DCPAPERS

Assessing Large Language Models: Architectural Archive Metadata and Transcription

Hannah Chavez Moutran[1,†], Devon Murphy[1,*,†], Karina Sanchez[1,*,†], Willem Borkgren[1,*,†], Katie Pierce Meyer[1,†] and Josh Conrad[1,†]

[1]*University of Texas at Austin Libraries, 101 E 21st Street, Austin, TX, United States*

Abstract

Our research explores whether Large Language Models (LLMs) can offer a solution for improving the efficiency of developing detailed, rich metadata for large digitized collections. We tested the ability of seven widely available LLMs to complete four metadata generation tasks for a selection of pages from the Southern Architect and Building News (1882-1932): assigning subject headings; creating short content summaries; extracting named entities; and writing transcriptions. Our cross-departmental team evaluated the quality of the outputs, the cost, and the time efficiency of using LLMs for metadata workflows. To do so, we developed a metadata quality rubric and scoring schematic to ground our results. Analysis suggests that models can perform interpretive metadata tasks well, but lack the accuracy needed for assigning terms from controlled vocabularies. With careful implementation, thorough testing, and creative design of workflows, these models can be applied with precision to significantly enhance metadata for digitized collections.

Keywords

large language models, artificial intelligence, metadata, archives, digital collections

1. Introduction

Our collections often remain frustratingly separate from one another. The promise of increased access to GLAMS (Galleries, Libraries, Archives, Museums, Special Collections), including efforts in digital scholarship and linked data, can only extend to those items that are thoroughly and accurately described. This work is labor and time intensive, and many collections are either minimally described or set aside in the hopes of securing more staff or funding. Advancements in artificial intelligence (AI) could be one of many tools to address this issue.

Exploration of LLMs in metadata and cataloging workflows is ongoing. Recent OCR studies show promise: Greif et al. (2025) [1] found Gemini 2.0 Flash and GPT-4o matched traditional OCR tools like Transkribus and Tesseract, while Thomas et al. (2024) [2] showed Llama 2 outperformed a fine-tuned BART in post-OCR error correction. Chen and Li's (2024) [3] survey of cataloging professionals revealed concerns about reliability and potential for de-professionalization of the field. The Library of Congress recommends members evaluate AI

DCMI-2025 International Conference on Dublin Core and Metadata Applications

*Corresponding author.

†These authors contributed equally.

hannah.chavezmoutran@austin.utexas.edu (H. C. Moutran); devon.murphy@austin.utexas.edu (D. Murphy); karina.sanchez@austin.utexas.edu (K. Sanchez); willem.borkgren@austin.utexas.edu (W. Borkgren); katie.meyer@austin.utexas.edu (K. P. Meyer); josh.conrad@austin.utexas.edu (J. Conrad)

https://doi.org/10.23106/dcmi.952583993

projects by questioning if implementation would be "responsible, effective, and practical" (Brador, 2024) [4]. Several studies emphasize the importance of human oversight. Huang et al. (2024) [5] found GPT-4o was cost-effective and scalable for web archive metadata generation but produced more inaccuracies and hallucinations than human-curated metadata. Toth et al. (2025) [6] found ChatGPT-4 valuable for revealing hidden structures in archival data despite occasional hallucinations that required human verification. Building on this work, we utilized core principles of the PROMPT design framework by creating a persona for the LLMs and contextualizing requests through detailed requirements, format, intended purpose, audience, and professional expectations such as tone (Hartman-Caverly, 2024) [7].

Our research assesses the performance of seven LLMs in the generation of metadata and transcriptions for archival materials. We tested 14 images of historical architecture magazine pages and previously created optical character recognition (OCR) text of those pages. Our study tested the LLMs via Application Programming Interface (API) in an automated workflow as well as a web-based interface. In this paper, we describe our process of evaluation, analysis of the efficacy of LLMs in various description tasks, recommendations for implementation, and ethical and practical considerations of AI metadata projects.

2. Process

Our cross-departmental team included six staff members and student workers with expertise in metadata, digital collections, AI implementation, archives, architectural history, and digital scholarship. This approach allowed us to develop comprehensive evaluation criteria and assessment methods. The project focused on the Southern Architect and Building News (SABN) collection (1882-1932), a monthly trade publication documenting architecture and building trades in the southern United States. It is comprised of 254 distinct issues that have been digitized with basic metadata but not fully indexed or described. More robust metadata will address recent research interest expressed at our university. The issues chosen contain illustrated covers, advertisements, and editorial content, and incorporate a diverse range of fonts, symbols, photographs, and drawings.

Through group discussion and review of local metadata best practices, we established three primary evaluation criteria: completeness, or whether descriptions fully described the contents of a page; accuracy, the correctness of the OCR text, descriptions, terms and entities when compared to the original item; and usefulness, or the increased value for search. Similar criteria have been employed in Zavalin and Zavalina (2025) [8] and Shyr et al (2024) [9]. Based off of these three primary criteria, we scored results using a scale of 1-4. Blank responses returned by models were given a 0.

We tested models' abilities to perform five key tasks: OCR text editing and transcription; table of contents entry generation; named entity extraction; and Library of Congress Subject Heading (LCSH) generation. The web and API testing utilized different page selections strategically in order to evaluate diverse content types. Results were manually compiled into Excel spreadsheets for evaluation. We conducted blind evaluations wherein model names were obscured in testing spreadsheets to reduce tester bias for or against any models or companies. Each team member independently scored all outputs, and we maintained notes to capture qualitative observations.

https://doi.org/10.23106/dcmi.952583993

Table 1
Evaluation Rubric

Score	Definition
1	Limited, partial or inaccurate result; many errors and/or formatting issues
2	Basic coverage of content and can improve discoverability, but errors present
3	Full and correct coverage of material with limited extraneous content and few errors
4	No errors, excellent formatting, improves accessibility and readability

Five models were tested through their official web interfaces: GPT-4o-mini, Google Gemini, Claude Sonnet, Haiku when Sonnet was briefly unavailable, Llama, and Microsoft Copilot (for the image input workflow only). Testing was conducted using free accounts, with the exception of Copilot, which was tested using an institutional account. We selected four pages from three different SABN issues: two full page editorials published in 1930, and two dual-column pages with multiple blocks of text published in 1928. Web interfaces were tested in two phases: text input testing (October 2024) and image input testing (February 2025). The API prompts were divided into sections due to limitations on unpaid accounts.

We tested six models via API: Claude Haiku, Claude 3 Sonnet, Claude Sonnet 3.5, GPT-4o-mini, GPT-4o, and Gemini 2.0 Flash Experimental (for the image input workflow only). We selected ten pages from a 1924 and a 1931 issue. The pages included a cover page; a contents page with advertising schedule; advertisements with visual elements and complex layouts; editorial content with multiple blocks of text; a page with a captioned photograph of a home in Alabama; and a text-heavy article. API testing was conducted in two phases: OCR text input testing (December 9, 2024) and image input testing (December 11-12, 2024). We developed Python scripts to automatically send input files with prompts to each LLM. The scripts captured the LLM outputs and wrote them to Excel spreadsheets for evaluation.

Sentiment analysis was performed on our scoring notes, 56 total, to evaluate trends and biases. Each project member's annotations were compiled into a text document and imported into R Studio for analysis, using the sentimentr package. While testers scored the LLMs differently on an absolute scale, with some perhaps more likely to give high or low scores, the relative sentiment expressed in comments was consistent across testers for the various models.

3. Model Comparison

Claude Sonnet 3.5 emerged as the top performer, achieving the highest overall scores (see Table 2 below.) The performance hierarchy was fairly consistent for each model regardless of testing method, but models performed better via API than web interface due to their relative stability. At the time of testing, Copilot only accepted PDFs, and Haiku was only available to unpaid accounts for a short time, resulting in some N/A values. Gemini and Llama returned few to no responses, with Gemini having difficulty in longer sessions and Llama not allowing file uploads. These are represented below with a 0.00 score.

In terms of specific tasks, OCR editing/transcription and content summarization were the greatest strengths of the LLMs (see Table 3). AI-generated OCR results outperformed our

https://doi.org/10.23106/dcmi.952583993

Table 2
Average Scores Across all Tasks with Different Input Formats

Model	API – Text	Web – Text	API – Image	Web – Image
Claude Sonnet 3.5	3.09	3.03	2.99	2.90
Claude 3 Sonnet	3.02	N/A	2.74	N/A
Copilot	N/A	N/A	N/A	3.00
Haiku	2.82	N/A	2.74	2.74
GPT-4o	2.81	N/A	3.15	N/A
GPT-4o-mini	2.76	2.58	2.75	2.80
Llama	N/A	2.54	0.00	0.00
Gemini	N/A	2.44	0.00	0.00

original OCR, in line with findings from Thomas et al (2024) [2]. Entity extraction was ranked as the second-lowest performing task, but testers felt it provided valuable metadata enhancement that could improve discoverability. Subject heading creation emerged as the weakest task across all models tested. The difficulty primarily stemmed from the complicated and variable syntax of generating exact matches of LCSH. Controlled vocabularies remain a challenge for LLMs, as found in Chow et al (2024) [10]. One surprise was that, though in general larger models produced the best results, Haiku performed best in content summarization.

Table 3
Average Scores Across by Task for Each Model (Text Input Score/Image Input Score)

Model/Access Point	OCR	ToC Entry	Entities	Subject Headings
Claude Sonnet 3.5/API	3.28/3.08	3.19/3.71	3.02/3.08	2.88/2.10
Claude 3 Sonnet/API	3.33/3.15	3.12/2.48	2.89/2.93	2.74/2.43
Haiku/API	2.88/2.76	3.34/2.88	2.59/3.15	2.49/2.18
GPT-4o/API	2.88/3.80	2.93/3.25	2.81/3.20	2.62/2.35
GPT-4o-mini/API	2.85/3.00	3.04/3.13	2.65/2.70	2.50/2.20
Llama/Web	2.14/0.00	2.80/0.00	2.83/0.00	2.41/0.00
Gemini/Web	2.00/0.00	3.15/0.00	2.00/0.00	2.63/0.00
GPT-4o-mini/Web	2.80/3.18	2.56/2.50	2.70/3.15	2.30/2.40
Claude Sonnet 3.5/Web	3.50/2.08	2.87/3.71	3.15/3.36	2.60/2.48

Some content presented more challenges. Cover pages received low scores in entity extraction and table of contents creation. This was likely due to the limited textual content available for analysis. Transcribing and editing OCR for text-heavy pages tended to have more small errors, though Claude 3.5 was an exception. Multi-column layouts and pages with mixed text and images could be a challenge for current LLMs, but they also showed strength when compared to traditional OCR technologies.

3.1. Model Responses to Offensive Historical Content

Models responded differently to a racist slur that was used in an advertisement published in the journal. In image-based API testing, Gemini and GPT-4o-mini refused to process the page.

https://doi.org/10.23106/dcmi.952583993

In text-based API testing, all models processed the page but Claude Sonnet 3.5 substituted the offensive term with a neutral word during OCR cleaning. Complete refusal to process materials containing offensive historical language or uncritical substitution of terms could result in collection gaps (Antracoli et al., 2019) [11], rendering important historical materials undiscoverable. Alternatively, Claude 3 Sonnet and Haiku wrote the term in the content warning and short summary. Repeating harmful terms can result in psychological damage known as vicarious trauma (Wright & Laurent, 2021) [12]. GPT-4o was the only model that transcribed the term across API workflows without repeating it. Collections and communities vary, and LLM use will need to be fitted to specific circumstances (Antracoli et al., 2019; Wilson Special Collections Library, 2022) [11, 13]. Importantly, web-based models carry the risk that input data will be used for training, risking exposure and replication of sensitive information.

3.2. Other Implementation Considerations

3.2.1. Cost

In general, larger models produced better output in our testing. Claude Sonnet 3.5 delivered the highest quality scores but at a higher cost, with image processing at $14.16 and text at $11.18 per 1,000 pages. GPT-4o-mini offered the best cost efficiency, with image at $1.15 and text at $0.52 per 1,000 pages, though it had lower overall quality. Notably, Haiku performed especially well in summarization despite being a smaller and more affordable model at $0.86 for image and $1.34 for text per 1,000 pages.

3.2.2. Processing Time

GPT-4o had the longest processing times regardless of format, followed by GPT-4o-mini for image processing. Running 1,000 pages with GPT-4o would take about 5.5 hours, while Haiku could handle the same volume in under two hours. Image processing was generally slower than text, though the difference was sometimes negligible. Gemini and Haiku were consistently the fastest models. For large-scale projects, batch processing can reduce costs by up to 50 percent while ensuring results within 24 hours.

3.2.3. Input Format

Text inputs are generally more cost-effective than images, but overall cost difference may vary depending upon collection size and model. For example, Claude 3 Sonnet costs about $6.27 more per 1,000 pages for image analysis, while Haiku costs $0.48 more. Anthropic models, including Claude Sonnet 3.5 and Haiku, struggled with smaller text in images, likely due to lower-resolution image ingest during pre-processing compared to OpenAI and Gemini models.

3.2.4. Recommendations

Budget-conscious projects can use smaller LLMs through APIs or choose to use web-based chatbots, though reliability, quality, and data privacy require consideration. We recommend choosing models based on demonstrated strengths in specific tasks, collection requirements,

https://doi.org/10.23106/dcmi.952583993

and institutional access. Batch processing can be used via API to reduce costs, and results can be improved through prompt optimization, rigorous testing, and formalized human oversight protocols. Pre-processing and human evaluation should strive to match the growing best practice of including community stakeholders in metadata creation (University of California Irvine Libraries, 2025; Trans Metadata Collective et al., 2022) [14, 15]. Community awareness of ethical pitfalls may be more acute than that of archivists, often a result of lived experience as well as subject expertise.

4. Conclusion

Our research demonstrates that LLMs can offer promising capabilities for enhancing metadata creation in archival collections, though with limitations that require strategic implementation. A significant challenge will be hosting and supporting search of the increased volume of the metadata generated, including fields describing AI usage and mechanisms allowing users to flag potential inaccuracies. Other challenges include keeping up with the AI landscape, as noted in Cushing and Osti (2022) [16], navigating the stability and security of the tools available, and processing harmful content. This could prove to be very challenging for smaller or under-resourced institutions. We conclude that LLM workflows designed carefully and in accordance with archival best practices can increase access, enrich metadata, and allow libraries and archives to address processing backlogs.

4.1. Generative AI Use

We employed Anthropic, OpenAI, Gemini, and Llama models for the following purposes: generating metadata for our research and generating and debugging code used in the API workflow. We evaluated the metadata output through the methods outlined in this paper and thoroughly tested and debugged the code. The authors assume all responsibility for the content of this submission.

References

[1] Greif, G., Griesshaber, N., & Greif, R. (2025). Multimodal LLMs for OCR, OCR post-correction, and named entity recognition in historical documents [Preprint]. arXiv. https://arxiv.org/abs/2504.00414

[2] Thomas, A., Gaizauskas, R., & Lu, H. (2024). Leveraging LLMs for post-OCR correction of historical newspapers. In Proceedings of the Third Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA) @ LREC-COLING-2024 (pp. 116–121). ELRA and ICCL.

[3] Chen, S., & Li, M. (2024). AI for cataloging and metadata creation: Perspectives and future opportunities from cataloging and metadata professionals. Technical Services Quarterly, 41(4), 317–332. https://doi.org/10.1080/07317131.2024.2394919

[4] Brador, I. (2024, November 19). Could artificial intelligence help catalog thousands of digital library books? An interview with Abigail Potter

https://doi.org/10.23106/dcmi.952583993

and Caroline Saccucci. The Signal. https://blogs.loc.gov/thesignal/2024/11/could-artificial-intelligence-help-catalog-thousands-of-digital-library-books-an-interview-with-abigail-potter

[5] Huang, A. Y., Nair, A., Goh, Z. R., & Liu, T. (2024). Web archives metadata generation with GPT-4o: Challenges and insights. arXiv.org.

[6] Toth, G. M., Albrecht, R., & Pruski, C. (2025). Explainable AI, LLM, and digitized archival cultural heritage: A case study of the Grand Ducal Archive of the Medici. AI & Society. https://doi.org/10.1007/s00146-025-02238-5

[7] Hartman-Caverly, S. (2024). PROMPT design worksheet. In Hidden layer: Intellectual privacy and generative AI. https://guides.libraries.psu.edu/berks/AI

[8] Zavalin, V., & Zavalina, O. L. (2025). Are we there yet? Evaluation of AI-generated metadata for online information resources. Information Research.

[9] Shyr, C., Grout, R. W., Kennedy, N., Akdas, Y., Tischbein, M., Milford, J., Tan, J., Quarles, K., Edwards, T. L., Novak, L. L., White, J., Wilkins, C. H., & Harris, P. A. (2024). Leveraging artificial intelligence to summarize abstracts in lay language for increasing research accessibility and transparency. Journal of the American Medical Informatics Association, 31(10), 2294–2303. https://doi.org/10.1093/jamia/ocae186

[10] Chow, E. H. C., Kao, T. J., & Li, X. (2024). An experiment with the use of ChatGPT for LCSH subject assignment on electronic theses and dissertations. https://arxiv.org/abs/2403.16424

[11] Antracoli, A., Berdini, A., Bolding, K., Charlton, F., Ferrara, A., Johnson, V., & Rawdon, K. (2019). Archives for Black Lives: Anti-Racist Descriptions. Archives for Black Lives in Philadelphia (A4BLiP).

[12] Wright, K., & Laurent, N. (2021). Safety, Collaboration, and Empowerment: Trauma-Informed Archival Practice. Archivaria, 91, 38–73.

[13] Wilson Special Collections Library. (2022). A Guide to Conscious Editing at Wilson Special Collections Library. The University of North Carolina at Chapel Hill University Libraries.

[14] University of California Irvine Libraries. (2025). Community-Centered Archives Practice: Transforming Education, Archives, and Community History. https://sites.uci.edu/ccap/

[15] The Trans Metadata Collective, Burns, J., Cronquist, M., Huang, J., Murphy, D., Rawson, K. J., Schaefer, B., Simons, J., Watson, B. M., & Williams, A. (2022). Metadata Best Practices for Trans and Gender Diverse Resources (1.0). Zenodo. https://doi.org/10.5281/zenodo.6686841

[16] Cushing, A. L., & Osti, G. (2022). "So how do we balance all of these needs?": how the concept of AI technology impacts digital archival expertise. Journal of Documentation, 79(7), 12–29. https://doi.org/10.1108/JD-08-2022-0170

A. Test Pages from Southern Architect and Building News

Denmark, E. R. (Ed.), & Harman, H. E. (Pres.). (1924, October). Southern Architect and Building News, 50(10). https://collections.lib.utexas.edu/catalog/utlarch:8fb93ee5-5b4b-4661-9a8c-ca0d76960182

Denmark, E. R. (Ed.), & Watts, B. H. (Ed.), & Harman, H. E., Jr. (Mgr.). (1928, February). Southern Architect and Building News, 54(2). https://collections.lib.utexas.edu/catalog/utlarch:db10bf40-401c-4cde-b845-0358fffcba54

https://doi.org/10.23106/dcmi.952583993

Denmark, E. R. (Ed.), Harman, H. E., Jr. (Bus. Mgr.), & Sorrow, F. H. (Adv. Mgr.). (1930, March). Southern Architect and Building News, 56(3). https://collections.lib.utexas.edu/catalog/utlarch:69db3ebf-90b4-41b4-9967-5aca81bfecd3

Denmark, E. R. (Ed.), Harman, H. E., Jr. (Bus. Mgr.), & Sorrow, F. H. (Adv. Mgr.). (1930, August). Southern Architect and Building News, 56(8). https://collections.lib.utexas.edu/catalog/utlarch:6485cc23-e9e9-41e3-8ece-6052422ba1a6

Denmark, E. R. (Ed.), Harman, H. E., Jr. (Bus. Mgr.), & Sorrow, F. H. (Adv. Mgr.). (1931, December). Southern Architect and Building News, 57(12). https://collections.lib.utexas.edu/catalog/utlarch:b17ca698-bfaa-4855-ae0c-168d01ce5568

https://doi.org/10.23106/dcmi.952583993