Luka Hribar[1]

WHEN AI MEETS ARCHIVES: TESTING COMMERCIAL GENERAL-PURPOSE LLMS FOR TECHNICAL AND CONTENT-RELATED DESCRIPTION

Abstract

Purpose: *To test the usability and examine the limitations of general-purpose large language models (LLMs) in archival description. The study was designed as a quantitative/qualitative assessment to monitor trends in this rapidly evolving field.*

Methodology: *The experiment involved testing five AI services on a set of archival records. The set of questions and tasks was divided into two categories: technical tasks (page counting, structure recognition, optical character recognition – OCR) and content-related tasks, such as language detection, content summarization, and title suggestions. Performance was evaluated using quantitative and qualitative methods, along with archivists' assessments.*

Results: *A significant discrepancy was found between the models' performance across different types of tasks. The tested models proved unreliable in seemingly simple technical tasks, such as determining the number of pages or detecting graphical elements, while showing greater utility in complex content-related tasks.*

Discussion: *The analysis highlights that the tested LLMs are currently unsuitable for automating precise technical description processes but represent a useful analytical and generative tool for producing content summaries and descriptions. By observing how AI systems perform, archivists also gain better insight into potential difficulties faced by users.*

Keywords: *Archival records, artificial intelligence (AI), large language models (LLM), digital humanities, archival cataloguing and description, OCR.*

1 Luka Hribar, PhD student of Archival Sciences at Alma Mater Europaea University, email: luka.hribar@almamater.si.

1 INTRODUCTION

Over the past decade, digital transformation has profoundly reshaped the operations of archives, libraries, and museums (the GLAM sector). With accelerated digitization and the growing creation of extensive born-digital records, institutions are facing an exponential increase in data that exceeds the capacity of traditional, manual cataloguing and processing methods. In this context, artificial intelligence (AI), particularly large language models (LLMs), offers potential solutions. The integration of AI into archival practice, such as transcription, description, and content analysis, represents one of the key challenges and opportunities currently faced by the profession. The development of AI also promises the opening of so-called "dark archives" (Decker et al., 2022), which remain inaccessible to the public due to insufficient metadata, sensitive content, or disorganization.

The purpose of this research was to conduct a systematic test of the capabilities of several commercially available general-purpose LLMs in performing tasks specific to archival description. The primary goal was to investigate to what extent these services are already useful for archivists and external users, and to identify their strengths and limitations. Special emphasis was placed on understanding where the models function effectively and where systemic shortcomings appear that could affect trust and reliability.

The aim was not to identify every correct and incorrect statement in detail, nor to determine a single "winner," but rather to establish a foundation for monitoring future developments and to provide guidelines for the integration of AI into archival practice.

2 CATALOGUING AND DESCRIBING ARCHIVAL MATERIALS

The purpose of cataloguing and describing archival records goes beyond merely creating inventories or lists. Its traditional core task is to establish intellectual access tools that can also serve as descriptive surrogates for the physical material (Pezzica, 2023). These tools enable users to discover and understand archival material through accurate descriptions of its content and context. Describing represents a crucial task of the archivist, as it transforms or supplements the material into an accessible source of knowledge. With the advent of the web and the

provision of digitized materials, this role has also started to change. Younger generations of external, non-professional users are often unaware of the existence of certain archival tools (Hankins, 2019), such as archival inventories, which are sometimes attached to higher levels (fonds, collections, and series) rather than to the individual archival unit.

To ensure consistency and interoperability, archivists around the world rely on international standards such as ISAD(G), RIC, and others. These standards provide a common framework and structure for describing archival records regardless of their form (written, printed, or electronic). The principles include multi-level description, ensuring relevance at each level, linking descriptions, and avoiding redundancy. In our experiment, we did not require the AI services to present information in the form of any of these established standards. However, when formulating prompts, we strived to achieve relatively simple mappings of the results to appropriate categories. This also allowed for easier and more effective evaluation of AI's capabilities in producing descriptions consistent with professional requirements.

Describing archival material nevertheless faces numerous challenges. The primary problem that AI seeks to address is the quadruple deficit of resources: a shortage of adequately trained staff, specialized knowledge, funding, and time (Bingham & Byrne, 2021). These factors contribute to significant backlogs in archival processing worldwide. As a result, much valuable and interesting material remains undescribed or only at a very basic level, hindering discovery and use.

2.1 THE STATE OF DESCRIBING IN SLOVENIAN ARCHIVES

This section briefly presents data on the state of describing in Slovenian archives. The data, drawn from the ScopeArchiv software used by the Archives of the Republic of Slovenia and regional Slovenian archives, highlight the scale of backlogs and the lack of resources, further justifying the need to explore new methods and tools such as AI.

As of August 2025, the ScopeArchiv system contains a total of **3,455,690 records** with reference codes (i.e., archival units). Titles have been entered for **3,449,376 records (99.82%)**. The language(s) of the records are entered for **2,266,651 records (65.59%)**; script(s) are recorded for **2,032,542 records (58.82%)**; and the content description is filled for **825,316 records (23.88%)**.

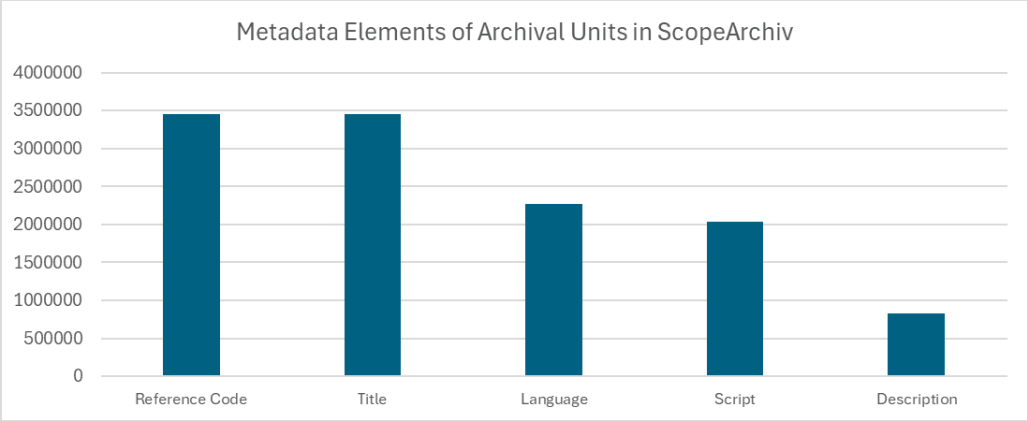

Figure 1: Metadata elements of archival units in ScopeArchiv.

As can be seen, the description field is filled with fewer than one quarter of all descriptive units. This does not necessarily mean that there is no information on content where this field is empty. Often, such information, if the record concerns a single document or a set of documents, appears at a higher level of archival hierarchy, such as in the series, collection, or fonds, or may be available in associated finding aids. Nevertheless, the data suggests that there is considerable room for improvement in content description.

Content descriptions are also of great value to users of the virtual, web-based, reading room (VAČ), many of whom struggle with search queries and are unfamiliar with the principles of archival description and data access.

3 ARTIFICIAL INTELLIGENCE IN ARCHIVAL WORK

3.1 A BRIEF OVERVIEW OF THE DEVELOPMENT OF ARTIFICIAL INTELLIGENCE

The broader development of AI began with attempts based on symbolic and rule-based systems, known as expert systems. These systems were designed to mimic human reasoning through predefined logical rules and knowledge bases. They were primarily developed to solve specific problems, for example in medicine or engineering (Brock, 2018).

Early systems were primarily designed for processing structured data. With the advent of machine learning, the focus of AI shifted from predefined rules to statistical methods, allowing computers to learn from large datasets. In this con-

text, Natural Language Processing (NLP) emerged, enabling computers to understand, interpret, and generate human language. NLP has proven to be a key technology for processing unstructured textual material, significantly reducing the limitations of early systems (Feng et al., 2024).

A revolutionary breakthrough in natural language processing came with the "transformer" architecture, which enabled the development of LLMs with hundreds of billions of parameters. This architecture made it possible for models to effectively process long sequences of data, resulting in advances in conversational AI, machine translation, text analysis, and text generation (Feng et al., 2024). These developments also open the door to the automation of archival processes such as transcription, cataloguing, description, and content analysis.

3.2 ARTIFICIAL INTELLIGENCE AND ARCHIVAL DESCRIPTION

In addition to scanning, Optical Character Recognition (OCR) forms a key foundation of most archival digitization projects, as it converts image-based material into machine-readable text. The quality of OCR is directly linked to the effectiveness of subsequent processes, including information extraction and search. LLMs build upon this foundation, not only recognizing text but also "understanding" and analyzing it within context.

Research on using LLMs for transcribing historical manuscripts indicates that they achieve significantly higher accuracy than traditional specialized software. However, challenges remain, particularly with low-resource languages. While some models have shown remarkable performance even in such situations, difficulties persist with older archival materials written in archaic language that are underrepresented in training datasets. This can result in inaccuracies in recognizing historical vocabulary and grammatical structures necessitating careful human verification (Kaluvilla et al., 2025; Khan et al., 2024; Humphries et al., 2024).

Nevertheless, the ability of large LLMs to combine concepts and expressions from multiple languages offers strong potential for working with multilingual archival fonds. Furthermore, LLMs perform well in summarization and named entity recognition, identifying key information such as topics, persons, places, and organizations (Zhang & Colavizza, 2025). These capabilities form the basis for automated archival description. Machine learning and NLP can also be lever-

aged to reveal connections within large corpora, thereby improving organization and contextualization of archival material (Cushing & Osti, 2023). Case studies where AI has been successfully introduced into different institutions suggest its suitability for automating archival descriptions (Arias Hernandez et al., 2024).

In addition, the integration of AI technologies promises progress in creating and enhancing metadata, which is crucial for ensuring accessibility of archival records. AI can reanalyze records when new material is added, enabling iterative updates and ensuring continued relevance (Fan et al., 2020). This is especially valuable in fast-changing situations, such as crises, where rapid action is required.

Despite clear advantages, concerns also exist regarding AI's use in archival description. The effectiveness of AI largely depends on the models being specifically trained for archival contexts. Given the opaque nature of many machine learning algorithms, explainability is also essential, particularly when establishing the context and meaning of records (Hou et al., 2022). These concerns are reflected in research on the ethical implications and potential biases embedded in AI processing frameworks, especially considering the complex nature of archival materials (Tenzer et al., 2024).

3.3 INTERNATIONAL PROJECTS AND ASSOCIATIONS

In recent years, several important international projects have emerged that focus on the research, ethical aspects, and practical application of AI in the management and accessibility of digital cultural heritage. These projects emphasize the crucial role of interdisciplinary collaboration among computer scientists, archivists, and humanists. Solutions for archival science are often developed within specialized communities, not solely by large commercial enterprises. **InterPARES Trust AI** is one of the fundamental projects addressing questions of AI usability and reliability in archival science. Ongoing studies include *Teachable AI for Arrangement and Description*. **AEOLIAN** and the related **AURA network** are focused on the challenge of "dark" digital archives, which remain closed to the public due to sensitive content, copyright restrictions, or other reasons. By leveraging AI-supported methods, they aim to selectively open collections and improve accessibility without requiring manual review of massive datasets. **LUSTRE** is a project focused on born-digital records, exploring the potential impact of AI on

these materials and on the work of archivists. **AI4LAM** is an open community initiative that serves as a hub for professionals and enthusiasts.

4 THE EXPERIMENT AND RESULTS

4.1 PURPOSE

The purpose of this experiment was to systematically test the capabilities of several publicly available LLMs in describing material from the Archives of the Republic of Slovenia, which may be the first such study in the Slovenian context. The rationale for this research stems from the following factors:

- The material is interwoven with texts in Slavic, Germanic, and Romance languages.
- Slovenian is a language underrepresented in large language models.
- Special emphasis was placed on older materials, which are also underrepresented in LLM training sets and are often difficult to understand for (external) users.
- To familiarize archivists with how currently popular AI services respond to users, so to better understand the issues and potentially adjust description strategies.

4.2 SELECTION OF MATERIAL

The material was chosen according to criteria that ensured independent replication of the study and legally and ethically unproblematic use. All material selected for the test is publicly available via VAČ and does not contain protected or sensitive data. This ensures that other researchers can repeat the analysis and test additional services.

Five descriptive units were selected for the test, each containing characteristics typical of real-world archival challenges (incomplete or missing OCR, mixed handwriting and print, text bleed-through, varied scripts and languages, archaic terminology, etc.).

1. **SI AS 730/2/1/1:** *Odloki, patenti, razglasi, okrožnice od 1725 do 1792* (Decrees, patents, proclamations, circulars, 1725–1792), file: SI_AS_730_2_1_1.pdf.
2. **SI AS 1073/II-37r:** *Vinogorski red iz leta 1543* (Viticultural Code, 1543), file: SI_AS_1073_495_(II-37r).pdf.

3. **SI AS 1080/1/2,3,7:** Three documents combined in one file:
   - Kapucinski provincial na Štajerskem, Koroškem in Kranjskem Silvester de Polcenico sprejme Burkharda Hitzinga [Hitzingkh] in njegovo ženo Sidonijo med duhovne otroke kapucinov, 1625 (Capuchin Provincial Silvester de Polcenico admits Burkhard Hitzing and his wife Sidonia among the spiritual children of the Capuchins, 1625), file: SI_AS_1080_I_2.pdf.
   - Papež Inocenc X. podeli odpustke bratovščini sv. Rešnjega telesa v župnijski cerkvi sv. Egidija v Višnji Gori, 1647 (Pope Innocent X grants indulgences to the Confraternity of the Holy Sacrament in the parish church of St. Giles in Višnja Gora, 1647), file: SI_AS_1080_I_3.pdf.
   - Kranjski deželni glavar Janez Gašper grof Cobenzl razsodi v zadevi zapuščine po pokojnem Francu Engelbrehtu pl. Zetschkerju, 1718 (Carniolan Governor Johann Caspar von Cobenzl decides in the inheritance case of Franz Engelbrecht von Zetschker, 1718), file: SI_AS_1080_I_7.pdf.
4. **SI AS 2048/HR DARI/2:** Zadeve mesta Reka (Rijeka), 1560-1714 (Affairs of the city of Rijeka, 1560–1714), file: SI_AS_2048_HR_DARI_2_273_08.pdf.
5. **SI AS 2048/IV/1:** Okrožnica Notranjeavstrijskega gubernija v Gradcu, s katero se zapoveduje, da cerkveni upravitelji ne smejo samovoljno ravnati s cerkvenim denarjem, 1788 (Circular of the Inner Austrian Government in Graz, ordering that church administrators may not independently manage church funds, 1788), file: SI_AS_2048_IV_1_in_Okroznica_in_1788.pdf.
6. **SI AS 2058/6:** Stenografske beležke Senata Kraljevine Jugoslavije, redni sklic, 16. rednega sestanka do 27. rednega sestanka, od 14. februarja 1933 do 10. marca 1933 (Stenographic records of the Senate of the Kingdom of Yugoslavia, regular sessions 16–27, February 14 – March 10, 1933), file: II-241_1933_knjiga_II.pdf.

Preprocessing included:

- Removing some pages (e.g., scanned covers with labels and notes) that could mislead recognition and analysis.
- Reducing the resolution of larger files to balance good OCR readability with processing speed.
- Removing descriptive metadata from file names, leaving only the reference code. This ensured that LLMs had to determine the content and context from the document itself, before potentially retrieving external information.

4.3 Selection of AI Services

The following services were tested, as they are currently popular and already widely used:

- **OpenAI ChatGPT 5 (CGPT5)** – Based on GPT-5, introduced mid-2025. Supports multimodal input (text, images, etc.), contextual analysis, and reduced hallucinations (https://chatgpt.com/)
- **Microsoft Copilot (COPLT)** – An AI assistant integrated into Office tools for document analysis and summarization, based on OpenAI and Microsoft technology (https://copilot.microsoft.com/).
- **Google Notebook LM (GNLM)** – A research tool that analyzes sources (PDFs, web pages, etc.), produces summaries, questions, and visualizations (https://notebooklm.google.com/).
- **Google AI Studio (GAIST)** – A platform for building multimodal AI apps, powered by Gemini 2.5 Pro and Gemma, emphasizing safety and responsible use (https://aistudio.google.com/).
- **Google Gemini (GGEM)** – A multimodal model (2.5 Flash/Pro), Google's direct competitor to OpenAI services, with browsing and integration capabilities (https://gemini.google.com/).

These services were chosen because they are widely used, multimodal, support multiple languages, and can perform OCR if needed. While GNLM, GAIST, and GGEM are all based on Gemini technology, they are optimized for different purposes (research, app development, general interaction).

4.4 COURSE OF THE EXPERIMENT

In this experiment, the goal was not precise measurability of correct and incorrect answers, since models are developing extremely quickly and such a detailed, time-consuming measurement would already be outdated by the time of publication. Instead, the emphasis was on a quantitative/qualitative approach and on establishing a reference baseline that will make it possible to monitor trends in suitability and performance in the future.

To ensure consistency and reproducibility of the experiment, the same protocols and questions were used for all services. The process began with a carefully considered prompt. Each LLM was first provided with a general instruction indicat-

ing expectations and specifying that the model should produce short, clear, and informative answers. This was followed by specific questions directing the LLM to extract precisely defined information, relevant for archival description. In this way, the outputs were oriented toward generating descriptive elements that could, without major difficulty, be mapped to a metadata schema such as ISAD(G).

The tasks were divided into two groups:

- **Technical tasks (1–6):** Determining the file name, number of pages, presence of text/images, assessment and execution of OCR.
- **Content tasks (7–16)[2]:** Language and script recognition, content summary, suggestion of a title, extraction of key entities (persons, places, time periods). Performance on content tasks was assessed by archivists with a four-point scale (from no good to fair), evaluations were divided into two categories: (A) usefulness/quality to the archivist describing the material and (B) perceived usefulness/quality to an external, non-expert user.

For each service we requested the use of the most up-to-date model available. Each service, in connection with each individual file, was given the same prompt in Slovenian language. Also, all responses were in Slovenian language.

The responses of all services for all tested files were copied into a single document, which grew to 168 pages, amounting to nearly half a million characters. Wherever possible, we enabled "deep reasoning" modes in the services, which in some services took up to 15 minutes to prepare a response. In the case of GNLM, we only asked questions 1–16 (without general instructions), as the service refused to accept the full prompt.

We partially edited the document and summarized the results in tabular form as meaningfully as possible. Summarization turned out to be extremely time-consuming and not without difficulties, since services sometimes returned not only the answer requested but also additional information or commentary, which occasionally enriched the response but sometimes made it ambiguous or even incorrect[3].

2 The processing of responses 12–16 has been reserved for subsequent articles due to their extent.

3 In a separate paper (in Slovenian language) in ATLANTI 2025 selected AI responses and tabulated results are published with running commentaries. In this paper only key findings are presented.

RESULTS FOR TECHNICAL QUESTIONS/TASKS

1. Question/task: State the name of the attached file

This question/task aimed to verify whether determining this parameter, which is important primarily for referencing, poses difficulties for the services. If AI processes only one file at a time, referencing is not problematic, but if it processes a bundle of files, it is very important that it always provides the correct file name. **General conclusion:** An apparently straightforward task showed that sometimes LLMs cannot access the file name. This fact must be considered particularly in cases where the file title itself is one of the key data elements. Such an example would be sensor measurement logs, where files are not necessarily equipped with appropriate headers.

2. Question/task: State the number of pages in the attached file

This question/task was intended to check whether determining this parameter, which is important above all for assessing the extent of material, causes difficulties for the services. **General conclusion:** A seemingly simple task demonstrated that even page counting, despite us explicitly asking for the number of pages in the PDF file, is not interpreted unambiguously by the tested LLMs. Problems occur when pages are also paginated. Some results are inexplicable; perhaps errors in the largest file relate to its extent (both in bytes and number of pages). Interestingly, some services sometimes returned the number of pages in words. Therefore, if we wanted services always to return data in numerical form (e.g., for automated procedures), we would apparently have to specify this explicitly.

3. Question/task: Identify on which pages there is text, and indicate page numbers without textual elements

This question/task was included to verify whether services can recognize document structure. This helps archivists in questions related to text extent and the ratio of text to graphic material, which is particularly pressing where OCR has not been performed, and it is difficult to determine the amount of text in characters and consequently the resources required for further processing. The third in the series of questions is essentially an extension of the first two, designed to test the usefulness of these services in determining the structure and form of scanned material. **General conclusion:** Even this supposedly simple task presented a hard challenge for many services. Possibly, in the case of the last file, the size of the

material exceeded the input capacity of the LLM, or the services failed to call the appropriate agents to perform the task. Yet at least one of the services completed it excellently. We may hope it is only a matter of time before the others follow.

4. Question/task: Identify on which pages there are images or photographs, if the file contains them

This question/task was posed for reasons like the previous one. There are situations, especially with long documents, where it is useful to know whether they contain visual material and how much. This allows an assessment of the necessary procedures for further processing, for example, preparing material for people with visual impairments, where documents must be supplemented with metadata for assistive tools or otherwise interpreted. **General conclusion:** When processing the results, it became evident that this question had been phrased somewhat clumsily. Perhaps we should have asked about the presence of graphic elements (photographs, images, illustrations, symbols, etc.). Nevertheless, the services performed the task relatively well. Once we received the answer that it was not clear whether the material contained images, once a service appeared to take a shortcut (trying to count image references). GAIST responded most consistently and correctly, drawing attention to aspects overlooked by the other services.

5.–6. Question/task: Does the file contain text recognized with OCR? Assess OCR quality. Perform OCR if necessary (for remaining tasks)

This question/task was included because searching archival material is much more effective if one can also search within the contents of archival units and not only in the description data prepared by the archivist. Materials are being digitized rapidly in many archival institutions, but OCR quality (when performed at all) varies widely. The hardest cases are manuscripts, older scripts, and older languages. Very few archives find the resources (staff, money, and time) to manually correct OCR. A rough estimate of OCR quality can be made by inspecting the extracted text; if it contains many garbled words, OCR is likely poor. A higher-quality evaluation requires re-running OCR with a better tool (or do it manually) and comparing results. We wanted to see how the services would tackle this task. **General conclusion:** The combined volume of responses concerning OCR presence, its quality, and the instructions to perform OCR (for the remaining tasks) if missing, amounted to over 25,000 characters, much more than expect-

ed. Services often ignored handwritten portions without OCR, focusing only on printed text with partial OCR. Where no OCR was present, they were confused, sometimes asserting that deficient OCR already existed. Some claimed to perform new OCR or corrections, others reported uncertainty whether they had executed OCR or not. Statements of confidence were ambiguous. Interestingly, some services tried to "fix" poor OCR (either preexisting or done by the service itself) by guessing from context, sometimes quite successfully. Overall, the detection of OCR presence and quality was highly inconsistent, and the services' self-assessment of confidence often did not align with actual performance.

7.–11. Content Questions/tasks

In this section, we sought to evaluate the adequacy of responses to questions concerning language, composition, content, and the suggested title for the material in the files. All five services responded to each of the six files with answers to the following questions:

- State the languages in which the text is written, if you recognize them.
- Does the file represent a single document, or does it contain multiple documents? How many, if more than one?
- Summarize the content of the file in a few paragraphs.
- Suggest a title or name that would be appropriate for the file if it were intended for a historian.

The services responded to these questions with varying levels of detail. We compiled all responses into a single document of about 40 pages (approximately 85,000 characters) and standardized the formatting to make it easier for archivists to conduct the evaluation. The quality of the answers was reviewed by archivists familiar with the material. Their task was not to determine all correct and incorrect statements in detail, but rather to provide two ratings for all four questions together, for the material assigned to them for assessment[4]. We asked them to rate (A) usefulness of the AI answers in assisting an archivist in descripting and (B) perceived usefulness of the AI answers in assisting an external, non-expert user. Every AI response was rated by two archivists (scale: 1 = no good, 2 = a little, 3 = fairly, 4 = very; "–" = the service could not process the file; format is A / B).

4 I would like to thank my colleagues Danijela Juričić Čargo, Žiga Koncilja, Branko Radulovič, Jure Volčjak, and Lilijana Žnidaršič Golec for their invaluable assistance and dedicated effort in evaluating the AI responses.

Table 1: Average ratings for questions/tasks 7–11.

File / Service	CGPT5	COPLT	GNLM	GAIST	GGEM
SI AS 730/2/1/1	2.5 / 2.5	2.5 / 2.5	3.0 / 2.5	3.0 / 2.5	3.0 / 2.5
SI AS 1073/II-37r	–	3.0 / 2.5	2.5 / 2.5	3.0 / 2.5	3.0 / 3.0
SI AS 1080/1/2,3,7	2.0 / 2.0	1.5 / 1.5	1.5 / 1.5	2.0 / 2.0	1.5 / 1.5
SI AS 2048/HR DARI/2	1.0 / 1.0	1.0 / 1.0	1.5 / 1.5	1.5 / 1.5	1.5 / 1.5
SI AS 2048/IV/1	3.0 / 3.0	3.0 / 3.0	2.5 / 2.5	3.5 / 3.5	3.0 / 3.0
SI AS 2058/6	3.0 / 3.0	–	3.0 / 3.5	3.0 / 3.5	3.5 / 3.5

In four cases, the AI was of no help at all (score 1). In twelve cases, it approached the threshold of being of little help (score 1.5). Twice it reached the value "a little helpful" (2), four times "fairly helpful," thirteen times between "a little" and "fairly helpful," sixteen times "fairly helpful," and seven times between "fairly" and "very helpful." The ratio of responses where it was not helpful to those where it was helpful is 1:2.5. The GAIST service received the highest number of top ratings.

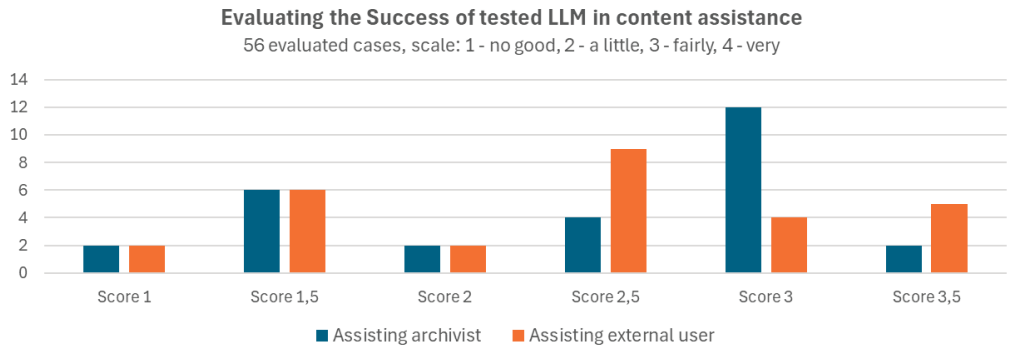

Figure 2: Chart showing the distribution of average ratings.

The average ratings by evaluated files: SI AS 730/2/1/1 (2.65); SI AS 1073/II-37r (2.75); SI AS 1080/1/2,3,7 (1.7); SI AS 2048/HR DARI/2 (1.3); SI AS 2048/IV/1 (3.0); SI AS 2058/6 (3.25). It is immediately evident that the best ratings were achieved by files that either already had OCR performed or where the service was able to successfully carry it out on its own.

5 DISCUSSION AND CONCLUSION

5.1 KEY FINDINGS

The results of the experiment clearly demonstrate that in tasks requiring technical precision, such as counting the number of pages, identifying blank or image-only pages, or determining the presence and quality of OCR, the services performed

poorly. Errors in these cases were frequent, often contradictory, and in many cases misleading. Such findings are highly significant because they indicate that archivists cannot rely on tested LLMs for these aspects of descripting, as the risk of incorrect data entry is too high.

By contrast, in content-related tasks the services performed considerably better. The ability to detect languages, summarize content, suggest titles, and extract key entities was judged positively. The services were shown to be useful both for archivists, as a support in creating initial descriptions, and for external users, who could gain a quick and reasonably accurate overview of the content. Nevertheless, the descriptions produced by tested LLMs are not flawless; they often contain errors, simplifications, or interpretative shifts. Therefore, they cannot replace professional descriptions, but they can be a valuable supplement.

The evaluations also revealed that the main factor influencing the quality of results is not the service itself but the type of document. Modern, clearly printed or those with OCR already performed, consistently received better ratings. On the other hand, older, handwritten, or multilingual documents with complex historical content received the lowest scores.

The comparison of the five tested services further showed that differences between them exist but are not decisive. In general, all services followed the same pattern: good results for modern and clear materials, poor results for older and complex ones. Individual deviations (e.g., GAIST receiving the highest number of top ratings) do not outweigh the overall trend.

5.2 LIMITATIONS OF THE STUDY

It is necessary to emphasize that this study has limitations. First and foremost, the experiment was limited to six files, which, although carefully chosen to represent a spectrum of challenges, cannot fully capture the entire diversity of archival material. For a comprehensive assessment, a significantly larger sample would be needed, covering a broader range of periods, languages, and types of records.

Second, the field of AI is developing extremely quickly; models are being updated or replaced every few months. The results therefore reflect the state at the time of testing and may soon become outdated. Future replications of the experiment with newer generations of services are therefore essential.

Third, the assessment of usefulness was carried out with the participation of a limited number of archivists. Although they were professionals familiar with the material, a larger and more diverse group of evaluators could provide an even more balanced assessment.

Finally, the experiment did not include a systematic analysis of errors or biases in AI-generated descriptions. While we noted examples of inaccuracies, this was not the focus of the study. A more detailed study of the types of errors, their frequency, and potential systemic causes would be an important step for future research.

5.3 DIRECTIONS FOR FURTHER RESEARCH

Based on the findings and limitations identified, several key directions for further research can be outlined:

1. Future studies should include a larger number of files that cover different historical periods, languages, scripts, and types of archival records. Particular attention should be paid to materials that are especially challenging for AI (e.g., manuscripts, mixed-language documents, records with archaic terminology).
2. Given the rapid development of AI, it will be crucial to repeat similar experiments regularly with newer models and services. This will make it possible to monitor progress over time and assess whether models are becoming more reliable and suitable for archival use.
3. More detailed research should be devoted to identifying and categorizing errors in AI-generated descriptions. This would allow a clearer understanding of where and why AI fails and how archivists can prepare to recognize and correct such errors.
4. Future work should explore how AI outputs can be more directly mapped to established metadata standards (e.g., ISAD(G), RiC). This would facilitate more effective integration of AI-generated descriptions into archival information systems.
5. It would also be useful to examine how different groups of users (professional archivists, historians, students, public) perceive and use AI-generated descriptions. Such studies could provide valuable insight into which functions are most useful and how AI can improve access to archival heritage.
6. While this study focused on general-purpose LLMs, it is reasonable to expect the emergence of specialized models trained specifically for archival and his-

torical material. Research should therefore also monitor this area and test the extent to which such models are more effective.

REFERENCES

AEOLIAN Network – Artificial Intelligence for Cultural Organisations. (s.d.). Retrieved from https://www.aeolian-network.net/ (accessed on 11. 9. 2025).

AI4LAM. (s.d.). Retrieved from https://sites.google.com/view/ai4lam (accessed on 11. 9. 2025).

Arias Hernandez, R., Fewster, K., & Penniman, S. (2024). Artificial Intelligence and Machine Learning Competencies for the Archival Professions. *Proceedings of the Association for Information Science and Technology*, *61*(1), 36–43. Retrieved from https://doi.org/10.1002/pra2.1006 (accessed on 11. 9. 2025).

AURA Network – AURA Network: Archives and AI in the UK and Ireland. (s.d.). Retrieved from https://www.aura-network.net/ (accessed on 11. 9. 2025).

Bingham, N. J., & Byrne, H. (2021). Archival strategies for contemporary collecting in a world of big data: Challenges and opportunities with curating the UK web archive. *Big Data & Society*, *8*(1), 2053951721990409. Retrieved from https://doi.org/10.1177/2053951721990409 (accessed on 11. 9. 2025).

Brock, D. C. (2018). Learning from Artificial Intelligence's Previous Awakenings: The History of Expert Systems. *AI Magazine*, *39*(3), 3–15. Retrieved from https://doi.org/10.1609/aimag.v39i3.2809 (accessed on 11. 9. 2025).

Cushing, A. L., & Osti, G. (2023). "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. Retrieved from https://doi.org/10.1108/JD-08-2022-0170 (accessed on 11. 9. 2025).

Decker, S., Kirsch, D. A., Kuppili Venkata, S., & Nix, A. (2022). Finding light in dark archives: Using AI to connect context and content in email. *AI & SOCIETY*, *37*(3), 859–872. Retrieved from https://doi.org/10.1007/s00146-021-01369-9 (accessed on 11. 9. 2025).

Fan, L., Yin, Z., Yu, H., & Gilliland, A. (2020). *Using Machine Learning to Enhance Archival Processing of Social Media Archives*. LIS Scholarship Archive. Retrieved from https://doi.org/10.31229/osf.io/gkydm (accessed on 11. 9. 2025).

Feng, C., Li, Y., Chen, Z., & Guo, L. (2024). The Evolution and Breakthrough of Natural Language Processing: The Revolution from Rules to Deep Learning. *Proceeding of the 2024 5th International Conference on Computer Science and Management Technology*, 307–311. Retrieved from https://doi.org/10.1145/3708036.3708089 (accessed on 11. 9. 2025).

Hankins, R. (2019). Information Literacy and Instruction: Embracing Informational and Archival Literacies: Challenges and Successes. *RUSQ: A Journal of Reference and User Experience*, *58*(3), 153–157. Retrieved from https://doi.org/10.5860/rusq.58.3.7042 (accessed on 11. 9. 2025).

Hou, Y., Kenderdine, S., Picca, D., Egloff, M., & Adamou, A. (2022). Digitizing Intangible Cultural Heritage Embodied: State of the Art. *Journal on Computing and Cultural Heritage*, *15*(3), 1–20. Retrieved from https://doi.org/10.1145/3494837 (accessed on 11. 9. 2025).

Humphries, M., Leddy, L. C., Downton, Q., Legace, M., McConnell, J., Murray, I., & Spence, E. (2024). *Unlocking the Archives: Using Large Language Models to Transcribe Handwritten Historical Documents* (Version 1). arXiv. Retrieved from https://doi.org/10.48550/ARXIV.2411.03340 (accessed on 11. 9. 2025).

InterPARES Trust AI - Artificial Intelligence. (s.d.). Retrieved September 11, 2025, from Retrieved from https://interparestrustai.org/trust/about_research/workinggroups

ISAD(G): General International Standard Archival Description - Second edition. (s.d.). *ICA*. Retrieved from https://www.ica.org/resource/isadg-general-international-standard-archival-description-second-edition/ (accessed on 11. 9. 2025).

Kaluvilla, B. B., Kalarikkal, S. A., & Thamilvanan, G. (2025). AI-driven extraction and intelligent retrieval of missionary archives in Malabar: Advancing preservation and accessibility with machine learning. *Performance Measurement and Metrics*, 1–15. Retrieved from https://doi.org/10.1108/PMM-02-2025-0008 (accessed on 11. 9. 2025).

Khan, A., Rai, U., Singh, S. S., Yamamoto, Y., Ibarreche, X. G., Meadows, H., & Gleyzer, S. (2024). OCR Approaches for Humanities: Applications of Artificial Intelligence/Machine Learning on Transcription and Transliteration of Historical Documents. *Digital Studies in Language and Literature*, *1*(1–2), 85–112. Retrieved from https://doi.org/10.1515/dsll-2024-0013 (accessed on 11. 9. 2025).

LUSTRE – Unlocking our Digital Past with Artificial Intelligence. (s.d.). Retrieved from https://lustre-network.net/

Pezzica, L. (2023). Archival inventories as a profession. *JLIS.It*, *14*(3), 64–71. https://doi.org/10.36253/jlis.it-563 (accessed on 11. 9. 2025).

Records in Contexts (RiC). (s.d.). *ICA*. Retrieved from https://www.ica.org/ica-network/expert-groups/egad/records-in-contexts-ric/ (accessed on 11. 9. 2025).

Tenzer, M., Pistilli, G., Bransden, A., & Shenfield, A. (2024). Debating AI in Archaeology: Applications, implications, and ethical considerations. *Internet Archaeology*, *67*. Retrieved from https://doi.org/10.11141/ia.67.8 (accessed on 11. 9. 2025).

Zhang, S., & Colavizza, G. (2025). *Named Entity Recognition of Historical Text via Large Language Model* (Version 1). arXiv. Retrieved from https://doi.org/10.48550/ARXIV.2508.18090 (accessed on 11. 9. 2025).