"Hey Google, Help Me do a Citation Analysis": Incorporating AI in the Citation Analysis Process to Identify Resources and Resource Types

Sarah G. Park and Lisa Romero

abstract: Citation Analysis is a use-based methodology that provides insight and can inform librarians' collection decisions. However, citation analysis can be labor-intensive and time-consuming. In addition, it primarily focuses on traditional library materials such as books, journals, book chapters, and conference proceedings, while often overlooking materials such as datasets, software, government documents, media, performances, and so on. This study integrates artificial intelligence (AI) in the citation analysis process to label citations according to resource type so that librarians might identify what resources scholars are using and citing. In addition to describing how to integrate AI into citation analysis, the authors evaluate the methodology's success, with an overall accuracy score of 97.16 percent.

Introduction

Librarians are often contacted by students and researchers about types of resources that will help them accomplish a specific research goal. If their library does not provide access, the librarian must then seek out possible resources for purchase. Collection development is a process that considers patrons' information needs, availability of resources, cost, format, and other factors. While librarians have access to tools such as publisher websites, selection platforms like GOBI, and book

portal: Libraries and the Academy, Vol. 26, No. 3 (2026), pp. 461–494.

catalogs, some resource types like datasets, software, government documents, and gray literature are underrepresented. An additional factor impacting collection development for these "non-traditional" resources might be a lack of librarian familiarity with what is available and what is used most by researchers. If librarians are not familiar with a specific type of resource, their collection development efforts may not sufficiently meet researchers' needs. Librarians need access to more comprehensive information and a more structured process that reveals what resources (representing all types) are used by researchers within specific subject area(s), to ensure that library collections reflect all types of resources used in research. The methodology proposed in the current study provides librarians with a means to make better informed collection decisions while meeting the researchers' needs.

A citation serves as a link between earlier and later works, revealing the sources used in the preparation of a text and enabling readers to locate those sources.[1] References appear at the end of a work and follow a "who-when-what-where format" so the author can give credit to their sources and provide a path for future readers to retrieve the sources.[2] Citation analysis has been utilized in a variety of ways, from quantifying and measuring the impact of sources and authors to evaluating scholarship in appointment, tenure, and promotion reviews.[3] For librarians, citation analysis has also been used as a tool for collection development, enabling assessment of the quality of library collections for whether resources adequately support specific disciplines.[4] The approach is insightful because it is often used to rank materials according to how many times they have been cited. Many believe that "high citation counts mean a statistical likelihood of high-quality research."[5] Because citations are use-based, they serve as a strong indicator, or reliable tool of what is used and, more importantly, if conducted regularly, could identify usage trends for different resources and resource types.

Given the complexity of scholarly literature and limited access to computational expertise, librarians often rely on proprietary databases such as Web of Science (WoS) and Scopus to analyze citations. However, these databases also have limited coverage and features that may not capture the full scope of scholarly output. They are designed primarily for search and discovery, focusing on indexing select journals. They often exclude a substantial body of relevant but unindexed work, such as non-journal and non-English publications.[6] Moreover, they provide greater coverage in natural science, engineering, and biomedical research than social sciences and arts and humanities.[7] As a result, they may have limitations in tracking and measuring scholarly impact across a wide range of resource types and disciplines that may be especially valuable to specific institutions and programs. Moreover, the databases' predetermined subject classification systems may fail to recognize small and multidisciplinary fields as distinct subject areas, resulting in a lack of field baselines for bibliometric analyses.[8] For example, actuarial science literature spans several major disciplines, including mathematics, statistics, finance, and economics, and is not in a single subject classification. In the authors' own experience, researchers wanted exhaustive lists of datasets, and the author was able to make several recommendations but was unable to meet the researcher's demand for comprehensiveness. These limitations underscore the need for a collection development methodology that will identify resource types and titles and offer solutions applicable to various subjects.

Artificial Intelligence (AI) is a bleeding-edge field that often lacks clear and commonly accepted definitions as researchers and practitioners continue to address its evolving challenges. Therefore, the authors attempt to define key terms in AI to provide a set of generally accepted concepts. AI is broadly defined as a field of study that aims to develop algorithmic methods capable of performing tasks that typically require human intelligence.[9] Machine learning (ML), a subset of AI, focuses on mimicking the human learning process by training algorithms on data to make predictions based on statistical analysis. Deep learning (DL), an advanced form of ML, utilizes complex layers of computation and inference to simulate human reasoning. Large language models (LLMs), such as OpenAI's ChatGPT and Google's Gemini, are advanced ML/DL models trained on massive datasets. These models enable machines to learn complex language patterns and interact with humans in a natural and conversational manner.[10] While LLM companies provide a chat interface for users to engage in conversations with AI models, they have also revolutionized AI engineering by offering access to foundation models via Application programming interfaces (APIs). Foundation models are "trained on broad data at scale such that they can be adapted to a wide range of downstream tasks" and support general-purpose AI applications across diverse domains.[11] This approach differs from traditional machine learning, which typically requires dataset collection, algorithm selection, model training, and iterative evaluation and refinement. By employing a foundation model, developers can bypass these model-building steps, thereby reducing the time and effort required to build an AI model from scratch.[12]

Building on the advancements in AI and growing accessibility to computational tools, tasks that were once only dreamed of are now becoming possible. For example, a layperson can take a photo of a flower or tree using a mobile app, which then identifies and returns the species of the plant. This raises the question: Can we develop an application that reads article references, analyzes the data, and extracts the necessary information from them? The authors of this study investigate a potential solution involving AI and determine whether AI can be used in citation analysis methodology to examine references and predict resource types in a subject, with the intention of meeting the researcher's needs and exploring the possibility of applying this new method to the collection development process. They will also evaluate how accurately AI predicts resource types. It is important to note that the terms "publication type," "resource type," and "material type" are used interchangeably throughout the literature. In this paper, the authors will use the term "resource type." However, when citing other works, the authors will use the term originally used by the cited source.

Can we develop an application that reads article references, analyzes the data, and extracts the necessary information from them?

Literature Review

Research and publishing are integral activities across all academic disciplines and are considered essential within academia. Librarians, through their collection development

efforts, support these important activities by providing the necessary resources that enable research and publishing. In fact, the Research Information Network found a direct relationship between an institution's library and its research performance. More specifically, easy access to high-quality resources directly impacts quality research and, when the library and researchers work in tandem, the results are better library service and top researchers.[13]

As Peggy Johnson explains, citation studies are a type of bibliometrics or collection analysis technique and are used primarily in academic and research libraries. Data from citation analysis can guide subscription, cancellation, and retention decisions. They are particularly useful in collections where journals are important and are used to develop core lists of primary journals, identify candidates for cancellation or storage, and identify trends in the literature and users' information-use behavior.[14]

Examination of Resource Types

Acquiring and providing access to resources deemed essential is often accompanied by numerous challenges for librarians. First, there is a seemingly endless number of resources available for purchase. In addition, the number of and preference for resource types used in scholarly research is also diverse and varies among the disciplines. In fact, numerous studies examining referencing behaviors across disciplines have identified up to 36 different types of resources used within disciplines.[15] Studies by Erika Alves dos Santos, Silvio Peroni, and Marcos Lui Mucheroni examine referencing behaviors across disciplines and identify what types of resources were used in each of the disciplines.[16] Svein Kyvik conducted a longitudinal study examining the publication behavior of faculty at a Norwegian university in 1982, 1992, and 2002 to identify trends in faculty publications by discipline.[17] In Mary C. Schlembach's bibliometric analysis of 12,065 references in chemistry and physics doctoral dissertations, they examined a random selection of 50 dissertations from each subject, spanning the years 1970 to 2020, to learn what resource types were most often cited.[18] Elina Late et al. studied the reading and writing behaviors of scholars in Finland, including the types of resources they used in their subject disciplines. The study concluded that scholars often read and cite non-academic resources beyond journal articles and books.[19] Studies have generally found that researchers in the sciences predominantly rely on journals while researchers in the humanities use more books and book chapters and tend to use broader types of resources for their research.[20] Relevant to the current study, Late points out that "there here has been a lack of studies taking into account different types of general publications such as newspapers, professional magazines, non-fiction, fiction or blogs."[21] All of these findings support the notion that many resource types exist, and the types of resources used among the disciplines varies. They can also inform academic librarians because they provide details on the broad range of resource types used in scholarship. More importantly, the studies affirm the challenges librarians face in their collection development efforts.

Librarians in the humanities, social sciences, and sciences have examined the diversity of resource types within their fields and/or have analyzed data representing resource citation patterns. Their studies often inform their collection management decisions relating to specific resources like journals or, on a broader scale, the variety of resource

types. To accomplish this goal, they usually rely on citation data. Studies examining the use of journals is more common because many tools are available to assist with finding citation data.[22] Mu-hsuan Huang and Yu-wei Chang conducted an extensive review of the literature documenting citation analyses conducted to determine resource types used across 17 subject areas in the humanities and social sciences. While they provided percentages of resource types cited within the various subjects, the data is limited to two resource types: books and journals. Their analysis does not include other resource types.[23] In their study of all references in psychology publications from German speaking countries, Günter Krampen, Peter Weiland, and Jürgen Wiesenhütter found limitations within the citation databases used for citation analysis. They determined that the limitations (selectivity) hindered evaluation of the different resource types, resulting in underrepresented scholarly work.[24]

Assessing Citation Tools

As mentioned, citation databases provide insight for data-driven collection development, with librarians often relying on databases such as Web of Science (WoS) and Scopus, to provide extensive citation data for journals.[25] WoS and Scopus represent the major citation databases for general-purpose scientific literature, including journal articles, conference proceedings, and books.[26] Both provide data for thousands of journal titles, representing decades of citations.[27] The data reveal which resources are used most or least, what subject areas are represented by the journals, and can indicate trends.[28]

Given the importance of considering all resource types when doing collection development, an obvious first step in the effort to determine their use in scholarly research would be to examine citation databases' coverage of the various resource types. Researchers have found that citation database, including WoS' and Scopus', coverage of other resource types is limited, and their analyses are biased toward journals. In their 2021 comprehensive comparison of citation databases, primarily focusing on WoS and Scopus, Raminta Pranckutė explains that WoS and Scopus both offer wide coverage of the highest quality journals, along with the additional tools for citation analysis. However, Pranckutė points out that while the two databases include coverage of conferences, books, patents, and trade publications, recent studies do not indicate any improvement in the coverage of books and conference proceedings, thus coverage of books and conference proceedings is still insufficient.[29] When comparing OpenAlex, WoS, Scopus, Pubmed, and Semantic Scholar, Nick Haupka, et al. found inconsistencies in resource and document type classification across all five databases and noted that the typologies used among the five databases was unclear. Understanding their differences is essential to selector decision-making.[30] Additional studies examined the strategies citation databases employed to assign resource types and found that these strategies differ and their accuracy is uncertain. More specifically, an analysis by Paul Donner found that 17 percent of resources in WoS had incorrect document type classification, with another analysis by Yu V. Mokhnacheva comparing document types of 3,843 resources in WoS and Scopus, finding differences in typification of resources from the databases compared to the resources' publisher websites.[31] The databases' primary focus is journals. Therefore, with regard to resource types other than journals, citation databases' use for citation analysis is

concerning and problematic because their selectivity "hinders" evaluations of the citation success of resource types such as books, chapters, journal articles, and non-traditional resources.[32] Additional studies also suggest that reliance on citation analysis databases can "hinder and complicate" evaluation of different languages of publications thereby resulting in "unfairness in scientometric evaluations."[33]

AI and Collection Development

The library and information science literature focusing on the integration of AI in libraries is mostly limited to library services. Related to service, Adebowale Jeremy Adetayo discusses the application of AI chatbots interacting with humans for reference consultations with the benefit being the ability to assist patrons at any time. They assessed that AI's contribution to collection development could involve ChatGPT. Specifically, a dataset of the library's holdings and details on its users could be used to train ChatGPT so that it could provide purchase recommendations.[34] On a larger scale, Muhammad Asim, Muhammad Arif, Muhammad Rafiq, and Rafiq Ahmad investigated the applications of AI in the university libraries in Pakistan, including library services, circulation, and collection development. The goal of their survey was to inform library administrators regarding the integration of AI in libraries. They found potential benefits and challenges of AI implementation in Pakistani libraries, and that library directors believe AI is expected to have a significant impact on library services and cataloging.[35] A more in-depth discussion of AI's potential application in academic libraries is provided by Andrew Cox. They discuss potential approaches to AI for knowledge discovery. One approach is offering the library collection as data for AI. However, Cox does not specify anything related to citation analysis.[36] More specific to archival collections, Sara Mannheimer et al. conducted a literature review on how AI is being used in library and archive practices. They identified 89 publications and found that AI is most often used for metadata extraction, reference and research services, but not for the purpose of collection development.[37] Alesia Zuccala, Maarten van Someren, and Maurits van Bellen experimented with machine learning to classify book reviews published in journals. They trained an intelligent system to recognize positive or negative reviews to inform librarians' collection development.[38] Ivan Portillo and David Carson evaluated the potential of four generative AI models in aiding health science librarians with collection development, specifically identifying collection gaps and recommending book titles. They found that LLMs are not yet suitable for information retrieval in the collection development process.[39] Ross Hanney applied Python scripting and machine learning algorithms to determine whether patrons save money by using public library materials instead of purchasing them. While the study applied machine learning, the purpose was not related to resource types or collection development in academic libraries.[40] Closely related to the current study is Schlembach's longitudinal analysis of chemistry and physics doctoral dissertations. While their classification of resource types relates to the current study, they did not use AI in their methodology. To date, nothing in the literature addresses the use of AI for citation analysis in the identification of resource types to support collection development.

Purpose

Issues relating to resource types and collection management efforts present challenges and impact librarians' efforts to build collections that meet the diverse needs of all researchers. Librarians managing collections need to keep appraised of what resource titles and types are used and most appropriate for all subject areas and resource types, not only for traditional resources such as books, journals, and newspapers. Selectors must also pay attention to emerging and nontraditional types, such as videos, digitized primary resources, photographs, advertisements, video games, datasets, maps, testing material, music, and preprints. It is important that the collection development process supports research and information needs while optimizing the library budget. Understanding trends of resource use by resource type helps librarians appropriate the budget according to needs. For example, if trends demonstrate a stronger need for datasets versus journals or other types in general, then librarians have use-based data to back up collection decisions.

Understanding trends of resource use by resource type helps librarians appropriate the budget according to needs.

If WoS and Scopus are lacking in citation data for all types of resources, and their classification of resource types is unreliable, then librarians are without tools to conduct reliable and comprehensive analyses. This especially impacts librarians who manage collections in the humanities and social sciences. While it is possible to perform citation analyses manually, it could prove time-consuming and labor-intensive. This issue would be to the detriment of regular evaluations.

Library selectors need a methodology that will provide a comprehensive list of what is cited within a subject or discipline that includes resource type and title information. As Peggy Johnson points out, citation analysis methodology focuses on analysis of journals and is used to develop "core lists of primary journals," but the proposed method is intended to expand analysis to broader resource types to help librarians develop core lists of primary books, primary datasets, preprints, grey literature, and so on, and inform efforts to guide users to more possibilities of resources. One goal of the current study is to develop a method for using AI to categorize references according to resource types, in hopes that the method could also be used in citation analysis for subjects not well represented in citation databases. The authors relied on Scopus because it accommodated the harvest of the large number of references within a specific subject area required to create a dataset for the AI model. Scopus also included coverage of the two actuarial science journals for the desired years' coverage. Therefore, the goals of the current study are to examine and determine the potential for AI to serve as a tool in citation analysis that can examine references and classify resource types. It is important to note that while the study applies its methodology to the discipline of actuarial science, the results are intended to guide practice in collection development generally. The following questions guided the research for this study:

- Can AI be used in citation analysis methodology to examine references and predict the resource types used in a subject?
- How accurately does AI predict resource types?

Methodology

This study employs a five-step methodology:

(1) **identification** of journal titles,

(2) **harvesting** references,

(3) **determining** a list of resource types,

(4) **classifying** resource types, and

(5) **review** and analysis.

The final evaluation step also involves applying the refined methodology to a new test dataset to assess its performance and replicability. These steps are illustrated in Figure 1, and a more detailed discussion of the evaluation process is provided in the Review and Analysis section.

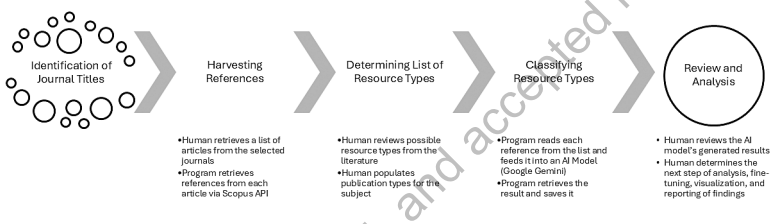

Figure 1. The five-step methodology for identifying resource types within a dataset.

Journal Title Identification

To identify top journals, the authors first consulted Clarivate's *Journal Citation Report (JCR)* and *SCImago Journal & Country Rank* for journal rankings and impact factor data. Unfortunately, the two databases lack a specific category for actuarial science. Instead, journals related to actuarial science are categorized under various subject areas, including decision sciences, economics, econometrics, finance, and mathematics. Consequently, the rankings of the journals are determined by the impact factors of other journals within their respective categories rather than by the subject area of actuarial science. Therefore, the researchers consulted a faculty member from the actuarial science program to identify leading journals in the field. Two journals were identified: the *North American Actuarial Journal* (*NAAJ)* and the *Journal of Risk and Insurance (JRI)*.

NAAJ began publication in 1997 and is a premier publication of the Society of Actuaries. It focuses on "domestic and international problems, interests and concerns of actuaries, their customers and public policy decision-makers."[41] *JRI* has a longer history than *NAAJ*, launching in 1933 as the *Proceedings of the Annual Meeting* (of the American Association of University Teachers of Insurance). In 1937, the title was changed to the *Journal of the American Association of University Teachers of Insurance*, which it retained until

1957. Following a merger with the *Review of Insurance Studies*, the journal was renamed the *Journal of Insurance* and was published under this title from 1957 to 1963. Since 1964, it has maintained its current title.[42] It is the "flagship journal of the American Risk and Insurance Association (ARIA)," focusing on theoretical and empirical research in insurance economics and risk management.[43]

To determine the research timeframe for the articles in these two journals, the authors consulted the existing literature and identified a leading researcher in citation analysis within actuarial science. L.L. Colquitt published five journal articles on citation analysis of risk and insurance journals in 1997, 1999, 2003, 2008, and 2017.[44] With the exception of the 1999 article, which examined articles published from 1987 to 1996, all of his studies analyzed publications from a five-year period. Following this series of work, the authors of the current study initially intended to conduct a citation analysis covering a five-year span from 2020 through 2024. However, because this period was directly affected by the Covid -19 pandemic, which may have influenced scholarly referencing and citation behavior, the authors decided to extend the analysis period to ten years, from 2015 to 2024.

Harvesting References

The process of harvesting references from the journals included identifying a source or a database for the journal articles and using that database to access and collect bibliographic information. Two abstracting and indexing (A&I) databases, WoS and Scopus, were evaluated for their coverage. Scopus was ultimately selected because its index of *NAAJ* began in 1997, whereas WoS only began indexing in 2019 and therefore does not fully cover the study's period (2015–2024). Table 1 lists the number of publications and references for the two journals during the period 2015 to 2024. A total number of 31,910 references were collected across both titles. Table 2 summarizes the history and indexing information of the two journals.

Once ten years' worth of articles' citation data from the two journals were collected from Scopus, the articles' reference lists were exported into a CSV file. Scopus provides an option to download the references cited within each article, but Microsoft Excel has a limitation of handling a maximum of 32,767 characters per cell, a limit that is often exceeded by lengthy references. To address this issue, a Python script was developed using the pybliometrics library to retrieve references by querying articles using their EID, a unique identifier in Scopus. [45] The EID of each journal article served as a primary key for its cited references, allowing the authors to merge the data and enable future analysis of the relationships between articles and their references. This script extracted specific reference elements via the *Scopus* API, including author names, article title, publication year, source title, Scopus ID, and the full reference text. All of this information was then saved into an Excel file.

Creating a List of Resource Types

Creating a clear list of resource types is a crucial step in guiding AI models to classify resources accurately and select the appropriate type without ambiguity. Therefore, it is essential to prepare a comprehensive list of resource types that represent the resources cited within the discipline. The authors needed to establish a list of resource types relevant

Table 1.

Number of publications and references from two actuarial science journals (2015-2024)

Journal Title	Number of Publications	Number of References
Journal of Risk and Insurance (JRI)	367	16,913
North American Actuarial Journal (NAAJ)	378	14,997

Table 2.

History and indexing information for two actuarial science journals, as reported from WoS and Scopus

Journal Title	First Year of Publication	WoS Coverage	Scopus Coverage
JRI	1933; 1964-	1966-	1978-
NAAJ*	1997-	2019-	1997-

* NAAJ was first indexed in WoS in 2019 so citations were harvested from Scopus instead of WoS.

to actuarial science which could be used in the experiment and subsequent evaluation to determine the methodology's effectiveness. They consulted four lists of resource types. One list of 46 types was provided by a colleague researching campus publications. Three other lists were available from Scopus, WoS, and PubMed. The authors selected resource types based on two criteria: types that related to actuarial science and types found on a majority of the four lists also related to actuarial science. The four lists of resource types informed the authors of the wide variety of types and ultimately motivated the authors to gear their list of types to actuarial science. The result was a list of 28 resource types (including the category "Unknown") that would be used to test the methodology. Figure 2 includes the list of resource types the authors created for the experiment.

The project's success was dependent upon how well the AI model was selected and instructed.

Classifying Resource Types

The process of classifying references by resource type was the part of the project that most relied on AI. The project's success was dependent upon how well the AI model

```
prompt = f""" Analyze the following text and identify citation information.

    For each citation/reference, provide:

    1. title

    2. author(s)

    3. source

    4. full-text

    5. category (must be one or two of these exact values: Journal article, Book, Book Chapter,
Magazine article, Newspaper article, Conference Paper, Dissertation and Thesis, Government
document from federal government agencies, Government document from state government
agencies, Government document from local government agencies, Government document
from foreign governments, Document from intergovernmental organizations (IGOs),
Document from international nongovernmental organizations (INGOs), Report from Non-
profit organizations, Court Case, Industry report, Business report, Working paper, Technical
report, Preprint, Standards document, Patents, Software, Database, Dataset, Unpublished
work, Website, Unknown)

"""
```

Figure 2. AI prompt that includes the full list of resource types used in the experiment.

was selected and instructed. In order to classify references into resource type categories, the authors utilized AI foundation models, pre-trained models for general purposes, instead of creating an AI model from scratch.

Selecting the AI Model

Numerous foundation models are readily available, such as OpenAI's ChatGPT and Google's Gemini. OpenAI's ChatGPT was evaluated first. However, the authors' institution had been in ongoing negotiations with OpenAI for a service contract and, as such, the tool was not available for research purposes. Google, however, provided educational credits to encourage researchers to experiment with its foundation models. In addition, Google's model offered Retrieval-Augmented Generation (RAG), marketed as, "Grounding with Google Search," enabling LLM applications to search for and obtain additional information from the internet and external sources. This capability can help explain the AI's reasoning and improve the classification process. Furthermore, this study chose Gemini 2.0 Flash from among available foundation models, as it was the most advanced version available at the time the experiments began.

In this study, RAG provided the research team with an added advantage in understanding and verifying the AI-generated classifications. For example, the LLM occasionally misidentified the correct resource types for government documents and reports, as these classifications often depend on the nature of the publishing body, for example nonprofit organization, business, or federal, state, and foreign governments. Several reports from nonprofit organizations were incorrectly classified as government documents by the AI model. One such case involved the National Association of Insurance Commissioners (NAIC), a nonprofit organization that standardizes insurance regulations, which the AI misclassified as a government agency. To gauge the model's performance, the researchers manually reviewed the classification results. Initially, the authors manually searched for publisher information to determine classification accuracy. When applying RAG, the prompt included instructions to search Google to verify the publisher and determine whether the resource was issued by a nonprofit organization, a business, or a government agency. The added line to the prompt read: "If the publica-

tion source and type are uncertain, search Google to verify the publisher and the document. Determine whether the source is from a non-profit organization, a business, or a government agency." This line asked the AI model to explicitly provide reasoning for its classifications and assisted human labelers in evaluating the model's performance.

AI Application

Based on the selected foundation model, an AI application was developed to communicate with Gemini 2.0 Flash through an API. A virtual machine (VM) was also set up on Google Cloud to allow the application to run continuously. For the AI program, the authors utilized prompt engineering in combination with Python code to automatically send a list of references to the AI model via an API. Prompt engineering involved crafting precise and plain-language instructions to elicit the desired outcomes, and the API enabled direct programmatic interaction with the AI model, eliminating the need for manual input by humans.[46]

A Python program was written to read each line of references harvested from Scopus via the Excel file. A prompt was constructed to send reference information, such as title, author(s), source title (journal title, book title, or website), and reference full text (as cited in the article) to the AI model, along with instructions for how to classify the references according to the resource types provided. Figure 2 shows the full prompt.

Review and Analysis

The goals of the study were to determine if and how well AI could be used in citation analysis to examine references and predict or label resource types in a subject for the purpose of informing collection development. In addition, the authors assessed the accuracy of the results for the purpose of encouraging other librarians to replicate the process for their collection evaluation needs. Therefore, the authors evaluated the stage within the process in which the AI model labeled references. The procedure involved an iterative process consisting of three major steps: (1) applying the methodology, (2) analyzing the results and making necessary adjustments, and (3) re-testing the revised methodology.

Comparing AI and Human Labeling

The first step in the process was applying the methodology to the dataset to classify references by type. After harvesting references from the *North American Actuarial Journal* (*NAAJ*), the authors had the AI model classify the references by resource type. To reduce the number of references requiring comparison between AI-generated labels and human-assigned labels, a 15 percent sample of all references was selected. Because the distribution of resource types was skewed toward traditional resources, stratified sampling was performed by AI-assigned resource type, with samples taken proportionally from each type to ensure representation across all categories.[47] For *NAAJ*, 2,251 references were manually labeled using a predefined set of resource type categories.

Next, the researchers analyzed the results for accuracy and made necessary adjustments. To do so, one of the authors, with the assistance of a graduate student, manually

labeled the sample references according to resource type. The results of the manual labeling were then compared against those generated by the AI model. In the test subset, 89 references (3.95 percent) exhibited discrepancies between the human-labeled results and the AI-generated labels. Reviewing several cases, the authors found that many of the discrepancies were due to the model's limited domain knowledge of publishing sources and organizational types. For example, publications from the Organization for Economic Cooperation and Development (OECD), an intergovernmental organization, were expected to be classified as "Documents from intergovernmental organizations (IGOs)." However, the AI model incorrectly labeled these cases as a "Report from Nonprofit Organizations."

In some cases, assigning a document to a single category was still not straightforward and, as a result, could not be considered a clear misclassification or error. For example, another document from the OECD was labeled as a "Technical Report" by the AI model because its title included the term "Technical Report," whereas human annotators classified it as a "Document from IGOs." Another case involved a report published on a government agency website, where it could reasonably be classified either as a "Website" or as a "Government Document." In such instances, a document could appropriately belong to either category. This suggested that the prompt for the AI model should be further refined to provide clearer instructions for determining the most suitable classification.

A small subset of references, particularly those citing websites or web pages, were no longer accessible, making it impossible to verify the original content. Additionally, some references remained difficult to assign to any category. For example, brief announcements from businesses did not clearly fit into either the "Business Report" or "Technical Report" categories, supporting the need to reconsider or refine the resource type categories. Figure 3 shows the full AI model workflow.

Performance Evaluation

When comparing human labeling to AI labeling, initial assessment revealed 89 discrepancies out of a sample of 2,251 references, resulting in an accuracy of 96.05 percent. While this level of accuracy falls within an acceptable margin, the authors noticed that discrepancies were less frequent among certain traditional resource types, like books and journals, and appeared to be more concentrated in other types, such as reports and government documents. Therefore, to gain a more comprehensive understanding of the model's accuracy with individual resource types and identify categories in need of modification, a performance evaluation was conducted for each resource type using machine learning evaluation metrics such as accuracy, precision, recall, and F_1-score.[48] Table 3 provides a list of the 26 resource type categories (including "Unknown) and includes the calculations for the values. The list excludes the categories "Government documents from local government agencies" and "Patents," along with their calculated values, because there were no results for these two resource types. The F_1 score is a single performance metric for classification that balances precision and recall, ranging from 0 to 1, where 0 represents the lowest and 1 indicates perfect performance. The interpretation of the F_1 score often varies based on its applications, and there is no universal standard for

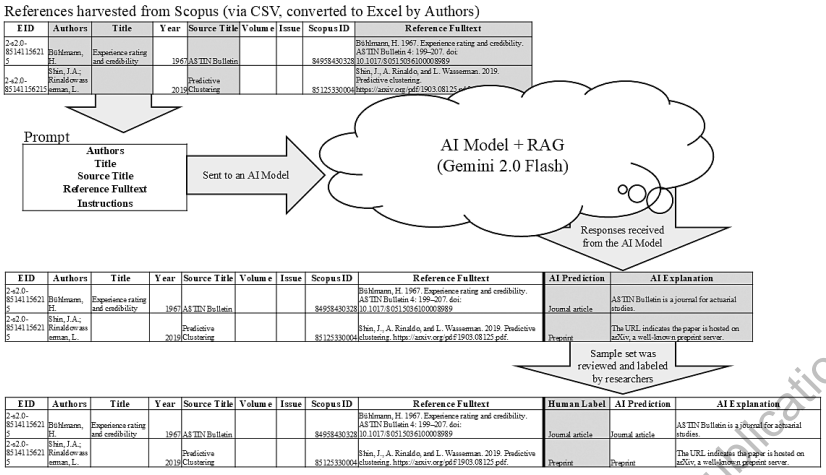

Figure 3. AI model workflow.

assessing what defines a good or poor score. Generally, values above 0.9 are considered excellent, those between 0.8 and 0.9 are deemed good, and values below 0.5 indicate poor performance.[49] For critical applications, such as medical diagnostics, even an F_1 of 0.9 may be insufficient.[50] For the purposes of this study, the authors defined ≥ 0.8 as excellent performance and < 0.5 as requiring improvement. Appendix A provides the formulas used in these calculations.

Categories in Need of Modification

As shown in Table 3, there was an overall accuracy score of 96.05 percent, with the first 17 categories demonstrating excellent classification performance from the AI model. In contrast, nine categories relating to three types of reports, three types of government documents, datasets, websites, and unknown exhibited comparatively lower F_1 scores (ranging from 0.4211 to 0.6019, including two N/A), highlighting several areas that required refinement of resource type and category definitions. In addition, the authors identified categories that could be consolidated to facilitate identification of resource type.

After examining the evaluation metrics, the authors noticed that three categories relating to government documents, including "Documents from international non-governmental organizations (INGOs)" and three categories of reports had lower F_1 scores. The authors examined references within these categories and found that AI incorrectly categorized the references because of the model's limited domain knowledge of publishing sources and organizational types, as previously discussed in "Comparing AI and Human Labeling." It is the opinion of the authors that with respect to these categories, there was no advantage to differentiating them. The four report categories (technical, business, industry, and non-profit reports) and all five government document categories (federal, foreign, intergovernmental, state, and INGOs) were redefined to two, more general categories: "Report" and "Government Document."

Table 3.

Calculations of accuracy, precision, recall, and F_1 score for resource types from *NAAJ* sample, ranked by F_1 Score

	True Positive	*False Positive*	*False Negative*	*True Negative*	*Accuracy*	*Precision*	*Recall*	*F1 Score*
Conference Paper	51	0	0	2200	1.0000	1.0000	1.0000	1.0000
Database	7	0	0	2244	1.0000	1.0000	1.0000	1.0000
Dissertation and Thesis	7	0	0	2244	1.0000	1.0000	1.0000	1.0000
Technical report	17	0	0	2234	1.0000	1.0000	1.0000	1.0000
Unpublished work	1	0	0	2250	1.0000	1.0000	1.0000	1.0000
Journal article	1646	9	1	595	0.9956	0.9946	0.9994	0.9970
Book	175	2	0	2074	0.9991	0.9887	1.0000	0.9943
Government document from federal government agencies	30	1	1	2219	0.9991	0.9677	0.9677	0.9677

Table 3, cont.

	True Positive	*False Positive*	*False Negative*	*True Negative*	*Accuracy*	*Precision*	*Recall*	*F1 Score*
Book Chapter	44	2	3	2202	0.9978	0.9565	0.9362	0.9462
Software	13	0	2	2236	0.9991	1.0000	0.8667	0.9286
Working paper	65	8	3	2175	0.9951	0.8904	0.9559	0.9220
Preprint	8	2	0	2241	0.9991	0.8000	1.0000	0.8889
Government document from foreign governments	17	0	5	2229	0.9978	1.0000	0.7727	0.8718
Newspaper article	3	0	1	2247	0.9996	1.0000	0.7500	0.8571
Magazine article	9	1	3	2238	0.9982	0.9000	0.7500	0.8182
Court case	2	0	1	2248	0.9996	1.0000	0.6667	0.8000
Standards document	2	1	0	2248	0.9996	0.6667	1.0000	0.8000
Report from Non-profit organizations	31	19	22	2179	0.9818	0.6200	0.5849	0.6019
Dataset	3	0	4	2244	0.9982	1.0000	0.4286	0.6000
Industry report	5	4	5	2237	0.9960	0.5556	0.5000	0.5263
Government document from state government agencies	2	5	0	2244	0.9978	0.2857	1.0000	0.4444

Business report	9	0	24	2218	0.9893	1.0000	0.2727	0.4286
Website	11	30	0	2210	0.9867	0.2683	1.0000	0.4231
Document from intergovernmental organizations (IGOs)	4	1	10	2236	0.9951	0.8000	0.2857	0.4211
Document from international nongovernmental organizations (INGOs)	0	1	0	2250	0.9996	0.0000	N/A	N/A
Unknown	0	3	4	2244	0.9969	0.0000	0.0000	N/A

Accuracy: 96.05 percent

It was also discovered that the category "Conference Paper" included conference papers, conference slides, and conference keynote speeches. This category was relabeled as "Conference Material" to encompass the varying types of conference materials. In addition, after re-examining the three references in the category, "Standards Document," the authors determined that the items could be classified as government documents or reports. As a result, the category "Standards Document" was omitted and the items were reassigned. One category "Newsletter and Press Release," was added to the list because the authors discovered announcements and newsletters within numerous other categories. Two categories, "Patents" and "Government Documents, Local," appeared in the original set of the resource type categories created by the authors (see Figure 2). However, because these categories were not found in the references, they were removed from the list. Based on the observations and findings, the authors revised the category list (Table 4), which now consists of 19 categories (including Unknown).

Enhancing the AI Model

RAG is a technique that enhances AI by searching external sources like Google or Google Scholar. The authors incorporated RAG to expand the AI model's ability to search and retrieve information online, with the goal of informing the AI model to more accurately categorize references. The revised prompt, that incorporated the RAG feature, not only labeled the resource types but also searched the sources of the resources, such as titles or publishers, and incorporated the information into the classification and evaluation processes.

Step three focused on re-testing the original dataset with the revised categories and revised prompt incorporating RAG. Table 4 provides the evaluation metrics, ranked in order of F_1 score, for the 19 revised resource type categories after refining categories and implementing RAG. It is worth noting that F_1 scores for Government Document and Report improved as a result of refining the list of resource types. The overall accuracy score was 97.16 percent, which was an improvement from 96.05 percent.

When applying citation analysis in collection evaluation, librarians will often apply the method to more than one journal. In addition, to validate an AI model, it is often recommended to test it with a new test set. To further determine the effectiveness of the refined categories and RAG, the authors applied the AI model to another dataset. For this case, they used the second actuarial science journal, *Journal of Risk and Insurance (JRI),* with a sample size of 2,536 references. Table 5 provides the results for the 19 categories, ranked according to F_1 score and showing overall accuracy of 98.62 percent. Six categories achieved a perfect score of 1.00, while nine categories scored from 0.8000 to 0.9995, and three scored from 0.6667 to 0.7200. Notably, none scored below 0.5.

The evaluation process provided additional insight regarding the methodology. First, it demonstrated the importance of having a list of resource types appropriate for the subject area under examination. Second, based on the evaluation metrics, incorporating AI tools in citation analysis enhances the process because AI can label references according to resource type.

Table 4.

Calculations of accuracy, precision, recall, and F1 score for resource types from *NAAJ* sample after refining the categories and implementing RAG, ranked by F_1 score

	True Positive	False Positive	False Negative	True Negative	Accuracy	Precision	Recall	F1 Score
Court case	3	0	0	2248	1.0000	1.0000	1.0000	1.0000
Database	7	0	0	2244	1.0000	1.0000	1.0000	1.0000
Journal article	1641	1	6	603	0.9969	0.9994	0.9964	0.9979
Book	169	0	6	2076	0.9973	1.0000	0.9657	0.9826
Book Chapter	43	1	4	2203	0.9978	0.9773	0.9149	0.9451
Software	13	0	2	2236	0.9991	1.0000	0.8667	0.9286
Government Document	62	4	8	2177	0.9947	0.9394	0.8857	0.9118
Report	102	17	8	2124	0.9889	0.8571	0.9273	0.8908

Table 4, cont.

	True Positive	False Positive	False Negative	True Negativ	Accuracy	Precision	Recall	F1 Score
Newsletter and press release	4	0	1	2246	0.9996	1.0000	0.8000	0.8889
Conference material	50	12	1	2188	0.9942	0.8065	0.9804	0.8850
Dissertation and Thesis	7	2	0	2242	0.9991	0.7778	1.0000	0.8750
Newspaper article	3	0	1	2247	0.9996	1.0000	0.7500	0.8571
Working paper	53	3	15	2180	0.9920	0.9464	0.7794	0.8548
Magazine article	9	2	4	2236	0.9973	0.8182	0.6923	0.7500
Preprint	8	6	0	2237	0.9973	0.5714	1.0000	0.7273
Unpublished work	1	1	0	2249	0.9996	0.5000	1.0000	0.6667
Dataset	3	0	4	2244	0.9982	1.0000	0.4286	0.6000
Website	9	13	2	2227	0.9933	0.4091	0.8182	0.5455
Unknown	0	2	2	2247	0.9982	0.0000	0.0000	N/A

Accuracy: 97.16 percent

Table 5.

Calculations of accuracy, precision, recall, and F_1 score for resource types from *JRI* sample, ranked by F_1 score

	True Positive	False Positive	False Negative	True Negativ	Accuracy	Precision	Recall	F1 Score
Book	94	0	0	2442	1.0000	1.0000	1.0000	1.0000
Court case	1	0	0	2535	1.0000	1.0000	1.0000	1.0000
Dissertation and Thesis	3	0	0	2533	1.0000	1.0000	1.0000	1.0000
Magazine article	5	0	0	2531	1.0000	1.0000	1.0000	1.0000
Software	2	0	0	2534	1.0000	1.0000	1.0000	1.0000
Unpublished work	3	0	0	2533	1.0000	1.0000	1.0000	1.0000
Journal article	1943	1	1	591	0.9992	0.9995	0.9995	0.9995
Working paper	118	0	7	2411	0.9972	1.0000	0.9440	0.9712
Report	129	7	2	2398	0.9965	0.9485	0.9847	0.9663

Table 5, cont.

	True Positive	False Positive	False Negative	True Negativ	Accuracy	Precision	Recall	F1 Score
Book Chapter	63	2	3	2468	0.9980	0.9692	0.9545	0.9618
Conference material	27	4	0	2505	0.9984	0.8710	1.0000	0.9310
Newsletter and press release	6	0	1	2529	0.9996	1.0000	0.8571	0.9231
Government Document	71	2	11	2452	0.9949	0.9726	0.8659	0.9161
Dataset	3	0	1	2532	0.9996	1.0000	0.7500	0.8571
Newspaper article	8	0	4	2524	0.9984	1.0000	0.6667	0.8000
Website	18	13	1	2504	0.9945	0.5806	0.9474	0.7200
Database	2	0	2	2532	0.9992	1.0000	0.5000	0.6667
Preprint	5	5	0	2526	0.9980	0.5000	1.0000	0.6667
Unknown	0	1	1	2534	0.9992	0.0000	0.0000	N/A

Accuracy: 98.62 percent

Findings

One goal of the current study is to determine whether, and how well, AI can be used in citation analysis to examine references and predict resource types in a subject. Therefore, the findings will discuss how well the AI model performed and contributed to citation analysis and will share insights the authors gained with the goal of encouraging others to integrate AI into their citation analyses. It is important to note that while the authors used the subject area of actuarial science as a prototype, the study does not discuss findings specific to the field itself.

AI Enhances the Citation Analysis Methodology

In the initial experiment using 28 resource type categories, the AI model achieved an accuracy of 96.05 percent, demonstrated when its predictions were manually evaluated against a 15 percent stratified sample drawn from a total of 14,997 references from *NAAJ*. After refining the resource type categories and improving the instructional AI prompts, accuracy increased to 97.16 percent using the same sample dataset. To test the method on a new dataset, references from *JRI* were analyzed, and the AI model achieved an even higher accuracy rate of 98.62 percent when compared with human-labeled data.

In addition to high accuracy, use of the AI model saved time. In the most recent experiment on the *JRI* dataset, the model completed predictions for 16,913 references in two hours and 54 minutes, an average speed of 5,872 per hour. In contrast, a human labeler with prior experience spent approximately seven hours manually reviewing a 15 percent sample (2,536 references). Extrapolating from this, labeling the full dataset would have required an estimated 47 human hours. Moreover, the cost of experimenting with both the *NAAJ* and *JRI* datasets using Google Cloud was $5.82. Based on these results, the authors conclude that the approach is effective and efficient for analyzing references.

Resource Types

The authors found that preparing a predefined list of resource types is essential to the process of using the AI model to identify resource types for three key reasons. First, the list provides clear guidance for the AI model on how to classify each reference without ambiguity. Second, as mentioned in the literature review, resource usage patterns vary across disciplines, making it necessary to understand the disciplinary context and tailor the list of resource types accordingly. Third, the granularity level in classification should reflect the information needs of researchers, enabling a more precise understanding of the types of resources being used.

. . . that preparing a predefined list of resource types is essential to the process of using the AI model to identify resource types . . .

Table 6 presents the results of AI predictions for the *JRI* dataset. While the most frequently cited resource type was journal articles, it is notable that grey literature, such as reports, working papers, government documents, and conference materials, also represent a large portion of the *JRI* dataset at 16.51 percent. Although some researchers

categorize many of these under a general "grey literature" or "others" category in their studies, this study originally aimed to examine the detailed types within grey literature and thus treated them as separate categories and further subdivided them within government documents and within reports. However, while reassessing the strategy, the authors realized that distinguishing the various government document types and report types might not be as necessary as previously assumed. Therefore, broader categories (government documents and reports) encompassing these subtypes were used. The results using this broader classification were found to be satisfactory for the purposes of this analysis.

Prompting for Better Results

While having a clear list of resource types is essential for classification, it is equally important to recognize the multi-faceted nature of resources. In other words, a single resource can be described in multiple ways, by format, source, or content. For example, an online report published by a government agency could be categorized as a webpage by format, a report by content type, and a government document by source type. In such cases, it is crucial to instruct the AI model on how to prioritize among these facets to ensure consistent and meaningful classification.

One of the consistent findings from the early experiments was that the AI frequently classified government documents published by intergovernmental organizations, such as the OECD or IMF, as reports. While it is understandable that the AI model might classify documents from organizations like the OECD as reports, this was not the intended classification in the context of this study, which aimed to identify such materials as government documents. This misclassification can be addressed by reinforcing the intended logic through more explicit instructions. An example would be, "If a report is published by an intergovernmental organization such as the UN or OECD, it should be classified as a government document."

A similar issue was observed with preprints. The AI model often predicted manuscripts from preprint servers as working papers, likely because many of these papers are later published in journals. However, when the research goal is to determine how many publications were cited while still in preprint status, it becomes critical to classify them accordingly. In such cases, the study team found it helpful to provide specific instructions to the AI model, for example: "If a paper is cited from a known preprint server such as arXiv, classify it as a preprint, not a working paper."

It is OK to Ask AI for an Explanation

One of the well-known challenges of AI models is their black-box nature, meaning they do not explain the reasoning behind their predictions. For example, an AI model may classify a resource as a report when it could be considered a government document without explaining how it reached that decision. Two approaches are being developed to address this issue: Explainable AI (XAI) and Interpretable AI (IAI). XAI focuses on explaining how models reach their decisions, often using a second model to explain the first model. IAI, on the other hand, aims to design models that are inherently interpretable, providing their own explanations without the need for post-hoc interpretation.[51]

The foundation model used in this study, Google's Gemini 2.0 Flash, does not provide native XAI or IAI capabilities but can simulate rationales when prompted. Therefore, in this study, the authors added an instruction asking the AI model to explain the rationale behind each prediction, combined with RAG. While this method is not yet perfect, the explanations can be compared with the prediction itself and help human reviewers determine the accuracy of the classification or identify cases that warrant further investigation.

The following is an example of an item that was misclassified as a report by the AI model, while human annotators correctly labeled it as a government document, and the AI model explained that it is a government document. Upon reviewing the AI model's explanation, it became clear that the model's reasoning aligned more closely with a government document type, even though it ultimately predicted it as a report. This further confirmed the accuracy of the human label. Such explanations, even when imperfect, can assist human reviewers in making more accurate and informed decisions.

- Reference Full text: Committee of Insurance and Occupational Pension Supervisors, (2008), Quantitative Impact Study 4. Available at www.ceiops.org
- Title: Quantitative Impact Study 4
- Source: Committee of Insurance and Occupational Pension Supervisors
- AI Prediction Type: Report
- AI Explanation: Committee of Insurance and Occupational Pension Supervisors (CEIOPS) was a predecessor to the European Insurance and Occupational Pensions Authority (EIOPA). Therefore, it can be considered a government agency of the EU. Although it is a report, because it is published by an intergovernmental organization, it is classified as a government document.

After noticing similar misclassifications within the government document and report categories, an additional instruction was added to ask the AI model to find the source of the resource, including the publishing body. While many journal articles and books do not specify such information in their references, reports and government documents often do. The inclusion of the identified source helped human reviewers scan and evaluate the list more efficiently.

References' Impact on Citation Analysis

References are the key to analyzing citations and bibliometrics, and studies have discussed issues related to inaccurate references, including spelling errors and variations in reference elements for decades.[52] Nevertheless, this study revealed the persistence of these issues, highlighting the importance of verifying references before conducting citation analysis. *The Geneva Papers on Risk and Insurance- Issues and Practice* is one of the major journals in actuarial science. This study identified 26 variations in spellings of the title. Some variations in the Scopus data included hyphens, long hyphens, the presence or absence of a colon before the subtitle, or a missing subtitle. Table 7 provides a list of variations of the journal title and the frequency with which they occurred in Scopus. While these variations do not appear to harm the references, a typical programming code, without cleaning the data, may interpret them as different titles, impeding accurate analysis. In addition to issues with title variations, many references were found to be incomplete or incorrect, often missing essential information or displaying inconsistencies in citation style. These issues reinforce the need for thorough data verification prior to

Table 7.

Example of title variations found in *JRI* dataset

Title Variations	Count
Geneva Papers	3
Geneva Papers for Risk and Insurance	1
Geneva Papers on Risk and Insurance	29
Geneva Papers on Risk and Insurance – Issues and Practice	2
Geneva Papers on Risk and Insurance Issues and Practice	3
Geneva Papers on Risk and Insurance: Issues and Practice	4
Geneva Papers on Risk and Insurance-Issues and Practice	11
Geneva Papers on Risk and Insurance–Issues and Practice	7
Geneva Papers on Risk and Insurance—Issues and Practice	15
Geneva Papers on Risk andInsurance-Issues and Practice	1
Geneva Papers on Risk andInsurance–Issues and Practice	1
Geneva Papers on Riskand Insurance-Issues and Practice	1
Geneva Papers onRisk and Insurance–Issues and Practice	1
Geneva Paperson Risk and Insurance	1
Geneva Paperson Risk and Insurance–Issues and Practice	1
GenevaPapers on Risk and Insurance-Issues and Practice	1
GenevaPapers on Risk and Insurance–Issues and Practice	1
The Geneva Papers	4
The Geneva Papers on Risk and Insurance	2
The Geneva Papers on Risk and Insurance - Issues and Practice	7
The Geneva Papers on Risk and Insurance – Issues and Practice	1
The Geneva Papers on Risk and Insurance Issues and Practice	1
The Geneva Papers on Risk and Insurance. Issues and Practice	2
The Geneva Papers on Risk and Insurance-Issues and Practice	8
The Geneva Papers on Risk and Insurance–Issues and Practice	3
The Geneva Papers on Risk and Insurance—Issues and Practice	32

a citation analysis. Any future studies conducting citation analysis should be mindful of such potential errors, as they may significantly affect the accuracy and reliability of the results.

Study Limitations

As mentioned, while integrating AI in library processes is not new, it has not been explored extensively within collection development. Collection evaluation is an essential function of libraries, and, to ensure collection relevancy, should be conducted on a regular basis. While citation analysis is a use-based methodology often employed by librarians, it can sometimes be time-consuming to conduct and analyze the results. The current study advances citation analysis methodology in numerous ways by integrating AI into the process. However, a requirement and limitation of the methodology is the need to create a list of resource types relevant to the discipline. This requires a deep human-conducted analysis before applying the AI model. This process could prove time-consuming if the librarian is not familiar with the discipline's research. An element that could facilitate this process would be a master list of resource types.

The first—and in many cases essential—step in the proposed methodology requires the researcher to select a source from which to harvest references. Using actuarial science, the authors identified two leading journals within the subject area. The goal was to identify journals that represented a variety of leading research within the discipline. Depending on the researcher's familiarity with the subject area, this process could be challenging. In addition, as is the case with WoS, the database often does not provide a specified list of journals for every subject category. To provide insight into the process, authors might consider harvesting references from journal(s) with high impact factors, journals published or sponsored by leading associations within the field, or journals whose scope is broad and includes research reflecting the majority of issues within the field, to name just a few. Another limitation of using WoS or Scopus for harvesting references is the limit to what is available from sciences, social sciences, and humanities resources. While researchers in the sciences primarily publish their findings in journals that are indexed in WoS or Scopus, those in the social sciences and the humanities also use books, journals, and non-traditional resources for research output. Therefore, WoS and Scopus may not be the most comprehensive source for harvesting references.

While the current study proposes a methodology that facilitates the citation analysis process, it did not attempt to evaluate the various AI models available. The authors selected the model that they were most familiar with and that was most available. The possibility exists that other AI models might perform better, offering an opportunity for future research.

Implications for Future Research

The proposed methodology presents new opportunities for collection managers. Incorporating AI provides librarians with an efficient way to analyze a greater volume of references. However, because this approach is new, it is essential that it is thoroughly examined or applied to a variety of subjects. The following provides numerous examples

of potential future research and practice that will inform and hopefully enhance the methodology.

There are elements of the authors' methodology that can be applied by other researchers to gain deeper insight into library resource usage. The current study provides a list of resource types that could be used by researchers or, alternatively, researchers could create their own lists adapted to other subjects or their own institution. The study's application of RAG demonstrates to future researchers the value of incorporating additional facets of information. The authors also applied an AI model, Gemini 2.0 Flash, that is widely available and that performed efficiently and effectively. The methodology provided in the current study opens the door for other researchers to explore using different AI models, as they become more available and easier to use without a computer science background. While this project evaluates the possibility of AI's integration into citation analysis, it more importantly invites librarians to explore AI innovations.

Incorporating AI provides librarians with an efficient way to analyze a greater volume of references.

The integration of AI into citation analysis facilitates the collection evaluation process and provides additional research opportunities across all subject areas. The more comprehensive and reflective of the subject area being examined the resource list is, the better the chance the AI model can thoroughly and accurately sort the references. Future research could strive to provide a comprehensive list of resource types or provide list(s) related to specific disciplines or subject areas.

The integration of AI into citation analysis facilitates the collection evaluation process and provides additional research opportunities across all subject areas.

The authors believe that the methodology's strength and contribution to the field lies heavily in how well it can be applied to a variety of subjects. While the authors of the current study used Scopus to harvest references, other databases like WoS could be used. As mentioned previously, Scopus and WoS focus on journals. Because researchers in the sciences publish their findings in journals, the current methodology could be easily applied to the STEM disciplines and other scientific subject areas like medicine, engineering, and public health. However, when applying the methodology to areas within the social sciences and humanities, findings obtained when harvesting references from WoS and Scopus should be considered only one part of the picture and, more importantly, researchers might consider identifying additional sources for harvesting references so that resources such as books and book chapters are included.

New AI models will continue to emerge as will improvements of current AI technology. These changes could impact how the models contribute to the current methodology. Another research possibility is regular evaluation of AI models applicable to the process of sorting and identifying resource types. Regularly evaluating these tools would help determine their strengths, weaknesses, and suitability for the citation analysis process. In addition, research assessing the various AI tools' suitability to a wide variety of subjects

would be helpful since new resource types will most likely continue to be established and integrated into research. This fine-tuning of the methodology will enable the citation analysis process to keep pace with the ongoing development of resource types and ultimately contribute to collection evaluation.

Through designing the study, the authors discovered that a key element of the methodology is the accuracy of references. This is especially relevant if a reference does not include correct information for the source title because the AI model uses this information to sort by resource type. Future research examining trends and errors of references could contribute to the current methodology.

Conclusion

As Late et al. pointed out in their analysis of reading practices in scholarly work, "scholars read a lot and they read a variety of publications" and as a result, librarians who support scholars' information needs must stay aware of what resources scholars are using.[53] This view is supported by Huang and Chang, who found diverse research output in the social sciences and humanities as compared to research in the natural sciences, and they concluded that output-based evaluations for social science and humanities research require different methodologies from those applied to the natural science disciplines and should specifically address diversity of resource types.[54] While tools exist to help librarians determine which resources are used more, many of them are biased toward journals. Unfortunately, for librarians building collections in the humanities and social sciences, these tools only provide part of the picture. With ARL data indicating an increase in library spending on resources, and collections costs outpacing inflation, it is essential that librarians have tools that provide reliable data to inform collection decisions and that reflect the wide variety of resources available.[55]

The methodology presented in the current study demonstrates that AI tools can facilitate and advance citation analysis methodology. AI can provide librarians relying on citation data with a more complete picture of the titles and types of resources used within the subject areas they manage, especially for research in the humanities and social sciences. The study outlines the important elements of the proposed process which includes a list of resource types geared to the subject area and relying on the most appropriate AI tool, provides details on the required steps, and describes ways to guarantee an efficient and informative analysis. More specifically, it provides guidance to librarians and enables them to understand their respective subjects more thoroughly through analysis of the resources and types of resources used in the discipline. By applying the current AI methodology, librarians will be able to adopt a more efficient and tailored approach to keeping up with what resources researchers are relying on and, more importantly, will be able to identify trends to better anticipate researchers' needs.

Acknowledgements

This research project used Google Gemini 2.0 Flash to classify resource types. The research project was supported by the Google Cloud for Researchers Program, which provided the necessary computational resources.

The authors are grateful to Dr. Volodymyr Kindratenko for their expert guidance on methodologies for evaluating AI performance.

Sarah G. Park is mathematics and computational science librarian at the University of Illinois Urbana-Champaign, email: gpark1@illinois.edu, ORCID: 0000-0002-5052-7252.

Lisa Romero is communications librarian and Latina/o Studies librarian at the University of Illinois Urbana-Champaign, email: L-Romero@illinois.edu, ORCID: 0009-0005-1102-4651.

Appendix A

Definitions and Formulas Relating to Performance Evaluation

- **True** Positive (TP): The AI correctly classifies the reference as belonging to a specific resource type (e.g., "journal article"), and this classification matches the human label.
- **False** Positive (FP): The AI incorrectly classifies the reference as belonging to the specific resource type, when the human label indicates it belongs to a different type.
- **False** Negative (FN): The AI incorrectly classifies the reference as not belonging to the specific resource type, when the human label indicates that it does.
- **True** Negative (TN): The AI correctly classifies the reference as not belonging to the specific resource type being evaluated, and this also matches the human label.

    - **Accuracy:** Proportion of correct classifications out of all references:
      $$\frac{(TP+TN)}{(TP+TN+FP+FN)}$$
    - **Precision:** Proportion of correctly identified positive cases out of all references predicted as positive:
      $$\frac{TP}{(TP+FP)}$$
    - **Recall:** Proportion of correctly identified positive cases out of all actual positive references:
      $$\frac{TP}{(TP+FN)}$$
    - **F1 Score:** Harmonic mean of precision and recall, balancing both metrics:
      $$2 \cdot \frac{(\text{precision} \cdot \text{recall})}{\text{precision}+\text{recall}}$$

Notes

1. Howard D. White, "Citation Analysis," in *Encyclopedia of Library and Information Sciences*, ed. John D. McDonald and Michael Levine-Clark, 4th ed. (CRC Press, 2017), 923.
2. *APA Style* (blog); "References Versus Citations," by Timothy McAdoo, posted September 20, 2017, accessed August 5, 2025.
3. Janet Dagenais Brown, "Citation Searching for Tenure and Promotion: An Overview of Issues and Tools," *Reference Service Review* 42, no. 1 (2014): 70, https://doi.org/10.1108/RSR-05-2013-0023; White, "Citation Analysis," 925.
4. White, "Citation Analysis," 923-939.
5. Michael Levine-Clark and Esther L. Gil, "A New Comparative Citation Analysis: Google Scholar, Microsoft Academic, Scopus, and Web of Science," *Journal of Business & Finance Librarianship* 26, no. 1-2 (2021): 145-146, https://doi.org/10.1080/08963568.2021.1916724.
6. Raminta Pranckutė, "Web of Science (WoS) and Scopus: The Titans of Bibliographic Information in Today's Academic World," *Publications* 9, no. 1 (March 2021), https://doi.org/10.3390/publications9010012.
7. Philippe Mongeon and Adèle Paul-Hus, "The Journal Coverage of Web of Science and Scopus: A Comparative Analysis," *Scientometrics* 106, no. 1 (January 2016), 213., https://doi.org/10.1007/s11192-015-1765-5.
8. "Reclassification of Papers in Multidisciplinary Journals for Creating Field Baselines," InCites Help Desk, accessed May 11, 2025, https://incites.zendesk.com/hc/en-gb/articles/22586272202513-Web-of-Science-Research-Areas#h_01HPQBWJQHP2CBCFSXFR914YQG.
9. Selmer Bringsjord and Naveen Sundar Govindarajulu, "Artificial Intelligence," in *The Stanford Encyclopedia of Philosophy*, ed. Edward N. Zalta and Uri Nodelman, Fall 2024 ed., https://plato.stanford.edu/archives/fall2024/entries/artificial-intelligence/.
10. Thimira Amaratunga, "Introduction," in *Understanding Large Language Models: Learning Their Underlying Concepts and Technologies*, ed. Thimira Amaratunga (Apress, 2023), 1-7, https://doi.org/10.1007/979-8-8688-0017-7_1.
11. Rishi Bommasani, et al, "On the opportunities and risks of foundation models," Preprint, 2021, https://crfm.stanford.edu/report.html.
12. Chip Huyen, AI Engineering: Building Applications with Foundation Models (O'Reilly, 2024).
13. Tenopir, Volentine, and King, "Scholarly Reading," 135-136.
14. Peggy Johnson, *Fundamentals of Collection Development and Management*, 4th ed. (ALA Editions, 2018), 300-302.
15. Erika Alves dos Santos, Silvio Peroni, and Marcos Lui Mucheroni, "Referencing Behaviours Across Disciplines: Publication Types and Common Metadata for Defining Bibliographic References," *International Journal on Digital Libraries* 25, no. 3 (September 2024), 443-445, https://doi.org/10.1007/s00799-023-00351-8; Svein Kyvik, "Changing Trends in Publishing Behaviour Among University Faculty," *Scientometrics* 58, no. 1 (September 2003): 35-48, https://doi.org/10.1023/A:1025475423482; Elina Late, Carol Tenopir, Sanna Talja, and Lisa Christian, "Reading Practices in Scholarly Work: From Articles and Books to Blogs," *Journal of Documentation* 75, no. 3 (2019): 478-499, DOI:10.1108/JD-11-2018-0178.
16. dos Santos, Peroni, and Mucheroni, "Referencing Behaviours," 443-455.
17. Kyvik, "Changing Trends in Publishing Behaviour," 35-48.
18. Mary C. Schlembach, "Doctoral Dissertations in Chemistry and Physics: A Longitudinal Study," *Science & Technology Libraries* 42, no. 4 (2023): 441-455, https://doi.org/10.1080/0194262X.2023.2208627.
19. Late, et al, "Reading Practices," 478-499.
20. Pei-Shan Chi, "Which Role do Non-Source Items Play in the Social Sciences? A Case Study in Political Science in Germany," *Scientometrics* 101, no. 2 (November 2014): 1195-1213, https://doi.org/10.1007/s11192-014-1433-1; dos Santos, Peroni, and Mucheroni,

"Referencing Behaviours," 448; Huang and Chang, "Characteristics of Research Output, 1819-1828; Kyvik, "Changing Trends," 35-48.

21. Late, et al, "Reading Practices," 481.
22. John Budd, "A Citation Study of American Literature: Implications for Collection Management," *Collection Management* 8, no. 2 (Summer 1986): 49-62, https://doi.org/10.1300/J105v08n02_04; Juris Dilevko and Keren Dali, "Improving Collection Development and Reference Services for Interdisciplinary Fields through Analysis of Citation Patterns: An Example Using Tourism Studies," *College & Research Libraries* 65, no. 3 (May 2004): 216-241, https://doi.org/10.5860/crl.65.3.216; dos Santos, Peroni, and Mucheroni, "Referencing Behaviours," 443-455; Tim C.E. Engels, Andreja Istenič Starčič, Emanuel Kulczycki, Janne Pölönen, and Gunnar Sivertsen, "Are Book Publications disappearing from Scholarly Communication in the Social Sciences and Humanities?" *Aslib Journal of Information Management* 70, no. 6 (2018): 592-607, https://doi.org/10.1108/AJIM-05-2018-0127; Björn Hammarfelt, "Interdisciplinarity and the Intellectual base of Literature Studies: Citation Analysis of Highly Cited Monographs," *Scientometrics* 86, no. 3 (March 2011): 705-725, https://doi.org/10.1007/s11192-010-0314-5; Alberto Martín-Martín, Enrique Orduna-Malea, Mike Thelwall, and Emilio Delgado López-Cózar, "Google Scholar, Web of Science, and Scopus: A Systematic Comparison of Citations in 252 Subject Categories," *Journal of Informetrics* 12, no. 4 (November 2018): 1160-1177, https://doi.org/10.1016/j.joi.2018.09.002; C. Keith Waugh and Margie Ruppel, "Citation Analysis of Dissertation, Thesis, and Research Paper References in workforce Education and Development," *Journal of Academic Librarianship* 30, 4 (July 2004): 276-284, https://doi.org/10.1016/j.acalib.2004.04.003; Gregory Youngen, "Multidisciplinary Journal Usage in Veterinary Medicine: Identifying the Complementary Core," *Science & Technology Libraries* 30, no. 2 (2011): 194-201, https://doi.org/10.1080/0194262X.2011.569257 .
23. Huang and Chang, "Characteristics of Research Output," 1819-1828.
24. Krampen, Weiland, and Wiesenhütter, "Citation Success of Different Publication Types," 827-840.
25. Ibid., 321-322.
26. Miika Kumpulainen and Marko Seppänen, "Combining Web of Science and Scopus Datasets in Citation-Based Literature Study," *Scientometrics* 127, no. 10 (October 2022): 5614, https://doi.org/10.1007/s11192-022-04475-7.
27. Pranckutė, "Web of Science (WoS) and Scopus," 3-27.
28. Johnson, *Fundamentals of Collection Development*, 300-302; Levine-Clark and Gill, "A New Comparative Citation Analysis," 145-156.
29. Pranckutė, "Web of Science (WoS) and Scopus," 8.
30. Nick Haupka, Jack H. Culbert, Alexander Schniedermann, Najko Jahn, and Philipp Mayr, "Analysis of the Publication and Document Types in OpenAlex, Web of Science, Scopus, Pubmed and Semantic Scholar," Unpublished manuscript, June 2024, 17.
31. Paul Donner, "Document Type Assignment Accuracy in the Journal Citation Index Data of Web of Science," *Scientometrics* 113, no. 1 (October 2017): 219-236, https://doi.org/10.1007/s11192-017-2483-y; Yu V. Mokhnacheva, "Document Types Indexed in WoS and Scopus: Similarities, Differences, and their Significance in the Analysis of Publication activity," *Scientific and Technical Information Processing* 50, no. 1 (March 2023): 40-46, https://doi.org/10.3103/S0147688223010033.
32. Günter Krampen, Peter Weiland, and Jürgen Wiesenhütter, "Citation Success of Different Publication Types: A Case Study on All References in Psychology Publications from the German-Speaking Countries (D-A-CH-L-L) in 2009, 2010, and 2011," *Scientometrics* 104, no. 3 (September 2015): 828, https://doi.org/10.1007/s11192-015-1573-y.
33. Gregorio González-Alcaide, Juan Carlos Valderrama-Zurián, and Rafael Aleixandre-Benavent, "The Impact Factor in Non-English-Speaking Countries," *Scientometrics* 92, no. 2 (August 2012): 297-311, https://doi.org/10.1007/s11192-012-0692-y; Günter Krampen, "Introduction and Some Ideas as Well as Visions on an Open Access European Psychology

Publication Platform," *Psychology Science Quarterly* 51, Sup. 1 (2009): 3-18; Krampen, Weiland, and Wiesenhütter, "Citation Success," 827-840.
34. Adebowale Jeremy Adetayo, "Artificial Intelligence Chatbots in Academic Libraries: The Rise of ChatGPT," *Library Hi Tech News* 40, no. 3 (2023): 18-21, https://doi.org/10.1108/LHTN-01-2023-0007.
35. Muhammad Asim, Muhammad Arif, Muhammad Rafiq, and Rafiq Ahmad, "Investigating Applications of Artificial Intelligence in University Libraries of Pakistan: An Empirical Study," *Journal of Academic Librarianship* 49, no. 6 (November 2023), https://doi.org/10.1016/j.acalib.2023.102803.
36. Andrew Cox, "How Artificial Intelligence Might Change Academic Library Work: Applying the Competencies Literature and the Theory of the Professions," *Journal of the Association for Information Science and Technology* 74, no. 3 (March 2023): 367-380, https://doi.org/10.1002/asi.24635.
37. Sara Mannheimer, et al., "Responsible AI Practice in Libraries and Archives: A Review of the Literature," *Information Technology and Libraries* 43, no. 3 (September 2024), https://doi.org/10.5860/ital.v43i3.17245.
38. Alesia Zuccala, Maarten van Someren, and Maurits van Bellen, "A Machine-Learning Approach to Coding Book Reviews as Quality Indicators: Toward a Theory of Megacitation," *Journal of the Association for Information Science and Technology* 65, no. 11 (November 2014): 2248-2260, https://doi.org/10.1002/asi.23104.
39. Ivan Portillo and David Carson, "Making the Most of Artificial Intelligence and Large Language Models to Support Collection Development in Health Sciences Libraries," *Journal of the Medical Library Association* 113, no. 1 (January 2025): 92-93, https://doi.org/10.5195/jmla.2025.2079.
40. Ross Hanney, "Reorienting Collection Analysis: Cost-Effective Item-Level Analysis and Machine Learning in Public Libraries," *Information Technology and Libraries* 42, no. 4 (December 2023), https://doi.org/10.5860/ital.v42i4.16987.
41. "About This Journal," North American Actuarial Journal, accessed May 14, 2025, https://www.tandfonline.com/journals/uaaj20/about-this-journal.
42. "Journal of Risk and Insurance," Ulrichsweb, accessed May 13, 2025, https://ulrichsweb.serialssolutions.com/title/1742929633217/48357.
43. "Journal of Risk and Insurance," American Risk and Insurance Association, accessed May 14, 2025, https://www.jri.pub/.
44. L.L. Colquitt and D.W. Sommer, "A Citation Analysis of Risk and Insurance Journals: 2011-2015," *Journal of Risk Education* 7, no. 1 (2016): 62-72; L.L. Colquitt, D.W. Sommer, and W.L. Ferguson, "A Citation Analysis of Risk, Insurance, and Actuarial Research: 2001 through 2005," *Journal of Risk and Insurance* 76, no. 4 (December 2009): 933-953, https://doi.org/10.1111/j.1539-6975.2009.01331.x; L.L. Colquitt, "An Analysis of Risk, Insurance, and Actuarial Research: Citations from 1996-2000," *Journal of Risk and Insurance* 70, no. 2 (June 2003): 315-338, https://doi.org/10.1111/1539-6975.00062; L.L. Colquitt, R.E. Dumm, and S.G. Gustavson, "Risk and Insurance Research Productivity: 1987-1996," *Journal of Risk and Insurance* 65, no. 4 (December 1998): 711-741, https://doi.org/10.2307/253808; L.L. Colquitt, "Relative Significance of Insurance and Actuarial Journals and Articles: A Citation Analysis," *Journal of Risk and Insurance* 64, no. 3 (September 1997): 505-527, https://doi.org/10.2307/253762.
45. Michael E. Rose and John R. Kitchin, "pybliometrics: Scriptable Bibliometrics using a Python Interface to Scopus," SoftwareX 10 (2019): 100263.
46. Huyen, "AI Engineering."
47. Han, Kamber, and Pei, "Data Mining," 109-110.
48. Luis G. Serrano, *Grokking Machine Learning* (Manning Publications, 2021), 177-204; Jiawei Han, Micheline Kamber, and Jian Pei, *Data Mining: Concepts and Techniques* 3rd ed. (Elsevier/Morgan Kaufmann, 2012), 364-369.

49. "F1 Score in Machine Learning," Encord, accessed August 15, 2025, https://encord.com/blog/f1-score-in-machine-learning/?utm_source=chatgpt.com.
50. "F1 Score in Machine Learning: All you Need to Know in 2025," Futurense, accessed August 15, 2025, https://futurense.com/uni-blog/f1-score-machine-learning.
51. Cynthia Rudin, "Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead," *Nature Machine Intelligence* 1, no. 5 (May 2019): 206-215, https://doi.org/10.1038/s42256-019-0048-x.
52. Robert A. Buchanan, "Accuracy of Cited References: The role of Citation Databases," *College & Research Libraries* 67, no. 4 (July 2006): 292-303, https://doi.org/10.5860/crl.67.4.292.; James H. Sweetland, "Errors in Bibliographic Citations: A Continuing Problem," *Library Quarterly: Information, Community, Policy* 59, no. 4 (October 1989): 291-304, http://www.jstor.org/stable/4308405.
53. Late, Tenopir, Talja, and Christian, "Reading Practices," 496.
54. Huang and Chang, "Characteristics of Research Output," 1819-1828.
55. Samantha Godbey and Starr Hoffman, "Characteristics of United States Academic Libraries in 2020 and Regional Changes from 1996 to 2020," *College & Research Libraries* 85, no. 2 (March 2024): 295; Association of Research Libraries, *ARL Statistics 2023* (Association of Research Libraries, 2023); Association of Research Libraries, *ARL Statistics 2022* (Association of Research Libraries, 2022); Association of Research Libraries, *ARL Statistics 2021* (Association of Research Libraries, 2021); Association of Research Libraries, *ARL Statistics 2018/19* (Association of Research Libraries, 2019).