Chris Holstrom, University of Washington

Large Language Models (LLMs) and Cataloging: Exploring How ChatGPT and Copilot Assign Subject Headings and Call Numbers

Abstract

Large Language Models (LLMs) have demonstrated some facility in language- and knowledge-intensive tasks that require domain knowledge, such as writing and computer programming. These syntactic facilities suggest that general purpose LLMs might be able to perform subject cataloging tasks like assigning subject headings and class numbers. This paper investigates how two commercially available LLMs (ChatGPT and Copilot) assign subject headings using the Library of Congress Subject Headings (LCSH) and the Sears List of Subject Headings, class numbers using Library of Congress Classification (LCC) and Dewey Decimal Classification (DDC), item numbers using Cutter numbers, and MARC fields for subject headings and call numbers. The paper finds that the LLMs show promise as automated catalogers, but exhibit numerous shortcomings, including: lacking specificity, using unauthorized terms, incorrectly assembling synthetic headings, assigning inaccurate headings and classes, and formatting MARC records incorrectly. Based on these findings, potential cataloging applications for current LLMs are primarily as aides and teaching tools, not as fully automated cataloging solutions. Additionally, collections considering LLMs for cataloging tasks should be aware of issues associated with these technologies, including environmental harm, de-skilling, intellectual theft, and bias.

1. Introduction

"If you need more precise cataloging details, you might want to check the official Library of Congress catalog or a professional cataloging resource." - Copilot

Knowledge organization systems (KOSs) such as thesauri, subject heading lists, classification schemes, and taxonomies are used to represent, to collocate, and to support retrieval of information, including library resources. Each KOS comprises a complex set of rules, vocabularies, syntax, and notation systems. For example, building the Library of Congress subject heading "Water resources development—Law and legislation—India—19th century" requires understanding the aboutness of a resource, translating that aboutness to authorized subject headings, and constructing, in this case, a lengthy, specific, and synthetic subject heading to represent the resource. Furthermore, the cataloger will likely assign multiple subject headings to represent the resource and—using complex notation systems—assign a class and call number for the resource. Finally, catalogers integrate these subject headings and class numbers with additional resource descriptions in highly formatted surrogates like MARC records. Catalogers undertake extended study and pursue ongoing practical experience to develop expertise with these KOSs so that they can catalog library resources accurately (Ranganathan, 1937), ethically (Bair, 2005; Alder and Tennis, 2013; Watson 2021), and with appropriate specificity (Cutter, 1904; Svenonius, 2000) and exhaustivity (Foskett, 1996; Lancaster, 2003).

These trained and experienced catalogers, while not perfect, represent the gold standard for working with complex KOSs to catalog resources; however, libraries and other collections in search of cost-saving efficiencies have pursued automated and semi-automated systems for assigning metadata to resources (Golub, 2021). Automated

systems have used techniques such as string matching (Kazi et al., 2021), Natural Language Processing (NLP) (Rogers et al., 2020), and supervised Machine Learning (Ghiassi et al., 2012; Cervantes et al. 2020). Recent research has begun to explore Large Language Models (LLMs) and generative AI for cataloging tasks (Brzustowicz, 2023; Moulaison-Sandy and Coble, 2024; Chow et al., 2024; Dobreski and Hastings, 2025), and this paper aims to contribute to that conversation.

Widely available commercial LLMs—while criticized for their environmental cost, for de-skilling human practitioners, for ethics and bias concerns, and for and "parroting" existing texts (Bender et al., 2021)—have demonstrated some facility in language- and knowledge-intensive tasks that require domain knowledge (Petroni et al., 2021), as human-like writers (Comez-Rodriguez and Williams, 2023), and writing "assistants" (Bibi and Atta, 2024). For example, LLMs like ChatGPT (OpenAI, 2025) and Copilot (Microsoft, 2025) can generate grammatically correct, albeit flaccid, prose in response to prompts from human users (Lozić and Štular, 2023; Hu et al., 2024). LLMs can also produce code that follows the complex syntax of programming languages (Coello et al., 2024). These syntactic facilities suggest that general purpose LLMs might also be able to identify patterns and generate language that follows the syntax and complex rules for indexing languages and controlled vocabularies.

Based on the promise that LLMs have shown, this paper asks: Can commercial LLMs—which are not designed specifically to use complex KOSs—"learn" the patterns and rules for assigning subject headings, classes, and call numbers for library resources? Can we characterize, classify, and understand the reasons that LLMs succeed and fail in these tasks? And what can LLM's responses tell us about these systems, how to prompt them, and how they might or might not be useful tools for subject cataloging now and in the future?

2. Literature Review

As stated above, this study aims to contribute to the emerging conversation around LLMs and generative AI for cataloging tasks. This section reviews some of these studies and identifies how the current study differs from and complements these studies.

Brzustowicz (2023) studies ChatGPT and its ability to "generate accurate MARC records using RDA and other standards such as the Dublin Core Metadata Element Set." While Brzustowicz raises issues related to intellectual property rights and bias, he is largely enthusiastic about the performance of ChatGPT in cataloging tasks, noting the similarity between WorldCat MARC records and MARC records generated by ChatGPT. His enthusiasm be based more on the novelty of the emergent technology than the quality of the results it produced because close inspection shows significant differences between the WorldCat and ChatGPT MARC records. For example, ChatGPT omitted key MARC fields, repeated values, did not follow name authorities, and generally produced lower quality metadata than human catalogers in the MARC records provided with the study. A group of catalogers and librarians wrote to the publishing journal, Information Technology and Libraries, to criticize the article and its findings (Amram et al., 2023). The current study differs from the Brzustowicz study because it focuses on subject cataloging and not other forms of metadata and because it considers two LLMs (ChatGPT and Copilot) and analyzes responses to a larger number and variety of resources.

Moulaison-Sandy and Coble (2024) are more critical of ChatGPT and generative AI systems, emphasizing the limitations and warning that these technologies are not an easy solution to cataloging problems. Moulaison-Sandy and Coble provide a broader introduction to the strengths and weaknesses of generative AI and then focus on how these systems likely train on the millions of MARC records and authority files available from the Library of Congress and other data sets. The find that "In general, ChatGPT can be prompted to generate largely well-formatted catalog records, but these still need to be checked for errors." They then discuss the deeper challenge of assigning subject headings—a primary focus of the current study—and suggest that plugins that provide LLMs with additional fine-tuning and training data can help with specialized tasks. The current study does not attempt to build or test an LLM plugin, but it heeds Moulaison-Sandy and Coble's caution that the rules and nuances of cataloging are not easily learned by LLMs like ChatGPT.

Chow et al. (2025) test ChatGPT's ability to generate LCSH subject headings for electronic theses and dissertations, choosing these texts because universities must perform time-consuming cataloging work for these texts and because LLMs likely have not been trained on existing catalog records for these texts. They find that ChatGPT succeeded at correctly formatting MARC records 90% of the time, but that ChatGPT produced valid LCSH headings for only 23.3% of records and largely failed to generate adequately specific subject headings for the texts, with only 53.3% of generated records having adequate specificity and exhaustiveness. The current study follows a similar tack by testing LLMs abilities to assign subject headings and class numbers for resources that lack existing catalog entries because they do not exist and online articles that, like electronic dissertations and theses are available in full text but likely not previously cataloged.

In their study, Dobreski and Hastings (2025) conduct a "performance test" of three LLMs: ChatGPT, Copilot, and Gemini, analyzing their performance in subject cataloging tasks. They find that all three LLMs perform poorly, noting in particular issues with assigning overly broad or incorrect class numbers. The Dobreski and Hastings study is most similar to the current study and, it must be admitted, beat the current study to press, as it was published while the current study was underway. That said, the current study can serve as a complement to the Dobreski and Hastings study because it provides additional data and analysis of how LLMs perform (and fail to perform) subject cataloging tasks and because it differs in design by considering the Sears List of Subject Headings, MARC records and item number, and, notably, tasks LLMs with cataloging fabricated resources and resources for which the full text is provided.

3. Methods

To evaluate the subject cataloging effectiveness of the popular LLMs, ChatGPT and Copilot, each LLM was prompted to catalog a series of ten resources from each of four categories: 1) well-known resources, 2) more obscure resources with less metadata on the web, 3) completely made-up resources, and 4) full-text articles. These categories were chosen in an attempt to isolate the types of information (i.e. title, author, full text, already assigned metadata, external resources) that these opaque LLMs used to catalog resources and to identify whether the LLMs simply copied existing cataloging work

performed by humans. (See the appendix for lists of all the real and fabricated resources used in the study.) The study used the free and publicly available version of each LLM, with no additional training or features. ChatGPT was accessed at https://chatgpt.com and Copilot was accessed at https://copilot.microsoft.com.

For each of ten resources from each category, ChatGPT and Copilot were prompted to assign subject headings using the Library of Congress Subject Headings (LCSH) and the Sears List of Subject Headings, class numbers using Library of Congress Classification (LCC) and Dewey Decimal Classification (DDC), item numbers using Cutter numbers, and MARC fields for subject headings and call numbers. The prompts for LCSH and Sears List headings were similar to the following: "You are an experienced library cataloger. Assign Library of Congress subject headings for the book 'Algorithms of Oppression' by Safiya Noble." The prompts for the LCC and the DDC were similar to the following: "You are an experienced library cataloger. Name a valid LCC class and assign a LCC call number for the book 'Algorithms of Oppression' by Safiya Noble. Use Cutter numbers for the item number and use MARC to represent the full call number." (See the appendix for the templates that were used for the prompts.)

ChatGPT and Copilot were prompted four times each for each of the 40 resources in the study, resulting in 320 combined total responses from the two LLMs. Each response was analyzed for 1) how well it conformed to the syntax and grammar of the LCSH, the Sears List, the LCC, the DDC, the Cutter Table, and MARC; 2) how well it met core cataloging goals of accuracy, specificity, exhaustivity; 3) how well it addressed ethical concerns; and 4) how closely it aligned with experienced catalogers (when possible). Each part of each response (i.e. each subject heading in the list provided by the LLM and each tag, indicator, and subfield in the MARC record) were scored as correct or coded in an emergent codebook. Responses for resources that were cataloged in the United States Library of Congress were compared to those catalog records. Resources that had not been cataloged were evaluated based on the author's judgment. These results were compiled for all four categories of resources, and the results are discussed below.

4. Results and Discussion

This section presents aggregate analysis and discussion of the 320 prompts and responses in the study. To contextualize the aggregate analysis, the section first discusses some representative prompts and responses, as shown in Figures 1 and 2. Note that both examples discussed in detail are from Copilot because ChatGPT's response were too verbose to easily show in an image.

In the prompt and response shown in Figure 1, Copilot assigned eight subject headings, all of which appear somewhat relevant to the text at first glance. This surface appearance of correctness and exhaustivity is also a trait of LLMs for non-cataloging tasks (Ji et al., 2023). However, closer examination shows that two of the suggested headings are not authorized LC subject headings: "Discrimination in information technology" and "Racism in technology," one of the headings is inaccurate: "Digital divide," and some lack specificity or are redundant (e.g. "Information technology–Social aspects" is unnecessary, especially with the more specific heading "Algorithms–Social aspects"). The unauthorized headings in particular suggest that Copilot struggled with the complex and inconsistent patterns for forming pre-coordinate headings and that

Copilot was not referencing (or not strictly following) the authorized headings in the vocabulary. Following this pattern of struggling with pre-coordinate headings, Copilot arguably overused the "Social aspects" subdivision by featuring it in four synthetic headings. Furthermore, Copilot was not consistent with its responses, as it assigned another unauthorized pre-coordinate heading, "Bias in computer algorithms," in one of its other responses to the same prompt. In contrast, the LC Catalog provides a more succinct, accurate, and authorized set of headings: "Google," "Search engines–Sociological aspects," and “Discrimination.”

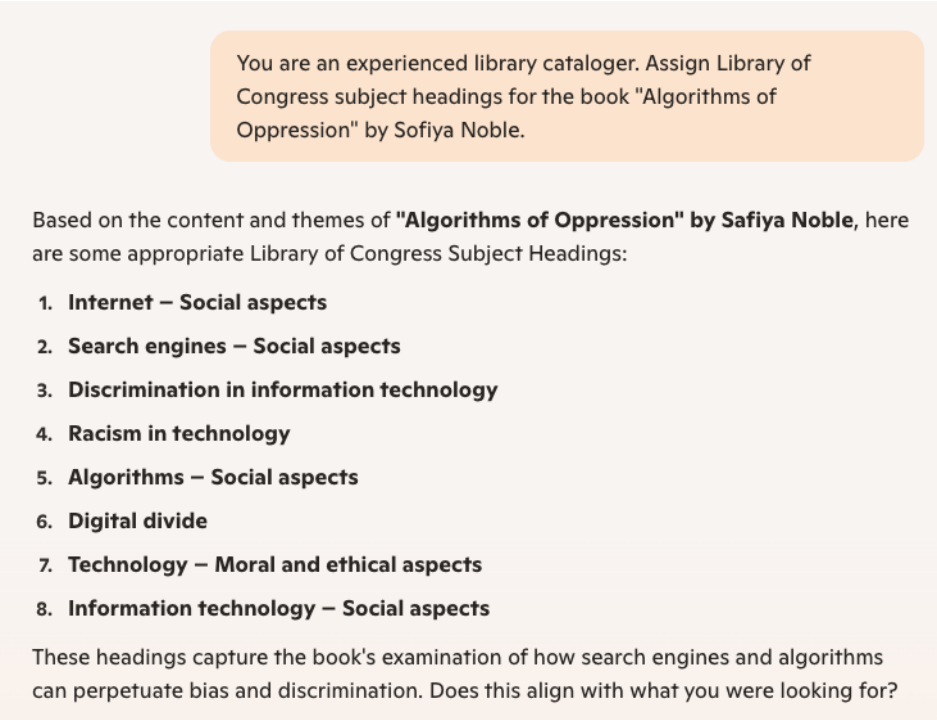

Figure 1. A prompt and Copilot's response as it assigns LC subject headings to the book *Algorithms of Oppression* by Safiya Noble.

Figure 2 shows how Copilot generated MARC field 050 for a LCC class for *Algorithms of Oppression*. Again, this response appears correct at first glance, with the caveat that Copilot provided placeholders “##” for the indicators because the prompt did not provide sufficient data to assign those values. In this case, closer examination shows that Copilot was more successful in assigning metadata to the resource. The full call number, including the class number ZA45230 and the item number .N63 2018 match the Library of Congress catalog entry, and the correct MARC field and subfields are used. This correct response, however, is revealing in its own way. Copilot cites a resource

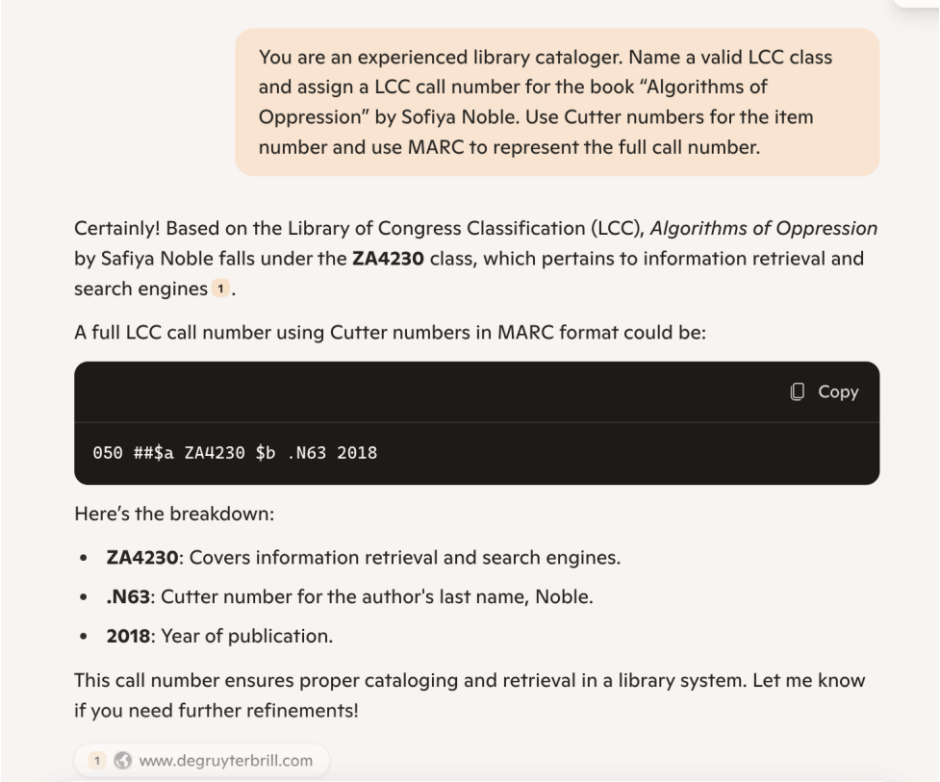

Figure 2. A prompt and Copilot's response as it assigns the MARC field for a DDC call number to the book *Algorithms of Oppression* by Safiya Noble.

("www.degruyterbrill.com"), which points to a PDF with the front matter of the book, including the LCC class number. Therefore, this assignment of a correct call number can largely be attributed to Copilot finding and directly copying cataloging work done by humans, raising the issue of stealing the human cataloger's intellectual work and illustrating how dependent LLMs are on human-created information. To its credit, Copilot determined a correct item number, even though the linked resource did not include a Cutter number. Also, Copilot's response indicated that it understood that the Cutter number was derived from the author's name.

These two examples show the types of responses provided by Copilot and ChatGPT and the analysis performed in this study. This section now moves to analyzing these responses in aggregate, moving through the results for LCSH subject headings, Sears List subject headings, LCC call numbers and MARC fields, and DDC call numbers and MARC fields.

4.1 LCSH Subject Headings

ChatGPT and Copilot struggled to assign accurate, specific, and exhaustive LCSH subject headings to all four types of resources in this study. Errors included inaccurate subject headings, use of unauthorized subject headings (especially unauthorized pre-coordinate phrase headings), both overly broad and overly specific subject headings, overuse and misuse of subdivisions when creating synthetic headings, redundant headings, and lack of exhaustivity in headings. Tables 1 and 2 summarize and quantify the types of errors observed across the ten resources that were cataloged in each category.

In all cases, Copilot and ChatGPT generated between six and ten subject headings and each subject heading was evaluated. The measures for redundancy and exhaustivity considered each set of subject headings collectively.

Table 1. Summary of the types of subject cataloging errors produced by Copilot while cataloging well-known resources, obscure resources, fabricated resources, and full-text articles.

	Well-Known Resources	**Obscure Resources**	**Fabricated Resources**	**Full-Text Articles**
Total subject headings	92	94	83	91
Inaccurate	6	8	-	3
Non-authorized	27	33	41	46
Overly broad	6	7	-	3
Overly specific	3	6	-	8
Overuse or mis-use of subdivisions	18	27	32	26
Redundancy	4	3	4	5
Lack of exhaustivity	2	1	-	4

Table 2. Summary of the types of subject cataloging errors produced by ChatGPT while cataloging well-known resources, obscure resources, fabricated resources, and full-text articles.

	Well-Known Resources	**Obscure Resources**	**Fabricated Resources**	**Full-Text Articles**
Total subject headings	87	91	72	90
Inaccurate	3	5	-	4
Non-authorized	22	23	19	24
Overly broad	7	2	-	5
Overly specific	2	1	-	4
Overuse or mis-use of subdivisions	14	19	12	10
Redundancy	3	5	2	2
Lack of exhaustivity	3	3	-	6

It is important to note that none of the responses for well-known and obscure resources matched the gold standard in the Library of Congress completely. All of these responses contained at least one of the types of errors listed in Tables 1 and 2. While many responses appeared valid at first glance and were generally accurate or about the correct subject, non-authorized terms and misuse of subdivisions (including unauthorized subdivisions) were particularly prevalent for both Copilot and ChatGPT. Overall, 40.8% of the LCSH subject headings that Copilot assigned were non-authorized terms. Copilot fared better, with 25.9% non-authorized terms. Many of the problems were related to the application of subdivisions, with 28.6% of Copilot headings and 16.2% off ChatGPT headings misusing subdivisions or using non-authorized subdivisions. These rates would be considered unacceptably high for even a novice subject cataloger. Finally, much of the redundancy observed for both Copilot and ChatGPT were related to overly broad subject headings, as the LLM assigned a both broader and narrower heading related to the same concept.

Despite these shortcomings, ChatGPT and Copilot performed relatively well in assigning LCSH subject headings for well-known books. For example, six of the ten subject headings assigned for *The Immortal Life of Henrietta Lacks* by Rebecca Skloot matched the Library of Congress record exactly. The detailed construction and specificity of some of these headings (e.g. “Cancer—Patients—Virginia—Biography” and “Human experimentation in medicine—United States—History”) suggest, however, that Copilot simply looked up human-generated subject headings. In contrast, ChatGPT provided detailed, heavily subdivided subject headings that did not exactly match the Library of Congress record (e.g. “Cancer—Patients—United States—Biography” and “Human experimentation in medicine—United States—History—20th Century”). While the ChatGPT-generated subject headings might be considered less “correct” than the Copilot subject headings because they did not match the gold standard exactly, this study considered these close matches to be more impressive because they suggest some type of processing beyond simply looking up an answer.

The study further explored Copilot and ChatGPT’s ability to assign subject headings without the aid of existing records by using fabricated book titles. In all cases, both Copilot and ChatGPT flagged these titles as “hypothetic” or “fictional” or noted that they “couldn’t find specific cataloging details” about the title. However, they still offered suggested subject heading—and, while the suggested subject headings largely sounded accurate, Copilot and ChatGPT (but especially Copilot) struggled to produce authorized headings for these resources in particular, using non-authorized terms and subdivisions at a much higher rate. This finding suggests that the LLMs, especially Copilot, rely heavily on existing cataloging records not to “learn” the rules and syntax of the LCSH vocabulary but to directly copy existing headings.

One positive finding from this part of the study was how the LLMs assigned more exhaustive subject headings for obscure resources. For example, the Library of

Congress record for *Biking Uphill in the Rain: The Story of Seattle from Behind Handlebars* contains only two subject headings: "Cycling—Washington (State)—Seattle—History" and "Urban transportation—Washington (State)—Seattle—History." These headings miss some of the concepts provided by (admittedly unauthorized) headings provided by LLMs: "Public health and transportation" and "Climate change and urban planning." While these proposed subject headings—which appeared to be based on web pages about the book—were not authorized headings and fit a pattern of attempting to construct non-enumerated pre-coordinate phrase headings, they might prompt catalogers to think of additional concepts when assigning subject headings to resources that are not as well-known and well-documented. Similarly, the Copilot response for the article "Indigenous Knowledge Has Been Warning Us About Climate Change for Centuries" suggested possibly redundant headings with different subdivided combinations of phrases related to the environment and Indigenous peoples, but these different suggestions might help a human cataloger arrive at a more exhaustive set of subject headings.

Overall, however, both Copilot and ChatGPT performed poorly in assigning LCSH subject headings, especially Copilot, which appeared to derive most of its "success" from copying existing catalog records.

4.2 Sears List Subject Headings

While the Sears List of Subject Headings is a different and smaller vocabulary than the LCSH, both Copilot and ChatGPT treated prompts for Sears List subject headings as similar to prompts for LCSH subject headings, often returning exactly the same lists of subject headings. These similarities meant that the results for the Sears List were, in many ways, largely the same as those for the LCSH, and results are not presented in full tables here.

It is worth noting, however, that the definition of "authorized heading" is much different for the Sears List, which encourages catalogers to add headings to their local catalog "as needed." Neither Copilot or ChatGPT seemed to recognize or address this nuance, but one big difference between the analysis of LCSH responses and Sears List responses is that this study marked headings as unauthorized only when a heading was assigned that covered the same topic as an existing. For example, Copilot assigned the subject heading "Composition (Language Arts)," which is not enumerated in the Sears List; however, "Writing" is, so adding redundant term would not be authorized by the Sears List. In contrast, ChatGPT used the heading "Northern Ireland—history—1969-1998" for the book *Belfast Diary: War as a Way of Life*. While the date range might be overly specific to the resource, the use of "Northern Ireland" and the "History" subdivision, while not enumerated in the Sears List are specifically authorized in the scope note for "Great Britain."

4.3 LCC and DDC Call Numbers and MARC Fields

Overall, Copilot and ChatGPT had more success with assigning LCC and DDC class numbers and (to a lesser extent) with producing MARC fields than they had with generating subject headings. One reason for this success might be the phenomenon exhibited in the *Algorithms of Oppression* example above, in which the call number was a value that the LLM could simply retrieve as an answer. However, being up to look up the class number was not necessarily the only path to success. For example, Copilot and ChatGPT agreed on the DDC class 808.027 ("Editing and scholarly writing") for the fabricated book title *Corrections: A Personal Guide to Editing Best Practices.* Interestingly, however, ChatGPT used MARC field 050, which is used for a Library of Congress call number for this MARC record and Copilot assigned the Cutter number N995 for the author Nguyen, which is not a particularly good match. Also, interestingly, Copilot assigned the date 2025 to this nonexistent book while ChatGPT omitted the date from the call number, presumably due to lack of data. Neither LLM assigned a value for subfield $2, which is commonly assigned in MARC field 082, presumably because the prompt did not indicate the version of DDC to use. These inconsistencies in formatting MARC fields were more common than issues with the class number assigned.

Tables 3 and 4 summarize data about LCC and DDC class numbers and MARC fields for all four categories of resources used in this study. In all cases, both Copilot and ChatGPT assigned one LCC class and one DDC class per resource, resulting in 20 total classes assigned for each category of resource. In 90.0% of cases where a comparison available, the DDC and LCC classes assigned by the LLMs matched those assigned by the Library of Congress. In the five cases where they did not match, the LLM made a reasonable alternative choice. For example, ChatGPT assigned the LCC class number G155.I2 ("Travel in Iceland") for the book *The Windows of Brimnes: An American in Iceland* by Bill Holm. The Library of Congress assigned the LCC class number PS35558.O3558, which specifically places the book in American Literature from 1961-2000 and reflects the author and subject of the book, Bill Holm. This study agrees that the human cataloger's choice is better, but would not go so far as to call the G155.I2 class number incorrect. Other deviations from the Library of Congress call numbers were similarly justified. For articles that were not cataloged in the Library of Congress, 10.0% of the assigned class numbers were considered inaccurate, suggesting that the LLMs had a more difficult time determining a class number without an existing catalog record. The assigned call numbers for the fabricated resources were not evaluated for accuracy because it was difficult to determine that a call number is "wrong" for a book that does not even exist. However, some class assignments showed that the LLMs struggled with the minimal data available for this task. For example, ChatGPT assigned the LCC class PS3606.A445 for the fabricated book *Here Kitty, Kitty* by Ernest Fallows. This class assumes that *Here Kitty, Kitty* is a work of American literature, which may or may not

be a good assumption. Even though both Copilot and ChatGPT generally performed well on assigning class numbers, catalogers should consider two important caveats: 1) In many cases that success might be based on simply copying the work of a human cataloger, and 2) Catalogers should not trust the class numbers assigned by LLMs just because they look detailed and well-presented in the user interface. In many cases, a better class number might be available, even if the LLM-generated class number is not "wrong."

Table 3. Summary of the types of cataloging errors produced by Copilot while assigning LCC and DDC class number and corresponding MARC fields for well-known resources, obscure resources, fabricated resources, and full-text articles.

	Well-Known Resources	**Obscure Resources**	**Fabricated Resources**	**Full-Text Articles**
Total classes	20	20	20	20
Matches Library of Congress catalog	19	18	-	-
Inaccurate	0	0	-	3
Incorrect mapping of class name to notation	0	0	0	0
MARC syntax incorrect	4	3	5	5
Incorrect application of the Cutter Table*	1	3	0	1

* Note that Cutter numbers do not always follow the Cutter Table perfectly, so this study considered reasonable matches to be correct applications of the Cutter Table.

Copilot and ChatGPT identified classes by both name and notation, and in all cases these representations mapped to each other, showing that the LLMs were able to look up LCC and DDC numbers and follow the complex rules for even LCC class numbers, which can be difficult for humans to assemble. The LLMs were somewhat less successful in applying MARC syntax for fields 050 and 082, with some form of error in 30.0% of cases. Copilot had a particularly high error rate of 42.5%. In some cases, the LLMs confused these fields 050 and 082. In other cases, they omitted commonly used subfields, like $2 for field 082. They were, however, consistent with applying subfield $a for the class number and subfield $b for the item number. In most cases they used ## placeholders for the two indicators, but in some cases, they assigned a value of 0 for one of the indicators.

Table 4. Summary of the types of cataloging errors produced by ChatGPT while assigning LCC and DDC class number and corresponding MARC fields for well-known resources, obscure resources, fabricated resources, and full-text articles.

	Well-Known Resources	Obscure Resources	Fabricated Resources	Full-Text Articles
Total classes	20	20	20	20
Matches Library of Congress catalog	20	15	-	-
Inaccurate	0	0	-	1
Incorrect mapping of class name to notation	0	0	0	0
MARC syntax incorrect	2	0	2	3
Incorrect application of the Cutter Table*	2	2	4	2

This study did not carefully critique the assignment of indicators because they are dependent on the library or the collection. It did, however, note that Cutter number generation was inconsistent, which is maybe unsurprising since item numbers are not always assigned strictly according to the Cutter Table. That said, despite lenient grading in this criterion, 18.9% of Cutter numbers assigned by the LLMs did not appear to be good matches for the author names. Overall, while Copilot and ChatGPT did better with assigning DDC and LCC call numbers and expressing them in MARC, they did not rise to the level of human catalogers—and appeared to be highly dependent on retrieving responses directly from pre-existing work of human catalogers.

5. Implications of LLMs Performing Cataloging Work

While Copilot and ChatGPT show promise as automated catalogers—especially considering they were not purpose-built for cataloging—they exhibit numerous shortcomings, including: lacking specificity, using unauthorized terms, incorrectly assembling synthetic headings, assigning inaccurate headings and classes, and formatting MARC records incorrectly. These shortcomings, even when only one error occurs in a cataloging task or record, are problematic because catalog records must be accurate to be useful and to be trusted. If that one error occurs, or even if a cataloger cannot trust the output of an LLM because of errors in other records, the LLM's cataloging output lacks value because the cataloger is compelled to review the entire output and remedy errors, work that is similar in scope and intellectual effort to the labor involved in cataloging without an LLM. Informal experimentation suggests that corrective prompt engineering can minimize these errors; however, the extra work to formulate and submit additional prompts undermines the time-savings that motivates the pursuit of automated cataloging. Furthermore, those times when Copilot and ChatGPT

produced “correct” responses or matched a gold standard, these responses appeared to rely on basic copying of human cataloging work.

Based on these findings, potential cataloging applications for current LLMs are limited to serving as aides and teaching tools, not as fully automated cataloging solutions. For example, LLMs could suggest subject headings and classes that human catalogers might have overlooked and could help humans appropriately skeptical humans navigate the complex class structure of the DDC and especially the LCC. Alternatively, LLMs could give human catalogers suggestions that might serve as starting points as they begin the cataloging process. ChatGPT appears to make a rhetorical move to embrace this role as aide by discussing a “Justification” for its cataloging choices in some responses. Additionally, because some of the mistakes that LLMs make are similar to the mistakes that novice human catalogers make, LLMs could be used as pedagogical examples and counterexamples. Specifically, students could work with LLMs on cataloging tasks and critique and correct the responses. Establishing an informed, critical relationship with LLMs is essential for cataloging students, as they might be expected to work with LLMs in the future, but they must avoid the de-skilling that comes from over-reliance on technology and remain critical of the outputs that LLMs generate.

While this study found current LLMs inadequate for subject cataloging work, the fast development of LLMs suggest that they might prove more capable of cataloging tasks in the coming years. One area of concern for cataloging and AI where both Copilot and ChatGPT appeared to do an adequate job was in their expression of bias. While their responses undoubtedly showed some biases, neither LLM generated subject headings or class names that the author could identify as problematic, even for sensitive topics around race, identity, and gender. It appears that these systems are designed and tested to be inoffensive. However, even improved cataloging performance would not eliminate the issues related to environmental impact, de-skilling, and intellectual property rights. And cataloging cannot progress without human innovation, as LLMs rely on human-created data for training and for direct access to correct cataloging records for resources. LLMs do not appear ready to replace humans in subject cataloging tasks anytime soon.

6. Conclusion

While ChatGPT and Copilot did not perform well enough to replace human catalogers, their partial success suggests that generative AI has some potential for working with KOSs and supporting subject cataloging work. That potential, however, should be pursued with caution because of the environmental impacts of LLMs, the de-skilling that could occur if LLMs take over all or part if subject cataloging work, the tendency of LLMs to perpetuate bias by using biased training data, and the threat of LLMs appropriating human catalogers’ intellectual property. The potential of generative AI should also remain focused on human agency and the value that humans bring to the subject cataloging process.

With this in mind, future research could pursue purpose-built SLMs (Bucher and Martini, 2024) that are trained by human catalogers to specifically to work with the rules and syntax of KOSs and to support KOS-specific calculations like deriving Cutter numbers. These smaller, purpose-built tools could assist catalogers in their cataloging processes and, importantly, afford them some control over these quickly evolving technologies. SLMs are also less data-intensive and, as a result, less energy-intensive

and less harmful to the environment. Additional research should also directly compare the time to complete subject cataloging tasks and the quality of cataloging work performed by catalogers working with their established methods to catalogers working with LLMs as aides because, while it is easy to ask LLMs to perform cataloging tasks and intriguing to critique their responses, it is unclear whether and how they can serve as effective aides for human catalogers. All of this research should be conducted with a appropriate skepticism for LLMs and appropriate respect for the longstanding effectiveness and ingenuity of human subject catalogers.

7. References

Adler, M. and J.T. Tennis. 2013. Toward a taxonomy of harm. *NASKO* 4: 1-19.

Ahn, J., R. Verma, R. Lou, D. Liu, R. Zhang, and W. Yin. 2024. Large language models for mathematical reasoning: Progresses and challenges. *arXiv preprint arXiv:2402.00157*.

Amram, T., R. Goodfellow Malamud, and C. Hollingsworth. Response to "From ChatGPT to CatGPT." *Information Technology and Libraries* 42, no. 4.

Bair, S. 2005. Toward a code of ethics for cataloging. *Technical services quarterly* 23, no. 1: 13-26.

Bender, E.M., T. Gebru, A. McMillan-Major, and S. Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? 🦜. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pp. 610-623.

Bibi, Z., and A. Atta. 2024. The role of ChatGPT as AI English writing assistant: A study of student's perceptions, experiences, and satisfaction. *Annals of Human and Social Sciences* 5, no. 1: 433-443.

Brzustowicz, R. 2023. From ChatGPT to CatGPT: The implications of artificial intelligence on library cataloging. *Information Technology and Libraries* 42, no. 3.

Bucher, M.J.J. and M. Martini. 2024. "Fine-Tuned 'Small' LLMs (Still) Significantly Outperform Zero-Shot Generative AI Models in Text Classification." *arXiv preprint arXiv:2406.08660*.

Chow, E.H.C., T.J. Kao, and X. Li. 2024. An Experiment with the Use of ChatGPT for LCSH Subject Assignment on Electronic Theses and Dissertations. *Cataloging & Classification Quarterly* 62, no. 5, pp. 574-588. https://doi.org/10.1080/01639374.2024.2394516.

Cervantes, J., F. Garcia-Lamont, L. Rodríguez-Mazahua, and A. Lopez. 2020. A comprehensive survey on support vector machine classification: Applications, challenges and trends. *Neurocomputing* 408: 189-215.

Coello, C.E.A., M.N. Alimam, and R. Kouatly. 2024. Effectiveness of ChatGPT in coding: a comparative analysis of popular large language models. *Digital* 4, no. 1: 114-125.

Cutter, C. A., and Ford, W. C. 1904. *Rules for a Dictionary Catalog*, vol. 10. US Government Printing Office.

Dobreski, B. and C. Hastings. 2025. AI Chatbots and Subject Cataloging: A Performance Test. *Library Resources & Technical Services*.

Foskett, A. C. 1996. *The Subject Approach to Information*. Facet Publishing.

Ghiassi, M., M. Olschimke, B. Moon, P. Arnaudo. 2012. Automated Text Classification Using a Dynamic Artificial Neural Network Model. *Expert Systems with*

Applications, vol. 39, no. 12, pp. 10967–76, https://doi.org/10.1016/j.eswa.2012.03.027.

Golub, K. 2021. Automated subject indexing: An overview. *Cataloging & Classification Quarterly* 59, no. 8: 702-719.

Gómez-Rodríguez, C., and P. Williams. 2023. A confederacy of models: A comprehensive evaluation of LLMs on creative writing. *arXiv preprint arXiv:2310.08433*.

Hu, J., K. Mahowald, G. Lupyan, A. Ivanova, and R. Levy. 2024. Language models align with human judgments on key grammatical constructions. *Proceedings of the National Academy of Sciences* 121, no. 36: e2400917121.

Kazi, N., N. Lane, and I. Kahanda. 2021. Automatically cataloging scholarly articles using library of congress subject headings. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop*, pp. 43-49.

Ji, Z., N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, YJj. Bang, A. Madotto, and P. Fung. 2023. Survey of hallucination in natural language generation. ACM Computing Surveys, 55(12):1–38.

Lancaster, F. W. 2003. *Indexing and Abstracting in Theory and Practice*, Third edition. University of Illinois, Champaign, Illinois.

Lozić, E., and B. Štular. 2023. Fluent but Not Factual: A Comparative Analysis of ChatGPT and Other AI Chatbots' Proficiency and Originality in Scientific Writing for Humanities *Future Internet* 15, no. 10: 336. https://doi.org/10.3390/fi15100336.

Microsoft. Copilot. Accessed 2025-05-16. https://copilot.microsoft.com.

Moulaison-Sandy, H. and Z. Coble. 2024. Leveraging AI in Cataloging: What Works, and Why?. *Technical Services Quarterly* 41, no. 4: 375-383.

OpenAI. ChatGPT. Accessed 2025-05-16. https://chatgpt.com/.

Petroni, F., A. Piktus, A. Fan, P. Lewis, M. Yazdani, N. De Cao, J. Thorne, Y. Jernite, V. Karpukhin, J. Maillard, et al. 2021. Kilt: a benchmark for knowledge intensive language tasks. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2523–2544.

Ranganathan, S. 1937. *Prolegomena to Library Classification*. Madras Library Association.

Rogers, A,, O. Kovaleva, and A. Rumshisky. 2021 A primer in BERTology: What we know about how BERT works. *Transactions of the association for computational linguistics* 8: 842-866.

Svenonius, E. 200. *The Intellectual Foundation of Information Organization*. MIT press.

Watson, Brian M. 2021. Advancing equitable cataloging. *NASKO 2021*: 1-28.

8. Appendix

This appendix lists the resources (real and fabricated) that were used in this study and the templates for the prompts submitted to ChatGPT and Copilot in this study.

8.1 Well Known Resources Cataloged in the Study

The following titles were chosen as well-known resources because they appeared on many "best of" lists and had prominent online presences, offering LLMs with a significant amount of metadata:

- *Algorithms of Oppression* by Safiya Noble
- *The Immortal Life of Henrietta Lacks* by Rebecca Skloot
- *Caste: The Origins of Our Discontent* by Isabel Wilkerson
- *The Information* by James Gleick
- *The New Jim Crow: Mass Incarceration in the Age of Colorblindness* by Michelle Alexander
- *The Emperor of All Maladies* by Siddhartha Mukherjee
- *The Omnivore's Dilemma* by Michael Pollan
- *The Sixth Extinction* by Elizabeth Kolbert
- *Thinking Fast and Slow* by Daniel Kahneman
- *Team of Rivals* by Doris Kearns Goodwin

8.2 Less Popular Resources Cataloged in the Study

The following titles were chosen as less popular resources because they had less than 300 reviews on Goodreads:

- *The Filing Cabinet: A Vertical History of Information* by Craig Robertson
- *Biking Uphill in the Rain: The Story of Seattle from Behind Handlebars* by Tom Fucoloro
- *The Windows of Brimnes: An American in Iceland* by Bill Holm
- *Masked by Trust: Bias in Library Discovery* by Matthew Reidsma
- *Belfast Diary: War as a Way of Life* by John Conroy
- *Style: Lessons in Clarity and Grace* by Joseph M. Williams
- *BKLN Manners: Positive Training Solutions for Your Unruly Urban Dog* by Kate Naito
- *The Art of Conversation: A Guided Tour of a Neglected Pleasure* by Catherine Blyth
- *How Italian Food Conquered the World* by John F. Mariani
- *Poplorica: A Popular History of the Fads, Mavericks, Inventions, and Lore that Shaped Modern America* by Martin J. Smith and Patrick J. Kiger

8.3 Fabricated Resources

The following titles were fabricated based on the author's imagination and designed to test how the LLMs would react to resources that did not exist and thus had no existing metadata:

- *An Irreverent of History Typewriters and Keyboards* by Anders Schlott
- *Moving West: Western Expansion and the Cruelty of Manifest Destiny* by Margaret Johnson
- *Corrections: A Personal Guide to Editing Best Practices* by Evie Nguyen
- *Modern Thinking, Modern Art* by Gwen Brimley
- *Reverberations* by Jasmine Young
- *Here Kitty, Kitty* by Ernest Fallows
- *Flora and Fauna of the Nigerian Savannah* by Naima Kurtz
- *Designing for Hygge* by Tove Johansen
- *Courting Disaster: Conservative Activism on the Bench* by Emile Rodriguez
- *Native Textiles* by June Cleaves

8.4 Full-Text Resources

The following titles were chosen as full-text articles that the LLMs could analyze in full at the provided URL. These articles were chosen from annual "Best of" lists from the website Longreads.

- "His Best Friend Was a 250-Pound Warthog. One Day It Decided to Kill Him" by Peter Holley. Available at: https://www.texasmonthly.com/news-politics/warthog-attack-texas-exotics/
- "Crying Myself to Sleep on the Biggest Cruise Ship Ever" by Gary Shteyngart. Available at: https://www.theatlantic.com/magazine/archive/2024/05/royal-caribbean-cruise-ship-icon-of-seas/677838/
- "Vigilantes for Views: The YouTube Pranksters Harassing Suspected Scam Callers in India" by Andrew Deck and Raksha Kumar. Available at: https://restofworld.org/2023/youtube-scam-call-vigilantes/
- "Edifice Complex" by Bench Ansfield. Available at: https://jewishcurrents.org/edifice-complex
- "What Happened to the Women Prisoners at Hickman's Farms" by Elizabeth Whitman. Available at: https://www.cosmopolitan.com/lifestyle/a42710907/women-prisoners-at-hickmans-farms/
- "Climate Signs" by Emily Raboteau. Available at: https://www.nybooks.com/online/2019/02/01/climate-signs/

- "Indigenous Knowledge Has Been Warning Us About Climate Change for Centuries" by Malcolm Harris. Available at: https://psmag.com/ideas/indigenous-knowledge-has-been-warning-us-about-climate-change-for-centuries/
- "The Doctor vs. #MeToo" by Caitlin L. Chandler. Available at: https://www.cjr.org/special_report/heiko-jessen-germany-me-too.php
- "Dying on the Waitlist" by David Armstrong and Marshall Allen. Available at: https://www.propublica.org/article/dying-on-the-waitlist
- "Invisible Child: Dasani's Homeless Life" By Andrea Elliott. Available at: https://www.nytimes.com/projects/2013/invisible-child/index.html#/?chapt=1

8.5 Templates for Prompts

The following templates were used for prompting ChatGPT and Copilot, with the appropriate titles, authors, and URLs inserted into the indicated locations in the templates.

8.5.1 Prompt Template for Library of Congress Subject Headings

You are an experienced library cataloger. Assign Library of Congress subject headings for the [*book/article*] "[*title of resource*]" by [*author name*]. [*You can access the full text of the article at URL*.]

8.5.2 Prompt Template for Sears List of Subject Headings

You are an experienced library cataloger. Assign Sears List subject headings for the [*book/article*] "[*title of resource*]" by [*author name*]. [*You can access the full text of the article at URL*.]

8.5.3 Prompt Template for Library of Congress Classification and MARC

You are an experienced library cataloger. Name a valid LCC class and assign a LCC call number for the [*book/article*] "[*title of resource*]" by [*author name*]. Use Cutter numbers for the item number and use MARC to represent the full call number. [*You can access the full text of the article at URL*.]

8.5.4 Prompt Template for Dewey Decimal Classification and MARC

You are an experienced library cataloger. Name a valid DDC class and assign a DDC call number for the [*book/article*] "[*title of resource*]" by [*author name*]. Use Cutter numbers for the item number and use MARC to represent the full call number. [*You can access the full text of the article at URL*.]