow-up user studies with practicing catalogers would provide more insight into how chatbot interactions are playing out in real working settings. Although validity was assessed for all provided headings and numbers, choosing only the best or closest to compare to the answer key may present an overly optimistic view of how these tools fully perform. Finally, as explained above, while premium versions of these tools may have performed better, they require costs that not all libraries are able to pay, particularly in support of what is already resource-intensive work.

Future work could and should address these limitations. In addition, a number of other opportunities exist for further study. This test could be repeated with future versions of these tools to gauge their improvement in performing subject cataloging work. The test could be modified to include some level of prompt engineering focused on instructing the chatbot to follow specific cataloging and classification rules, provide a certain number of headings, or take into account additional information such as summaries or tables of contents. This might better reflect *in situ* cataloger interactions with these tools. In addition, testing other systems for subject cataloging would be of use, particularly for specialized libraries and libraries outside of the United States. Using a simpler system, such as the FAST vocabulary, or a more domain-specific system, such as the Medical Subject Headings (MeSH), is another worthwhile direction for research. Comparative testing may also yield useful results. Testing free versions versus premium versions may help libraries decide whether the investment in a paid tool would really be justifiable. Finally, testing could be useful to compare performance among three different groups: AI chatbots, beginning catalogers, and beginning catalogers paired with chatbots. Results of such a study may yield more actionable results for libraries interested in incorporating AI tools into their existing workflows.

Conclusion

Working from a well-established cataloging text, researchers tested three free and commonly used AI chatbots on a series of subject cataloging tasks, finding none of these tools currently adequate in their ability to assign DDC, LCC, or LCSH. These results add further empirical evidence into ongoing conversations about AI and library work and offer a starting point for continuing observation of the development of AI cataloging. Of particular interest are the kinds of errors observed during this study, which provide both caveats for catalogers already working with these tools, as well as indications of the kinds of knowledge still needed by library staff moving forward. Part of the challenge in subject cataloging is, after all, its subjectivity and the lack of any real-life answer key. Fluency in subject cataloging systems remains critical. All of this underscores the continued importance of human labor in subject cataloging work. In the future, AI tools may prove more valuable in assisting catalogers,

especially in subject heading assignment, but continued testing and assessment will be needed to demonstrate this. The present study suggests a number of promising directions for future study, including the repetition of the Broughton test on future versions of these and other AI tools as a means of tracking progress and comparing performance. Careful, ongoing assessment is required to support the responsible incorporation of AI into libraries, not only in subject cataloging, but in all areas of information work.

Notes

1. Varun Gupta and Chetna Gupta, "Leveraging AI Technologies in Libraries through Experimentation-Driven Frameworks," *Internet Reference Services Quarterly* 27, no. 4 (October 2, 2023): 211–22, https://doi.org/10.1080/10875301.2023.2240773.
2. Koraljka Golub et al., "Automated Dewey Decimal Classification of Swedish Library Metadata Using Annif Software," *Journal of Documentation* 80, no. 5 (2024); 1057–70; Ex Libris, "The AI Metadata Assistant in the Metadata Editor," Ex Libris Knowledge Center, November 13, 2024, https://knowledge.exlibrisgroup.com/Alma/Product_Documentation/010Alma_Online_Help_(English)/Metadata_Management/005Introduction_to_Metadata_Management/The_AI_Metadata_Assistant_in_the_Metadata_Editor.
3. *Survey of Use of Bard, Bing & ChatGPT for Academic Library Cataloging* (Primary Research Group, 2023), https://books.google.com/books?id=R7ke0AEACAAJ.
4. Saba Inamdar, "Impact of Artificial Intelligence Text Generators (AITGs) on Libraries," *Library Hi Tech News* 40, no. 8 (January 1, 2023): 9–13, https://doi.org/10.1108/LHTN-03-2023-0048.
5. Marshall Breeding, "The Systems Librarian" Computers in Libraries, 43, no. 4 (May 2023): 17–19, https://www.proquest.com/trade-journals/systems-librarian/docview/2809560722/se-2?accountid=15172; Richard Brzustowicz, "From ChatGPT to CatGPT: The Implications of Artificial Intelligence on Library Cataloging," *Information Technology and Libraries* 42, no. 3 (2023): 1–22, https://doi.org/10.5860/ital.v42i3.16295.
6. Brian Dobreski, "Descriptive Cataloging: The History and Practice of Describing Library Resources," *Cataloging & Classification Quarterly* 59, no. 2–3 (April 13, 2021): 225–41, https://doi.org/10.1080/01639374.2020.1864693; Ralph M. Holley and Daniel N. Joudrey, "Aboutness and Conceptual Analysis: A Review," *Cataloging & Classification Quarterly* 59, no. 2–3 (April 13, 2021): 159–85, https://doi.org/10.1080/01639374.2020.1856992.
7. Arlene G. Taylor and Daniel N. Joudrey, "On Teaching Subject Cataloging," *Cataloging & Classification Quarterly* 34, no. 1–2 (September 1, 2002): 221–30, https://doi.org/10.1300/J104v34n01_13.
8. Dessy Harisanty et al., "Leaders, Practitioners and Scientists' Awareness of Artificial Intelligence in Libraries: A Pilot Study," *Library Hi Tech* 42 (April 4, 2022), https://doi.org/10.1108/LHT-10-2021-0356.
9. Saba Inamdar, "Impact of Artificial Intelligence Text Generators (AITGs) on Libraries," *Library Hi Tech News* (May 4, 2023), https://doi.org/10.1108/LHTN-03-2023-0048; Parbat Chhetri, "Analyzing the Strengths, Weaknesses, Opportunities, and Threats of AI in Libraries," *Library Philosophy and Practice*, 2023, 1–14.
10. Amy B. James and Ellen Hampton Filgo, "Where Does ChatGPT Fit into the Framework for Information Literacy? The Possibilities and Problems of AI in Library Instruction," *College & Research Libraries News* 84, no. 9 (October 2023), 334, https://doi.org/10.5860/crln.84.9.334; Sanjay Kumar Jha, "Application of

Artificial Intelligence in Libraries and Information Centers Services: Prospects and Challenges," *Library Hi Tech News* 40, no. 7 (January 1, 2023): 1–5, https://doi.org/10.1108/LHTN-06-2023-0102.

11. Sharesly Rodriguez and Christina Mune, "Uncoding Library Chatbots: Deploying a New Virtual Reference Tool at the San Jose State University Library," *Reference Services Review* 50, no. 3/4 (January 1, 2022): 392–405, https://doi.org/10.1108/RSR-05-2022-0020.
12. "Yokohama Library System Introduces Japan-1st AI Book Search," *Mainichi Daily News*, January 16, 2024, https://mainichi.jp/english/articles/20240116/p2a/00m/0na/006000c; John Emeigh, "Butte Library Using Artificial Intelligence to Educate and Create," *KXLF*, July 25, 2023, https://www.kxlf.com/news/local-news/butte-library-using-artificial-intelligence-to-educate-and-create; "A.I. Uncovers Unknown Play by Spanish Great in Library Archive," *Reuters*, January 31, 2023, sec. Oddly Enough, https://www.reuters.com/lifestyle/oddly-enough/ai-uncovers-unknown-play-by-spanish-great-library-archive-2023-01-31/.
13. Caroline Saccucci and Athena Salaba, "Introduction to Artificial Intelligence (AI) and Automated Processes for Subject Access," *Cataloging & Classification Quarterly* 59, no. 8 (December 21, 2021): 699–701, https://doi.org/10.1080/01639374.2021.2022058.
14. *Survey of Use of Bard, Bing & ChatGPT for Academic Library Cataloging*.
15. Paul R. Pival, "How to Incorporate Artificial Intelligence (AI) into Your Library Workflow," *Library Hi Tech News* 40, no. 7 (January 1, 2023): 15–17, https://doi.org/10.1108/LHTN-03-2023-0052.
16. Breeding, "AI: Potential Benefits."
17. Alex Woodie, "Hallucinations, Plagiarism, and ChatGPT," *Datanami*, January 18, 2023, https://www.datanami.com/2023/01/17/hallucinations-plagiarism-and-chatgpt/; Alex P. Watson, "Hallucinated Citation Analysis: Delving into Student-Submitted AI-Generated Sources at the University of Mississippi," *The Serials Librarian* (December 2024): 1–9, https://doi.org/10.1080/0361526X.2024.2433640; Mikaël Chelli et al., "Hallucination Rates and Reference Accuracy of ChatGPT and Bard for Systematic Reviews: Comparative Analysis," *Journal of Medical Internet Research* 26 (2024): e53164, https://doi.org/10.2196/53164.
18. Breeding, "AI: Potential Benefits."
19. Brzustowicz, "From ChatGPT to CatGPT."
20. Christine DeZelar-Tiedman, "Response to 'From ChatGPT to CatGPT'," *Information Technology and Libraries* 42, no. 4 (2023); Tess Amram, Robin Goodfellow (Puck) Malamud, and Cheryl Hollingsworth, "Response to 'From ChatGPT to CatGPT'," *Information Technology and Libraries* 42, no. 4 (2023); David Floyd, "Response to 'From ChatGPT to CatGPT'," *Information Technology and Libraries* 42, no. 4 (2023).
21. Jenny Bodenhamer, "The Reliability and Usability of ChatGPT for Library Metadata," (2023), https://openresearch.okstate.edu/entities/publication/98b121d2-1f87-4824-b5c9-d204dfe87ced.
22. Bodenhamer, "The Reliability and Usability."
23. Eric H. C. Chow, T. J. Kao, and Xiaoli Li, "An Experiment with the Use of ChatGPT for LCSH Subject Assignment on Electronic Theses and Dissertations," *Cataloging & Classification Quarterly* 62, no. 5 (2024): 574–88.
24. Vanda Broughton, *Essential Classification* (Facet Publishing, 2015).
25. Chow et al., "An Experiment."

26. *Survey of Use of Bard*.
27. Broughton, *Essential Classification*, 142.
28. "Introduction to the Dewey Decimal Classification" OCLC, https://www.oclc.org/content/dam/oclc/dewey/versions/print/intro.pdf.
29. Bodenhamer, "The Reliability and Usability."
30. Pival, "How to Incorporate."
31. Chow et al., "An Experiment."
32. Broughton, *Essential Classification*.