EREA: ENHANCED RESEARCH EXPLORATION AND ANALYSIS
CONFERENCE SUBMISSIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

The increasing volume of scientific publications poses challenges for researchers in efficiently identifying relevant literature, synthesizing research trends, and exploring emerging ideas. Manual search and analysis processes are time-consuming and often insufficient for capturing complex citation relationships. This project presents an open-source Python-based system, EREA (Enhanced Research Exploration and Analysis), that integrates generative artificial intelligence, automated information retrieval, semantic vector search, and citation-based visualization to support enhanced research exploration. User-defined queries are processed to extract structured keywords, retrieve scholarly articles from Google Scholar, and supplement metadata using OpenAlex. Retrieved data are structured, and embedded in a vector database for semantic retrieval, and visualized through interactive, offline HTML graphs. A research report is generated through large language model-assisted synthesis. Developed according to the FAIR (Findability, Accessibility, Interoperability, and Reusability) Data Principles, the system accelerates research exploration, provides structured thematic insights, facilitates understanding through visual citation networks, and supports the identification of research gaps and future directions.

1 INTRODUCTION

The pursuit of knowledge is invariably accompanied by the development and innovation of tools. As the volume of scientific literature continues to proliferate in all academic disciplines(Bornmann & Mutz, 2015; Bornmann et al., 2021; Walsh & Rowe, 2023; Andersen et al., 2024; Gusenbauer & Haddaway, 2020), the process of identifying relevant research, understanding scholarly trends(Bornmann et al., 2021), and organizing knowledge into actionable insights(Peng et al., 2023) has become increasingly complex(Gusenbauer, 2019; Gusenbauer & Haddaway, 2020; Jin et al., 2024). This unwelcome revelation has prompted exploration of new solutions and tools by the scientific community. Whenever the magnitude of tasks at hand exceeds human beings can handle, either in their minds or with tools currently available, new tools and methods are created (Alordiah et al., 2023). The utilization of Enhanced Research Exploration and Analysis (EREA) is particularly pertinent in this context.

Before the emergence of general-purpose online academic search engines, scholars and researchers rely on a cumbersome, time-consuming, and expensive pathway to undertake research work, such as paid libraries and databases (Alordiah et al., 2023; Mierzecka et al., 2017; Gusenbauer, 2019). As research needs can hardly be met under such circumstances and the advancement of digitalization, information retrieval (IR), lower-cost, but faster mass storage and computing, interested stakeholders, including academia, industry, and the general public, turned to Journal Storage (JSTOR) (Alordiah et al., 2023), Google Scholar (Gusenbauer, 2019), ResearchGate (Francke & Hammarfelt, 2022), and Elsevier (Gusenbauer, 2019) for help with their research work (Mierzecka et al., 2017; Gusenbauer & Haddaway, 2020). Then, the advent of technological innovation makes a profound impact on tools and platforms like Semantic Scholar, OpenAlex and The Lens along with citation tracking tools like Connected Papers (Culbert et al., 2025; Singh et al., 2023; Ammar et al., 2018; Penfold, 2020; Gusenbauer & Haddaway, 2021).

Although these services are effective in retrieving indexed documents, they offer limited assistance in structured exploration(Gusenbauer & Haddaway, 2020), thematic analysis(Gusenbauer & Gauster, 2025), and visualization of the citation network(González-Márquez et al., 2024). These limitations inevitably lead to inefficiencies in workflows of the literature review and a reduced capacity to generate data-driven research directions (Gusenbauer & Haddaway, 2020). Our project EREA is developed to address this issue by synthesizing large language models (LLMs) and website scraping.

To be specific, LLMs like Gemini and ChatGPT, they do provide researchers related research work, however, not all the citing papers are available for researchers to study. Existing LLMs do not take into account a considerable number of important papers (Wu et al., 2025; Zhang et al., 2025). EREA can not only find the most important and relevant papers on targeted topics and themes, also provide an interactive citation graph with multiple accessible functions to users as a research panorama. In particular, we differentiate our method from Connected Papers (https://www.connectedpapers.com) in that we use a citation tree rather than organize papers according to their similarity by presenting papers that directly cite each other and complementing the function Connected Papers does not include.

Our system introduces an open-source Python-based method that facilitates optimized research exploration and analysis through a modular, automated pipeline that integrates generative artificial intelligence (Gen AI), web-based information retrieval, semantic vector search, and graph-based visualization. The method is designed to extract and structure bibliographic data from publicly available sources, build a personalized research database, and generate interpretive output that summarizes research trends, citation relationships, and related work.

Key components of EREA include topic extraction using the function-calling capabilities of LLMs, automated paper retrieval through website scraping tools, citation expansion of retrieved articles, semantic embedding of paper summaries for similarity search, and generation of structured outputs such as interactive citation graphs and textual research reports. These components are designed for modular reuse and are published as part of an open-source framework.

Data, especially research data, are the building blocks of science. The Open Science Framework (OSF) (https://osf.io) and the FAIR Guiding Principles (https://www.go-fair.org/fair-principles/) to encourage and enhance the realization of research data sharing. The entire EREA workflow adheres to the FAIR Principles, ensuring that research data and output are findable, accessible, interoperable, and reusable (Wilkinson et al., 2016; Mons et al., 2017; Wilkinson et al., 2019). Our system is designed to support reproducible and scalable literature analysis and to promote transparent workflows in the exploration of scientific knowledge. In addition, our method can be merged into other research infrastructures or extended by the scientific community for enhanced capabilities, including long-term tracking of research topics, collaboration mapping, or integration with data repositories. The synthesis of multiple platforms and the use of Unified Retrieval Interface (URI) contribute greatly to the Findability Principle. The application of FAIR principles avoids the disadvantage that data are ”hiding in the basement” in an old and rigid research environment, and makes it accessible to a wider audience (Wilkinson et al., 2016; Mons et al., 2017; Wilkinson et al., 2019).

We store EREA in both Zenodo (https://zenodo.org) and GitLab (https://gitlab.com/). Zenodo provides the Digital Object Identifier (DOI) for EREA to make it citable and findable, and GitLab offers a better platform for sharing, displaying, and discussing. The two platforms make EREA searchable, open-free, and universally implementable, which in turn enables EREA to be more FAIRer. It is crucial to break the barrier of access permission or right, simplify the research process, avoid inefficiency in using research resources such as computational resources, and lower coordinating cost and research time. We present how to use EREA in detail in Section Methods of this paper and also provide documentations and examples in Zenodo and GitLab to demonstrate how to use EREA.

In this paper, we first present results that include our solution, data sources, output data, and use cases. Then, we discuss benefits, limitations, and future work. In the following, we show the methods to implement the proposed system.

2 RESULTS

2.1 OUR SOLUTION

EREA is open-source and Python-based. It employs to assist various users to acquire knowledge, explore potential research opportunities, or deepen understanding of their current work. The entire system incorporates multiple components - Gen AI, automatic IR, semantic vector search, citation-based visualization, and reporting - to support and assist research process. Each module is designed with adherence to the FAIR principles to ensure transparency, extensibility, and reproducibility (Wilkinson et al., 2016; Rocca-Serra et al., 2023).

The system begins by processing user queries through Gen AI function-calling to extract structured keywords, which guide the retrieval of scholarly articles using website search scraping tools - SerpApi (https://serpapi.com). Missing metadata, such as author information, is supplemented through OpenAlex (Culbert et al., 2025; Cao et al., 2025). For each article, a summary is generated based on open-access content, including scholarly journals and articles made freely available online, or other publicly available information, such as public records and government reports, depending on accessibility.

Citation relationships are mapped by scraping citing papers from Google Scholar and retrieving cited references from OpenAlex (Culbert et al., 2025; Cao et al., 2025). All data — including article metadata, generated summaries, and citation links — are structured into a pandas DataFrame (Wes McKinney, 2010) and exported as a comma-separated values (CSV) file for future use. To support semantic search, the article information is embedded using LLMs and stored in Chroma (https://www.trychroma.com), an open-source vector database (Pan et al., 2024).

An interactive citation network is generated using NetworkX and Plotly (Hagberg et al., 2008), with the visualization exported to a hypertext markup language (HTML) file for offline access. This enables dynamic exploration of research landscapes while preserving usability in different environments.

By integrating structured data extraction, semantic search capabilities, and interactive visualization within a reproducible framework, EREA provides researchers with a scalable method to explore the academic literature and identify research directions.

2.2 DATA SOURCES

Google Scholar is used as the primary bibliographic data source for this study, due to its unparalleled coverage and scope. Google Scholar indexes a wide range of scholarly literature in all disciplines, including not only journal articles but also conference proceedings, theses, books, and other academic documents (Martín-Martín et al., 2018; Gusenbauer, 2019; Martín-Martín et al., 2021; Gusenbauer, 2021; Delgado-Quirós & Ortega, 2024). Several studies have shown that the Google Scholar database is the most comprehensive among academic search platforms (Gusenbauer, 2019; Gusenbauer & Haddaway, 2020): An estimated 389 million records are indexed in Google Scholar, making it larger than conventional citation databases (Gusenbauer, 2019). In terms of citation coverage, Google Scholar has been shown to encompass essentially all citations found in Web of Science and Scopus, plus additional ones that these databases miss (Culbert et al., 2025; Cao et al., 2025). In other words, its citation index is effectively a superset of content in major commercial databases (Culbert et al., 2025; Cao et al., 2025). This broad coverage ensures that our analysis captures as many relevant publications and citations as possible, including publications not indexed elsewhere.

However, despite its superior size and coverage, Google Scholar has well-documented limitations regarding metadata completeness and data accessibility. Google Scholar's bibliographic records are often incomplete or inconsistently formatted compared to those in curated scholarly databases (Céspedes et al., 2025). For example, Google Scholar can sometimes misidentify or omit metadata elements; past analyses have noted cases of missing author names and other parsing errors in Google Scholar's records. In particular, Google Scholar does not provide full reference lists (the list of works cited by a given publication) through its interface. It only displays the citing works (the "Cited by" list), and offers no public application programming interface (API) to retrieve structured reference data for a given item. This means that while Google Scholar can tell how many times a paper has been cited, it does not readily expose which references that paper itself contains. These deficiencies

in metadata and reference information necessitate additional measures to obtain a complete picture of the bibliographic data.

To complement Google Scholar and address these gaps, we have incorporated OpenAlex (Culbert et al., 2025; Cao et al., 2025) as a secondary data source. OpenAlex is an open index of scholarly works that provides rich, structured metadata and open citation data for a vast number of publications (Culbert et al., 2025; Cao et al., 2025). Each record in OpenAlex comes with standardized metadata fields, for example title, authors, publication venue, publication date, DOI, as well as lists of references and citations, allowing direct extraction of an article's reference list.

In summary, Google Scholar's expansive coverage provides a broad foundation for identifying relevant literature and citations, while OpenAlex supplies the detailed metadata and open citation information that Google Scholar cannot furnish. This combination leverages the strengths of both sources: Google Scholar ensures that no important records are overlooked, and OpenAlex allows us to retrieve standardized metadata (including each publication's reference list) for comprehensive analysis. Such a multi-source approach is in line with recommendations in bibliometric research to combine databases in order to compensate for the limitations of any single data source (Culbert et al., 2025; Cao et al., 2025). Using Google Scholar as the primary search platform and OpenAlex to supplement missing bibliographic details, we maintain both breadth and precision in our data, which is essential for a robust research analysis.

2.3 Output Data

The output data of this study are structured and disseminated using multiple standard formats to ensure accessibility and ease of analysis. First, all tabular results are archived as CSV files. The use of CSV maximizes compatibility, allowing the data to be opened and processed by a wide array of software tools without specialized conversion.

Second, a vector database (Chroma) is employed to enable semantic information retrieval from the results. In this approach, high-dimensional embedding vectors (encodings of textual or numeric data features) are stored in the database, so that queries can retrieve records by meaningful similarity rather than exact keyword matching (Pan et al., 2024). The vector database supports fast nearest-neighbor search in the embedding space using approximate algorithms, such as hashing and graph-based indexes, to find the closest vectors efficiently. This allows researchers to quickly locate relevant entries including documents and data points with related content, based on semantic proximity, improving the effectiveness of information retrieval from the output corpus.

Third, the study's interactive figures and charts are distributed as self-contained HTML files. Each HTML file bundles the visualization code (JavaScript/HTML/CSS) and the data, enabling the charts to render interactively in any modern web browser without requiring an Internet connection or a separate server-side component. This approach (using libraries such as NetworkX and Plotly) ensures that collaborators can explore the visualizations offline. It offers practical benefits for dissemination – for example, an interactive HTML report can be emailed to stakeholders, who can open it directly and examine dynamic features like tooltips, zooming, or filtering, all without installing additional software. By leveraging an open web format, the interactive output remains widely accessible and preserves full functionality across different operating systems.

Finally, an automatic textual report is generated to summarize and contextualize the findings, using a LLM as the writing assistant. Recent advances in LLM technology have demonstrated the capability to produce coherent, contextually relevant summaries of complex scientific content, effectively blending extractive and abstractive summarization strategies (Zhang et al., 2025; Laskar et al., 2023b;a). In this work, the LLM is prompted with the key results and metadata, allowing it to synthesize a narrative that highlights the main insights. The generated summary is generally fluent and semantically rich, benefiting from the model's pre-trained knowledge of scientific language and concepts (Zhang et al., 2025; Laskar et al., 2023b;a). This automation significantly accelerates the report-writing process and helps ensure consistency in how results are described. The combination of structured CSV data, a semantic vector database, portable HTML visualizations, and LLM-assisted summaries provides a comprehensive and user-friendly record of the research outputs, facilitating both reuse by other researchers and effective communication of the results.

2.4 USE CASES

To demonstrate the applicability of the system, we present a representative use case in which the user query is: "*I would like to know more about human trafficking in economics.*" The system first applies its function-calling mechanism to extract the primary research topic, identifying "*human trafficking in economics*" as the target area.

Following the multi-stage pipeline, the system produces four distinct outputs. First, a structured dataset as Table 1 is generated, containing metadata such as article identifiers, titles, publication years, author names, citation counts, summaries, and article links. Second, a semantic vector database as Table 2 is constructed from the same dataset using customized embeddings, allowing contextual queries and retrieval of semantically related materials. Third, an interactive citation graph is produced to visualize citation relationships between the retrieved articles, shown as Figure 1. Finally, a research report is generated, summarizing current research trends, identifying gaps in the literature, and suggesting potential future research directions, shown as Figure A.1.1.

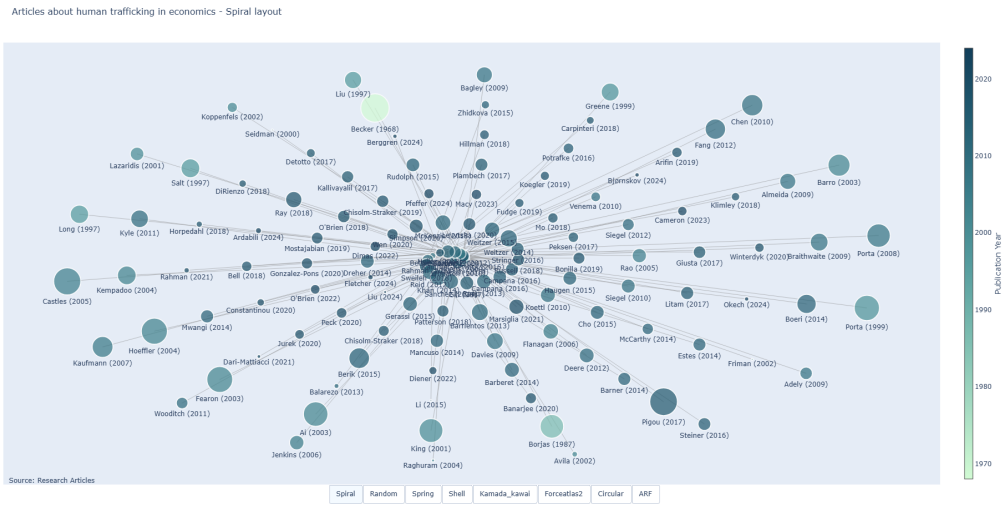

Figure 1: Interactive graph. Each node represents a research article, and each directed edge denotes a citation from one article to another. Node size is proportional to the article's citation count — larger nodes indicate more frequently cited publications. Node color reflects publication year: more recent articles are shown with more intense colors, while older articles appear in lighter shades. Users may switch between layout styles by clicking the layout button to optimize the network's spatial arrangement.

Additional sample outputs for other research topics are available in the samples folder within the project's GitLab repository: https://github.com/csmichael-ai/erea/tree/main/examples. Further technical details on each stage of the process can be found in the Section Methods of this article.

3 DISCUSSION

This project presents a modular open-source system to support enhanced research exploration through the integration of Gen AI, automated information retrieval, semantic vector databases, and interactive visualization of citation graphs. Google Scholar is selected as the primary data source for this study owing to its superior coverage of scholarly publications across disciplines compared with other bibliographic databases (Martín-Martín et al., 2018; Gusenbauer, 2019; Martín-Martín et al., 2021; Gusenbauer, 2021; Delgado-Quirós & Ortega, 2024). However, it should be noted that Google Scholar applies extensive safeguards against automated extraction. To address this, the system employs SerpAPI to reliably extract information from Google Scholar (Gusenbauer & Gauster, 2025).

article_id	*article_name*	*publication_year*	*author_name*	*citations*	*summary*	*link*	*citing_papers*
12348...	Economics of human traffi...	2010	['EM Wheat...	403	The econom...	https://on...	['12792315...
15082...	The economics of human tr...	2010	['C Trebes...	227	The study ...	https://ww...	['58334777...
38435...	Human trafficking, modern...	2009	['J Koettl...	93	Human traf...	https://do...	['15471832...
17041...	Human trafficking by the ...	2016	['M Kammer...	88	The "Human...	https://re...	['17766271...
61423...	Human trafficking in Euro...	2004	['Gisjbert...	57	The articl...	https://ec...	['35878196...
83416...	The determinants of human...	2012	['Alicja J...	87	The articl...	https://on...	['10943393...
14145...	Understanding human traff...	2012	['S Rao']	125	The articl...	https://ww...	['16925584...
45272...	Is human trafficking the ...	2018	['LR Helle...	22	The articl...	https://ww...	['64804140...
12350...	The economics of slavery ...	2008	['Patrick ...	9	"The Econo...	https://ww...	['11709456...
18511...	Human trafficking in the ...	2011	['MA Rahma...	76	"Human Tra...	https://ww...	['87951270...
...	...	...	...	...	...	...	...
14684...	Comparing Freedom House d...	2016	['ND Stein...	106	The articl...	https://ww...	['14145794...
10447...	EDUCATING WOMEN FOR DEVEL...	2009	['FJ Adely...	96	The articl...	https://do...	['14145794...
14500...	The quality of government	1999	['R La Por...	9498	The articl...	https://do...	['45272563...
11456...	The economic consequences...	2008	['RL Porta...	4834	The articl...	https://do...	['45272563...
55833...	Religion and economic gro...	2003	['RJ Barro...	2975	The study ...	https://do...	['45272563...
17993...	The developing world is p...	2010	['S Chen']	2311	The articl...	https://do...	['45272563...

Table 1: The output structured data containing article metadata and summaries. The table cannot show all details for the data, and "..." means there is more information in the data. The *article_id* is a unique number that can identify the article; the *article_name* is the article name; *author_name* is a list that contains author names; citations is the citation number; summary is a generated summary of the article; link is a website link that can open the article; *citing_papers* is a list that contains other citing or cited articles with *article_id*.

EREA is also explicitly designed to comply with the FAIR Data Principles (Findability, Accessibility, Interoperability, and Reusability) with the full source code, documentation, and examples made publicly available on GitLab and Zenodo. The use of structured CSV files ensures compatibility and reusability across a wide range of scientific workflows, while the deployment of a semantic vector database (Chroma) allows users to retrieve information based on conceptual similarity rather than exact matching. Exporting of interactive visualizations as offline HTML files further facilitates portability and broadens access to dynamic citation analyses without dependency on specific computational environments. In addition, automated generation of a research report through LLM assistance streamlines the synthesis of research trends, gaps, and directions, improving researchers' ability to quickly navigate complex academic landscapes.

Together, these features contribute to a scalable, extensible, and reproducible framework for research analysis, supporting both individual researchers and collaborative research teams seeking to conduct literature reviews, trend analyses, and research gap identification with increased efficiency and consistency.

3.1 Limitations

While EREA alleviates issues to substantially assist structured exploration, thematic analysis, and visualization of the citation networks, several limitations must be acknowledged. First, although EREA, by integrating OpenAlex into the system thus complementing missing metadata, it may not cover all records indexed by Google Scholar in some rare cases, potentially resulting in incomplete author or reference information for certain publications. Second, the semantic embeddings used for vector search depend on the quality of the LLM-generated summaries and metadata; inaccuracies or inconsistencies in source content may propagate through the vector representations. Additionally, while the generated research reports are generally coherent, LLM-based summarization is inherently probabilistic and may occasionally produce minor factual inaccuracies if not carefully reviewed by the user (Hosseini et al., 2024; Moëll & Sand Aronsson, 2025).

3.2 Future work

From the initial design stage of EREA, we have a goal of contributing to the vision of Open Science and maximizing the use of research resources and providing powerful and efficient assistance to researchers with their work. Future development of EREA will focus on the integration of LangGraph (https://www.langchain.com/langgraph), an open-source AI agent framework to construct, implement and manage stateful graph-based applications on LLM. Incorporating LangGraph will allow EREA to maintain user interaction histories, enabling iterative and memory-enhanced exploration. This can significantly improve the overall user experience during research navigation.

ids	embeddings	documents
0	[0.02994205, -0.0012782, -0.00452825, ..., -0.06318714, 0.02563175, -0.00735727]	*article name*: Economics of human trafficking, *publication year*: 2010, *author name*: ["EM Wheaton"], *citations*: 416, *summary*: The "Economics of Human Trafficking" is a study that presents an economic model encompassing factors affecting human trafficking across and within national borders. It envisions human trafficking as a monopolistically competitive industry where traffickers act as intermediaries between vulnerable individuals and employers, supplying differentiated products, which are human beings ...
1	[0.01773802, 0.00254536, -0.09226844, ..., -0.04077326, 0.02636305, 0.00223326]	*article name*: The economics of human trafficking and labour migration: Micro-evidence from Eastern Europe, *publication year*: 2010, *author name*: ["C Trebesch"], *citations*: 228, *summary*: The study "The economics of human trafficking and labour migration: Micro-evidence from Eastern Europe" analyzes the economics of human trafficking and labor migration using microdata from household surveys in Belarus, Bulgaria, Moldova, Romania, and Ukraine ...
2	[-0.02992962, 0.03398041, -0.01619028, ..., -0.04797086, 0.08280791, 0.00298904]	*article name*: Human trafficking, modern day slavery, and economic exploitation, *publication year*: 2009, *author name*: ["J Koettl"], *citations*: 97, *summary*: Human trafficking, also referred to as trafficking in persons or modern-day slavery, involves the exploitation of individuals for profit through forced labor or commercial sex. It is a grave crime and human rights abuse that undermines national and economic security, the rule of law, and the well-being of communities ...
...	...	...

Table 2: The first three records from the vector database. Due to the length of the embeddings and stored information, these entries are truncated using "...", and some metadata is omitted for brevity. The *ids* column contains the unique identifiers in the vector database, the *embeddings* column contains the corresponding embedding vectors generated from the stored information in *documents*, and the *documents* column stores each record from the previously structured dataframe.

Furthermore, EREA has the potential to not only apply to academic research, but also to multi-industrial processes. It can be used in a number of scenarios such as providing professional solutions and strategies in consulting industry, market and industry analysis of financial services and more. With the help of further development, EREA can also be extended to medical research, pharmaceutical innovation, legal services and other fields. We aim to develop a system standing to benefit the scientific community, professionals and technical people with research needs at all levels, extending to individuals with a variety of interests and needs. These endeavors and works will also be part of supporting the Open Science movement and the FAIR Principle.

4 METHODS

4.1 EREA SYSTEM

The EREA system is a modular, open-source framework designed to support automated literature discovery, exploration, and summarization based on user-defined research queries. Upon receiving a query, EREA executes a multi-stage pipeline to generate multiple coordinated outputs, as Figure 2. This pipeline includes automated function calling, topic extraction, relevant paper retrieval, citation network expansion, summary generation, structured data organization, semantic vector database creation, interactive visualization, and research report generation. These functions are discussed in the following sections.

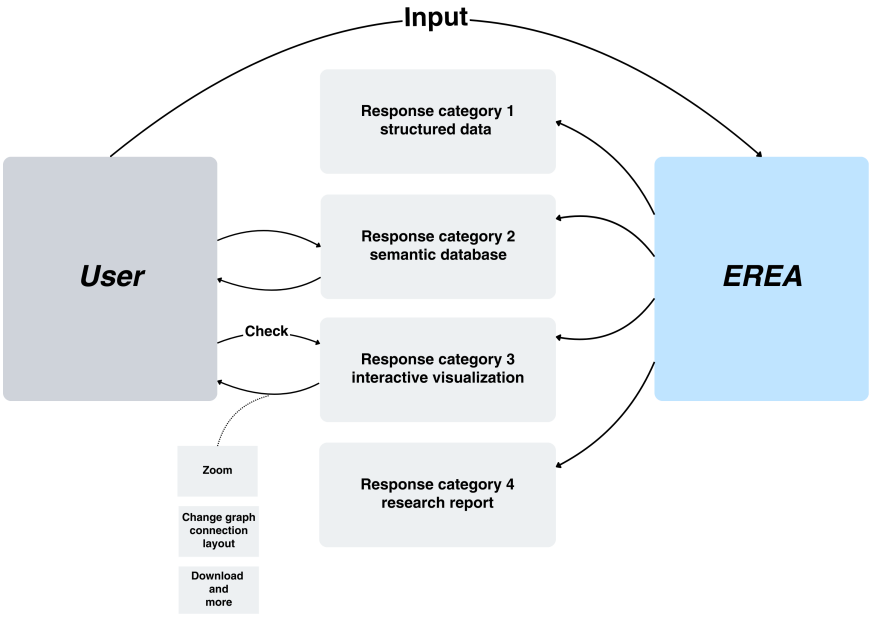

Figure 2: EREA overview.

The system produces four main types of responses. First, it generates a structured dataset in tabular format (CSV), which includes article metadata, citation counts, summaries, and citation relationships. Second, a semantic database is created using Chroma. This database allows users to interact with the embedded content using natural language queries, enabling the retrieval of semantically relevant articles and information beyond exact matches.

Third, an interactive citation graph is generated using NetworkX and Plotly (Hagberg et al., 2008). The graph enables users to explore citation relationships visually, with functionality for zooming, panning, switching between graph layout (e.g., spiral, spring, shell), and exporting the visualization as a static image. Each node in the graph corresponds to a scholarly article, with dynamic tooltips that reveal additional metadata and summaries upon hovering.

Finally, EREA generates a text-based research report that synthesizes trends, gaps, and directions within the retrieved corpus. Together, these outputs offer a comprehensive and reproducible research exploration tool that facilitates both a high-level overview and detailed literature interrogation.

4.1.1 Automated function calling

To enable structured and context-aware interaction between the LLM and specific tasks, we implement a function-calling mechanism that allows the LLM to autonomously invoke predefined functions as needed, shown as Figure 3.

A list of available functions is registered within the execution environment and exposed to the LLM through its function-calling interface. During execution, the LLM dynamically determines whether a function call is required based on the user's input and the system prompt, and selects the appropriate function and provides the corresponding arguments. The current implementation includes some core functions. For example, the *get_topic_keywords* function extract research-relevant keywords from free-text user input. This function is used to standardize query terms for downstream paper retrieval and semantic search tasks. Also, the *get_summary* function generates a concise textual summary from a given input text for integration into structured data.

4.1.2 Topic extraction

The system initiates the LLM chat with automatically function calling enables. Then, the system welcomes the user and briefly telling the capability of the system to user. After the system receives

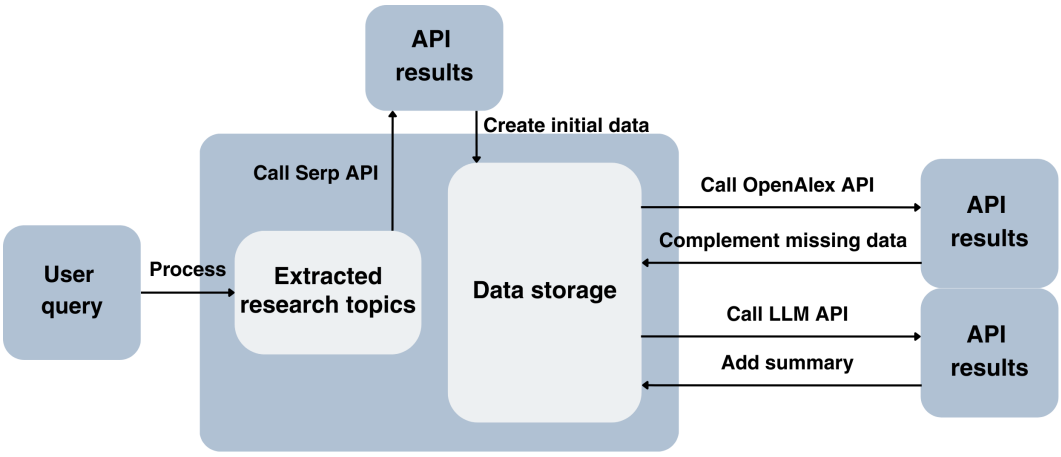

Figure 3: Function calling overview

the input from user, the LLM chat would automatically call the predefined function to extract the research topic for further processing.

4.1.3 Relevant information retrieval

After we extract results from Google Scholar via SerpAPI (Sultan & Abdullah, 2022; Gusenbauer & Gauster, 2025) based on the extracted research topics, we have the information for article unique number, article name, author names, citation number, the year of publication and external link for each article in the extracted results. When there is missing information from the Google Scholar, we use OpenAlex (Culbert et al., 2025; Cao et al., 2025) to complement the missing data. Then, we store key information into pandas DataFrame (Wes McKinney, 2010). For the generated summary, we will discuss it in the summary generation section.

The DataFrame contains following attributes:

- *article_id* : a unique number that can identify the article.
- *article_name* : the article name.
- *publication_year* : the year of publication.
- *author_name* : author names.
- *citations* : the citation number.
- *summary*: a generated summary of the article.
- *link*: a website link that can open the article.

For example, the initial results based on human trafficking in economics are presented in Table 3.

4.1.4 Citation network expansion

For each article retrieved in the initial results, a subsequent search query is executed to identify papers that cite it. The results of these searches are systematically recorded, and a dedicated attribute (*citing_papers*) is established within the structured dataset to store the unique identifiers (*article_id*) of these citing articles. This information facilitates subsequent visualization of citation networks.

Since Google Scholar does not provide comprehensive references (i.e., cited papers) directly within its search results, OpenAlex (Culbert et al., 2025; Cao et al., 2025) is utilized to extract the reference lists from each article. Titles of referenced papers obtained via OpenAlex (Culbert et al., 2025; Cao et al., 2025) are then used as inputs for additional Google Scholar queries to retrieve relevant metadata and ensure consistency and completeness within the dataset.

article_id	*article_name*	*publication_year*	*author_name*	*citations*	*summary*	*link*
12348...	Economics of human traffi...	2010	['EM Wheat...	403	The econom...	https://on...
15082...	The economics of human tr...	2010	['C Trebes...	227	The study ...	https://ww...
38435...	Human trafficking, modern...	2009	['J Koettl...	93	Human traf...	https://do...
17041...	Human trafficking by the ...	2016	['M Kammer...	88	The "Human...	https://re...
61423...	Human trafficking in Euro...	2004	['Gisjbert...	57	The articl...	https://ec...
83416...	The determinants of human...	2012	['Alicja J...	87	The articl...	https://on...
14145...	Understanding human traff...	2012	['S Rao']	125	The articl...	https://ww...
45272...	Is human trafficking the ...	2018	['LR Helle...	22	The articl...	https://ww...
12350...	The economics of slavery ...	2008	['Patrick ...	9	"The Econo...	https://ww...
18511...	Human trafficking in the ...	2011	['MA Rahma...	76	"Human Tra...	https://ww...

Table 3: The initial structured data containing article metadata and summaries. The table cannot show all details for the data, and "..." means there is more information in the data. The *article_id* is a unique number that can identify the article; the *article_name* is the article name; *author_name* is a list that contains author names; citations is the citation number; *summary* is a generated summary of the article; link is a website link that can open the article.

4.1.5 SUMMARY GENERATION

We utilize prompt engineering techniques within the LLM to automatically generate article summaries based on the provided document links (Laskar et al., 2023b;a). With internet search capabilities enabled in the LLM, the system attempts to access each article's direct link to retrieve the full-text content and produce a precise summary. However, certain articles may not be accessible directly due to subscription requirements imposed by publishers. In such cases, the LLM employs general internet search functionality to identify alternative sources, such as openly available abstracts, invited presentations, or scholarly discussions, to synthesize an accurate summary. Each generated summary is immediately recorded in the dedicated *summary* attribute of the structured dataset. To enhance efficiency, the summary generation process is executed concurrently with the information extraction procedure, reducing processing time and streamlining the overall workflow.

4.1.6 DATA ORGANIZATION

The extracted data, including *article_id*, *article_name*, *publication_year*, *author_name*, *citations*, *summary*, *link*, and *citing_papers*, are organized into a structured format using a pandas DataFrame. The DataFrame is then exported to a CSV file for future use or archival purposes (Wes McKinney, 2010).

Based on human trafficking in economics topics, the final structured results are presented in Table 1.

4.1.7 VECTOR DATABASE CREATION AND QUERYING

After the structured data has been stored, the system creates a semantic vector database to facilitate efficient information retrieval and contextual search. The previously extracted and structured data, comprising article identifiers, article titles, publication years, author names, citation counts, summaries, and article links, are stored in Chroma, an open-source vector database (Pan et al., 2024). Before insertion, each data entry is transformed into a high-dimensional embedding using a customized embedding function, enabling semantic similarity search capabilities.

After the creation, the vector database supports interactive querying through natural language input, eliminating the need for traditional database query languages such as SQL (Pan et al., 2024). Users can write queries in plain language and receive semantically relevant results ranked by contextual similarity. This approach enhances the discovery of related research materials and supports more intuitive information retrieval compared to keyword-based searches alone.

4.1.8 INTERACTIVE VISUALIZATION

To facilitate the exploration of citation relationships among scholarly articles, this system generates an interactive citation graph using the NetworkX and Plotly libraries (Hagberg et al., 2008), as Figure 1. Each node in the graph represents a single article, while directed edges denote citation

relationships, where an edge from node A to node B indicates that article A cites article B. This graph-based representation supports intuitive navigation and structural analysis of citation networks.

The visual attributes of each node encode key bibliometric metadata. Node size is proportional to the number of citations received, allowing highly cited articles to be immediately distinguished by their larger appearance. Node color is determined by the publication year: recent articles are assigned more intense colors, while older publications appear in lighter hues. This temporal gradient enables users to visually assess the recency of influential publications within the network.

Interactivity is a central feature of the visualization. When the user hovers over a node, a tooltip appears showing article metadata including the title, author list, publication year, citation count, and a brief summary, as Figure 4. To accommodate different network structures, the interactive graph also supports multiple layout algorithms, including spiral layout, shell layout, random layout, circular layout, and more. Users may dynamically change the layout to optimize visual clarity or highlight specific patterns.

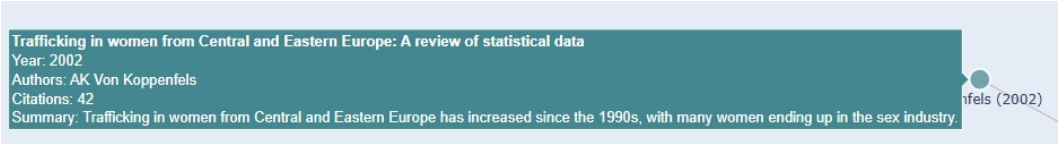

Figure 4: Hovering effect. When hovering the cursor over a node, a tooltip appears displaying key metadata about the corresponding article. This includes the article title, list of authors, publication year, citation count, and a brief summary. This feature enables users to access detailed contextual information without leaving the visual interface.

Additional interactive controls enable zooming, resetting axes, autoscaling, and exporting the current view as a static image. The final visualization is exported as a self-contained HTML file, ensuring that it remains accessible and fully interactive offline. This approach supports reproducibility, ease of sharing, and broad compatibility with modern browsers without requiring server-side infrastructure.

4.1.9 Research report generation

The workflow concludes its analytical workflow by producing a structured research report that synthesizes the results of the prior stages. This report serves as a textual summary of the retrieved and processed literature, consolidating key findings into a format suitable for rapid review and high-level interpretation, as Figure A.1.1.

The report includes three main components. First, it provides a summary of current research trends, identified through the analysis of keyword frequencies, citation prominence, and topical clustering within the extracted corpus. Second, it highlights existing research gaps by identifying underrepresented themes, low-density citation clusters, and topics with sparse publication coverage. Third, it offers suggestions for potential research directions, informed by semantically adjacent areas with limited citation activity and by extrapolation from recent trends in the field.

To enhance contextual understanding, the report can be combined with the interactive citation graph. This pairing enables users to cross-reference textual insights with the underlying citation network, facilitating both narrative interpretation and structural exploration. The graph allows dynamic access to article metadata and summaries, which supports a deeper engagement with the research landscape.

This integration of automated text generation and visual analytics offers several benefits. It enhances interpretability by aligning qualitative summaries with visual structures, supports informed exploration of the literature by highlighting unexplored or emerging areas, and reduces the manual burden of literature review through automated synthesis.

4.2 System deployment

The system is implemented in Python and deployed through two primary interfaces: standalone Python scripts and Jupyter Notebooks. This dual-format design accommodates a range of user preferences and usage scenarios, supporting both automated execution and interactive exploration.

The Python script-based deployment enables streamlined, end-to-end execution of the workflow. It is structured functionally, allowing individual components (e.g., data extraction, vector embedding, visualization, report generation) to be run independently or in sequence. In parallel, a Jupyter Notebook version is provided to facilitate transparency, educational use, and step-by-step interaction. The notebook interface enables users to inspect intermediate outputs, adjust parameters, and visualize results at each stage of the process. It is particularly useful for demonstrations, testing, and adaptation of the workflow to domain-specific research tasks.

4.3 Dependency management and installation

To ensure reproducibility and minimize potential conflicts due to version mismatches or package dependencies, users are recommended to run the project within a dedicated virtual environment. This approach isolates the project's dependencies from those of the host system or other projects, providing a controlled and consistent runtime environment. All required Python packages and their respective versions are listed in a requirements.txt file.

4.4 Unit and functional tests

4.4.1 Validation of installation

The installation process begins by specifying all necessary software packages along with their exact versions in a requirements.txt file, ensuring a consistent and reproducible environment. Following package installation, the system explicitly validates the installed version of the packages upon startup to confirm alignment with the defined requirements. If discrepancies between installed packages and the specified versions are detected, the system will halt execution and raise an informative error message, prompting the user to reinstall the required packages with the correct versions. This validation step ensures reliability and prevents potential runtime issues arising from incompatible or incorrectly versioned software dependencies.

4.4.2 Validation of function calling

To ensure the correct operation of automated function calls within the system, a dedicated function *check_function_call*, has been implemented to check the validaty of function callings. This function takes the chat object as an argument and inspects the LLM's chat history to verify that function calls are executed as intended.

For instance, consider the case where the user input is: "I am new to economics, and I would like to know more about human trafficking.". The *get_topic_keywords* function is designed to extract the research topic from such an input, returning "human trafficking" as the identified topic. An example output of the validation process may appear as function response.

The *name* field confirms that the LLM has automatically invoked the *get_topic_keywords* function, while the *result* field shows that the function has correctly returned "human trafficking" as its output. This process verifies both the correct invocation of the intended function and the accuracy of its result. By systematically validating function calls, the system ensures reliable integration of automated function execution into the overall research exploration workflow.

4.4.3 Data quality control

During data processing, missing values are systematically identified and addressed to maintain dataset completeness and quality. Initially, the dataset is checked for missing entries, and targeted strategies are applied to fill them where possible. In rare cases where Google Scholar does not provide author information in an extractable format, OpenAlex (Culbert et al., 2025; Cao et al., 2025) is employed as a supplementary source to obtain the missing author details. Similarly, instances occur where the publication year field in Google Scholar contains extraneous content. In such cases, regular expressions are applied to filter and extract the valid publication year directly from the Google Scholar page.

For example, for each missing summary, the system iteratively invokes the *fill_missing_summary* function to fill in the absent entry. This process continues until either the missing summary is filled or a predefined maximum number of iteration rounds is reached.

After running all applicable missing value filling procedures, any records with unresolved missing values are removed from the dataset to maintain quality control. Additionally, a post-processing step is applied to further ensure data integrity. Specifically, duplicate entries are eliminated using the drop duplicates function, as the *article_id* serves as a unique identifier for each article.

4.5 Documentation

The documentation can be found in the GitLab repository at: https://github.com/csmichael-ai/erea/. The documentation provides instructions on setting up the environment and running the program step-by-step.

5 Data availability

The data generated by the system is produced dynamically and stored in the output directory upon execution of the source code. These outputs include structured CSV files, a semantic vector database, an interactive graph, and a generated research report. As the system operates based on user-defined input queries and real-time retrieval from external sources, the resulting data may vary with each execution. All outputs are saved in standardized formats to support reuse and reproducibility. Sample output data generated from user queries can be accessed via the project's GitLab repository at: https://github.com/csmichael-ai/erea/tree/main/examples.

6 Code availability

The full source code, including example workflows, documentation, and configuration files, is publicly available in the project's GitLab repository at: https://github.com/csmichael-ai/erea/tree/main. To ensure long-term accessibility and reproducibility, a snapshot of the repository has been archived with Zenodo.

References

Caroline Ochuko Alordiah, Mercy Afe Osagiede, Florence Chiedu Omumu, Isabella Ezinwa Okokoyo, Helena Tsaninomi Emiko-Agbajor, O Chenube, and John Oji. Awareness, knowledge, and utilisation of online digital tools for literature review in educational research. *Heliyon*, 9(1), 2023. doi: https://doi.org/10.1016/j.heliyon.2022.e12669.

Waleed Ammar, Dirk Groeneveld, Chandra Bhagavatula, Iz Beltagy, Miles Crawford, Doug Downey, Jason Dunkelberger, Ahmed Elgohary, Sergey Feldman, Vu Ha, Rodney Kinney, Sebastian Kohlmeier, Kyle Lo, Tyler Murray, Hsu-Han Ooi, Matthew Peters, Joanna Power, Sam Skjonsberg, Lucy Lu Wang, Chris Wilhelm, Zheng Yuan, Madeleine van Zuylen, and Oren Etzioni. Construction of the literature graph in semantic scholar. In Srinivas Bangalore, Jennifer Chu-Carroll, and Yunyao Li (eds.), *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 3 (Industry Papers)*, pp. 84–91, New Orleans - Louisiana, June 2018. Association for Computational Linguistics. doi: https://doi.org/10.18653/v1/N18-3011.

Mikkel Zola Andersen, Philine Zeinert, Jacob Rosenberg, and Siv Fonnes. Comparative analysis of cochrane and non-cochrane reviews over three decades. *Systematic Reviews*, 13(1):120, 2024. doi: https://doi.org/10.1186/s13643-024-02531-2.

Lutz Bornmann and Rüdiger Mutz. Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references. *Journal of the association for information science and technology*, 66(11):2215–2222, 2015. doi: https://doi.org/10.1002/asi.23329.

Lutz Bornmann, Robin Haunschild, and Rüdiger Mutz. Growth rates of modern science: a latent piecewise growth curve approach to model publication numbers from established and new literature databases. *Humanities and Social Sciences Communications*, 8(1):1–15, 2021. doi: https://doi.org/10.1057/s41599-021-00903-w.

Zhe Cao, Lin Zhang, Ying Huang, and Robin Haunschild. How does the academia refer to open research information data sources? a review study based on openalex and microsoft academic series. *Scientometrics*, pp. 1–27, 2025. doi: https://doi.org/10.1007/s11192-025-05347-6.

Lucía Céspedes, Diego Kozlowski, Carolina Pradier, Maxime Holmberg Sainte-Marie, Natsumi Solange Shokida, Pierre Benz, Constance Poitras, Anton Boudreau Ninkov, Saeideh Ebrahimy, Philips Ayeni, et al. Evaluating the linguistic coverage of openalex: An assessment of metadata accuracy and completeness. *Journal of the Association for Information Science and Technology*, 2025. doi: https://doi.org/10.1002/asi.24979.

Jack H Culbert, Anne Hobert, Najko Jahn, Nick Haupka, Marion Schmidt, Paul Donner, and Philipp Mayr. Reference coverage analysis of openalex compared to web of science and scopus. *Scientometrics*, pp. 1–18, 2025. doi: https://doi.org/10.1007/s11192-025-05293-3.

Lorena Delgado-Quirós and José Luis Ortega. Completeness degree of publication metadata in eight free-access scholarly databases. *Quantitative Science Studies*, 5(1):31–49, 2024. doi: https://doi.org/10.1162/qss_a_00286.

Helena Francke and Björn Hammarfelt. Competitive exposure and existential recognition: Visibility and legitimacy on academic social networking sites. *Research Evaluation*, 31(4):429–437, 2022. doi: https://doi.org/10.1093/reseval/rvab043.

Rita González-Márquez, Luca Schmidt, Benjamin M Schmidt, Philipp Berens, and Dmitry Kobak. The landscape of biomedical research. *Patterns*, 5(6), 2024. doi: https://doi.org/10.1016/j.patter.2024.100968.

Michael Gusenbauer. Google scholar to overshadow them all? comparing the sizes of 12 academic search engines and bibliographic databases. *Scientometrics*, 118(1):177–214, 2019. doi: https://doi.org/10.1007/s11192-018-2958-5.

Michael Gusenbauer. The age of abundant scholarly information and its synthesis–a time when 'just google it'is no longer enough. *Research synthesis methods*, 12(6):684–691, 2021. doi: https://doi.org/10.1002/jrsm.1520.

Michael Gusenbauer and Sebastian P Gauster. How to search for literature in systematic reviews and meta-analyses: A comprehensive step-by-step guide. *Technological Forecasting and Social Change*, 212:123833, 2025. doi: https://doi.org/10.1016/j.techfore.2024.123833.

Michael Gusenbauer and Neal R Haddaway. Which academic search systems are suitable for systematic reviews or meta-analyses? evaluating retrieval qualities of google scholar, pubmed, and 26 other resources. *Research synthesis methods*, 11(2):181–217, 2020. doi: https://doi.org/10.1002/jrsm.1378.

Michael Gusenbauer and Neal R Haddaway. What every researcher should know about searching–clarified concepts, search advice, and an agenda to improve finding in academia. *Research synthesis methods*, 12(2):136–147, 2021. doi: https://doi.org/10.1002/jrsm.1457.

Aric Hagberg, Pieter Swart, and Daniel Chult. Exploring network structure, dynamics, and function using networkx. In *Proceedings of the 7th Python in Science Conference*, 06 2008. doi: https://doi.org/10.25080/TCWV9851.

Mohammad Hosseini, Lisa M. Rasmussen, and David B. Resnik. Using ai to write scholarly publications. *Accountability in Research*, 31(7):715–723, 2024. doi: https://doi.org/10.1080/08989621.2023.2168535.

Qiao Jin, Robert Leaman, and Zhiyong Lu. Pubmed and beyond: biomedical literature search in the age of artificial intelligence. *EBioMedicine*, 100, 2024. doi: https://doi.org/10.1016/j.ebiom.2024.104988.

Md Tahmid Rahman Laskar, M Saiful Bari, Mizanur Rahman, Md Amran Hossen Bhuiyan, Shafiq Joty, and Jimmy Huang. A systematic study and comprehensive evaluation of ChatGPT on benchmark datasets. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 431–469, Toronto, Canada, July 2023a. Association for Computational Linguistics. doi: https://doi.org/10.18653/v1/2023.findings-acl.29.

Md Tahmid Rahman Laskar, Xue-Yong Fu, Cheng Chen, and Shashi Bhushan TN. Building real-world meeting summarization systems using large language models: A practical perspective. In Mingxuan Wang and Imed Zitouni (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pp. 343–352, Singapore, December 2023b. Association for Computational Linguistics. doi: https://doi.org/10.18653/v1/2023.emnlp-industry.33.

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(4):1160–1177, 2018. doi: https://doi.org/10.1016/j.joi.2018.09.002.

Alberto Martín-Martín, Mike Thelwall, Enrique Orduna-Malea, and Emilio Delgado López-Cózar. Google scholar, microsoft academic, scopus, dimensions, web of science, and opencitations' coci: a multidisciplinary comparison of coverage via citations. *Scientometrics*, 126(1):871–906, 2021. doi: https://doi.org/10.1007/s11192-020-03690-4.

Anna Mierzecka, Małgorzata Kisilowska, and Andrius Šuminas. Researchers' expectations regarding the online presence of academic libraries. *College and research libraries.*, 78(7):934–951, 2017. doi: https://doi.org/10.5860/crl.78.7.934.

Birger Moëll and Fredrik Sand Aronsson. Harm reduction strategies for thoughtful use of large language models in the medical domain: perspectives for patients and clinicians. *Journal of Medical Internet Research*, 27:e75849, 2025. doi: https://doi.org/10.2196/75849.

Barend Mons, Cameron Neylon, Jan Velterop, Michel Dumontier, Luiz Olavo Bonino da Silva Santos, and Mark D Wilkinson. Cloudy, increasingly fair; revisiting the fair data guiding principles for the european open science cloud. *Information Services and Use*, 37(1):49–56, 2017. doi: https://doi.org/10.3233/ISU-170824.

James Jie Pan, Jianguo Wang, and Guoliang Li. Survey of vector database management systems. *The VLDB Journal*, 33(5):1591–1615, 2024. doi: https://doi.org/10.1007/s00778-024-00864-x.

Rob Penfold. Using the lens database for staff publications. *Journal of the Medical Library Association: JMLA*, 108(2):341, 2020. doi: https://doi.org/10.5195/jmla.2020.918.

Ciyuan Peng, Feng Xia, Mehdi Naseriparsa, and Francesco Osborne. Knowledge graphs: Opportunities and challenges. *Artificial intelligence review*, 56(11):13071–13102, 2023. doi: https://doi.org/10.1007/s10462-023-10465-9.

Philippe Rocca-Serra, Wei Gu, Vassilios Ioannidis, Tooba Abbassi-Daloii, Salvador Capella-Gutierrez, Ishwar Chandramouliswaran, Andrea Splendiani, Tony Burdett, Robert T Giessmann, David Henderson, et al. The fair cookbook-the essential resource for and by fair doers. *Scientific data*, 10(1):292, 2023. doi: https://doi.org/10.1038/s41597-023-02166-3.

Amanpreet Singh, Mike D'Arcy, Arman Cohan, Doug Downey, and Sergey Feldman. SciRepEval: A multi-format benchmark for scientific document representations. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 5548–5566, Singapore, December 2023. Association for Computational Linguistics. doi: https://doi.org/10.18653/v1/2023.emnlp-main.338.

Nagham A. Sultan and Dhuha B. Abdullah. Scraping google scholar data using cloud computing techniques. In *2022 8th International Conference on Contemporary Information Technology and Mathematics (ICCITM)*, pp. 14–19, 2022. doi: https://doi.org/10.1109/ICCITM56309.2022.10032044.

Isabelle Walsh and Frantz Rowe. Bibgt: combining bibliometrics and grounded theory to conduct a literature review. *European Journal of Information Systems*, 32(4):653–674, 2023. doi: https://doi.org/10.1080/0960085X.2022.2039563.

Wes McKinney. Data Structures for Statistical Computing in Python. In Stéfan van der Walt and Jarrod Millman (eds.), *Proceedings of the 9th Python in Science Conference*, pp. 56 – 61, 2010. doi: https://doi.org/10.25080/Majora-92bf1922-00a.

Mark D Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E Bourne, et al. The fair guiding principles for scientific data management and stewardship. *Scientific data*, 3(1):1–9, 2016. doi: https://doi.org/10.1038/sdata.2016.18.

Mark D Wilkinson, Michel Dumontier, Susanna-Assunta Sansone, Luiz Olavo Bonino da Silva Santos, Mario Prieto, Dominique Batista, Peter McQuilton, Tobias Kuhn, Philippe Rocca-Serra, Merce Crosas, et al. Evaluating fair maturity through a scalable, automated, community-governed framework. *Scientific data*, 6(1):174, 2019. doi: https://doi.org/10.1038/s41597-019-0184-5.

Kevin Wu, Eric Wu, Kevin Wei, Angela Zhang, Allison Casasola, Teresa Nguyen, Sith Riantawan, Patricia Shi, Daniel Ho, and James Zou. An automated framework for assessing how well llms cite relevant medical references. *Nature Communications*, 16(1):3615, 2025. doi: https://doi.org/10.1038/s41467-025-58551-6.

Haopeng Zhang, Philip S. Yu, and Jiawei Zhang. A systematic survey of text summarization: From statistical methods to large language models. *ACM Comput. Surv.*, 57(11), June 2025. ISSN 0360-0300. doi: https://doi.org/10.1145/3731445.

A Appendix

A.1 Output data

A.1.1 Generated Research Report

Research Report: Human Trafficking in Economics

Introduction

Human trafficking is a grave violation of human rights and a complex global issue with significant economic dimensions. This report examines the economics of human trafficking, identifies current research trends and gaps, and suggests potential directions for future research.

Current Research Trends

- **Economic Analysis of Human Trafficking:** A significant portion of research focuses on applying economic frameworks to understand the dynamics of human trafficking. This includes analyzing the costs and benefits for traffickers, the supply and demand factors that drive the trade, and the role of market structures in facilitating exploitation.
- **Macroeconomic Impacts:** Studies explore the macroeconomic consequences of human trafficking, such as its effects on labor markets, economic growth, and income inequality.
- **Policy Evaluation:** Research assesses the effectiveness of different anti-trafficking policies, including law enforcement strategies, victim support programs, and prevention efforts.
- **Regional Focus:** Research often focuses on specific regions or countries, examining the unique economic and social factors that contribute to human trafficking in those areas.
- **Quantitative Analysis:** Researchers use statistical methods and econometric models to estimate the scale of human trafficking, identify risk factors, and evaluate the impact of interventions.

Research Gaps

- **Data Scarcity:** A major challenge in studying the economics of human trafficking is the lack of reliable data. The clandestine nature of the crime makes it difficult to collect accurate statistics on the number of victims, the profits generated by traffickers, and the flows of money involved.
- **Complexity of the Issue:** Human trafficking is a multifaceted problem that intersects with various economic, social, and political factors. Research needs to consider these complexities and adopt interdisciplinary approaches to fully understand the issue.
- **Limited Focus on Prevention:** While much research focuses on law enforcement and victim support, there is a need for more studies on effective prevention strategies. This includes addressing the root causes of trafficking, such as poverty, inequality, and lack of education.
- **Impact of Technology:** With the rise of the internet and social media, technology plays an increasing role in human trafficking. More research is needed to understand how technology facilitates trafficking and how it can be used to combat it.

Potential Directions for Future Research

- **Develop better data collection methods:** Researchers need to develop innovative methods for collecting data on human trafficking, such as using surveys, interviews, and case studies. They should also explore the use of big data and machine learning techniques to identify patterns and trends in trafficking.
- **Conduct more rigorous evaluations of anti-trafficking interventions:** There is a need for more rigorous evaluations of anti-trafficking interventions, using experimental or quasi-experimental designs. This will help to identify which interventions are most effective and how they can be improved.
- **Examine the role of the private sector:** The private sector can play a significant role in preventing and combating human trafficking. Future research should explore how businesses can implement ethical supply chain management practices, raise awareness among employees and customers, and support anti-trafficking initiatives.
- **Explore the intersection of human trafficking and other forms of crime:** Human trafficking is often linked to other forms of crime, such as drug trafficking, money laundering, and corruption. Research should examine these links and develop integrated strategies for combating these crimes.
- **Focus on the long-term impact of human trafficking on victims:** Human trafficking can have long-lasting physical, psychological, and economic consequences for victims. Future research should examine these consequences and develop effective strategies for helping victims recover and reintegrate into society.

Conclusion

The economics of human trafficking is a complex and challenging field of study. By addressing the research gaps and pursuing the potential directions outlined in this report, researchers can contribute to a better understanding of this issue and inform the development of more effective anti-trafficking policies and programs.

Figure 5: Generated Research Report.