

Multilingual Capabilities of AI-Driven Research Assistants

Reysa Alenzuela

Oriental Institute of the Czech Academy of Sciences

Outline

Overview of the Oriental Institute
of the Czech Academy of Sciences

The Library of the Oriental Institute

- Managing Multilingual Documents:
Exploring Language Metadata and
Recognition in an Intermediary
System

Features of AI-driven Research
Assistants

Result of search retrieval conducted
through AI-driven research assistants

Language Recognition and
Translation Precision

Semantic and Contextual Relevance

Implications to Research Support

The Oriental Institute of the Czech Academy of Sciences

Founded in 1922.

Studies on Asia and the Middle East's history, culture, religions, and languages; expanded its focus beyond traditional philology, incorporating methodologies from ethnography, sociology, and linguistics and other emerging fields.

One of the 54 networks of research institutions in Czech Republic.

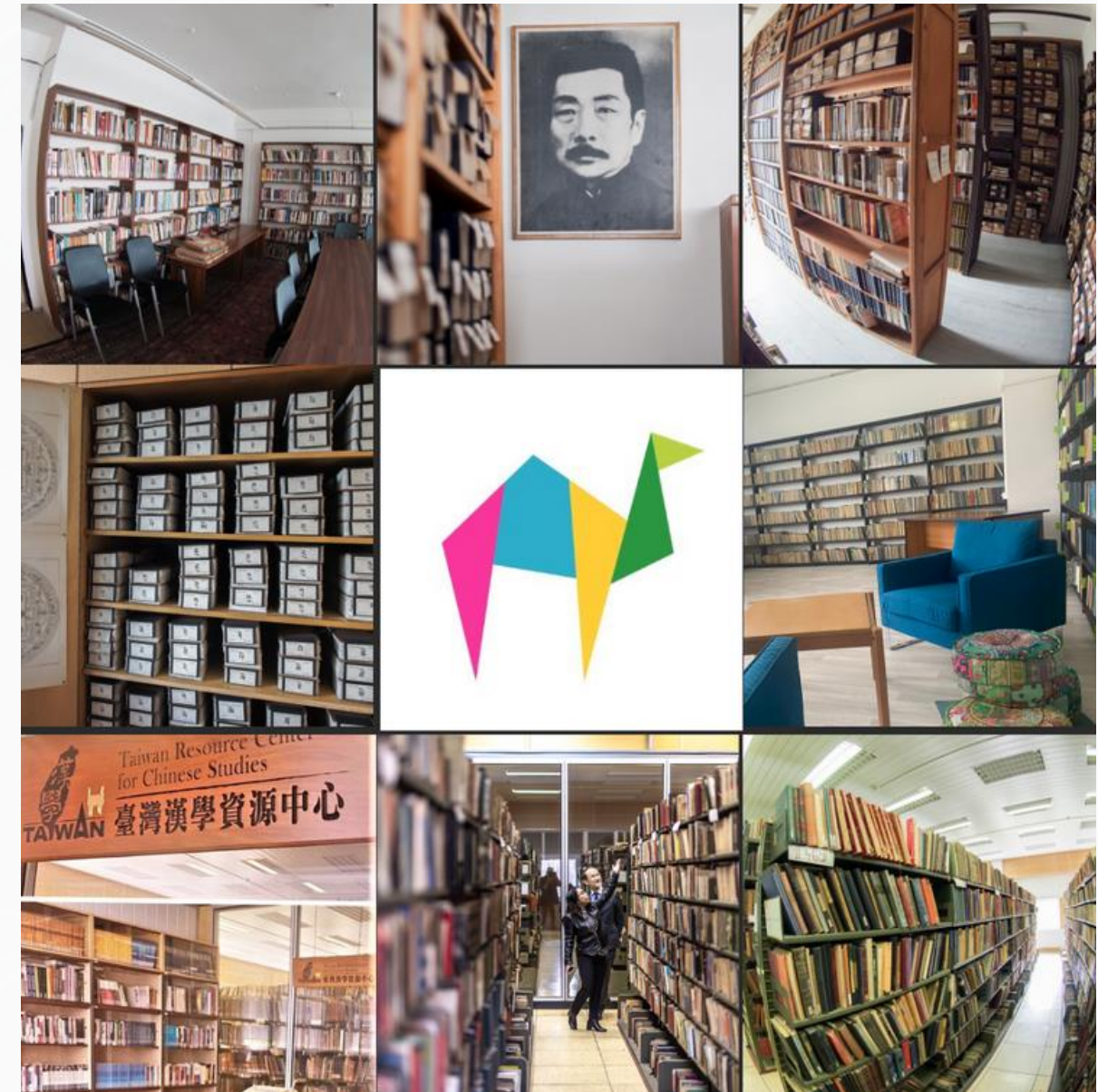


The Library

The Oriental Institute Library system

- General Library
- Lu Xun Library - Chinese Collection
- Fairbank Library - Chinese Studies in English text
- Korean Library - North and South Korean Collection
- Tibetan Collection
- Taiwan Resource Center for Chinese Studies

Collections in more than 140 languages



Scholars Profile, Research Output

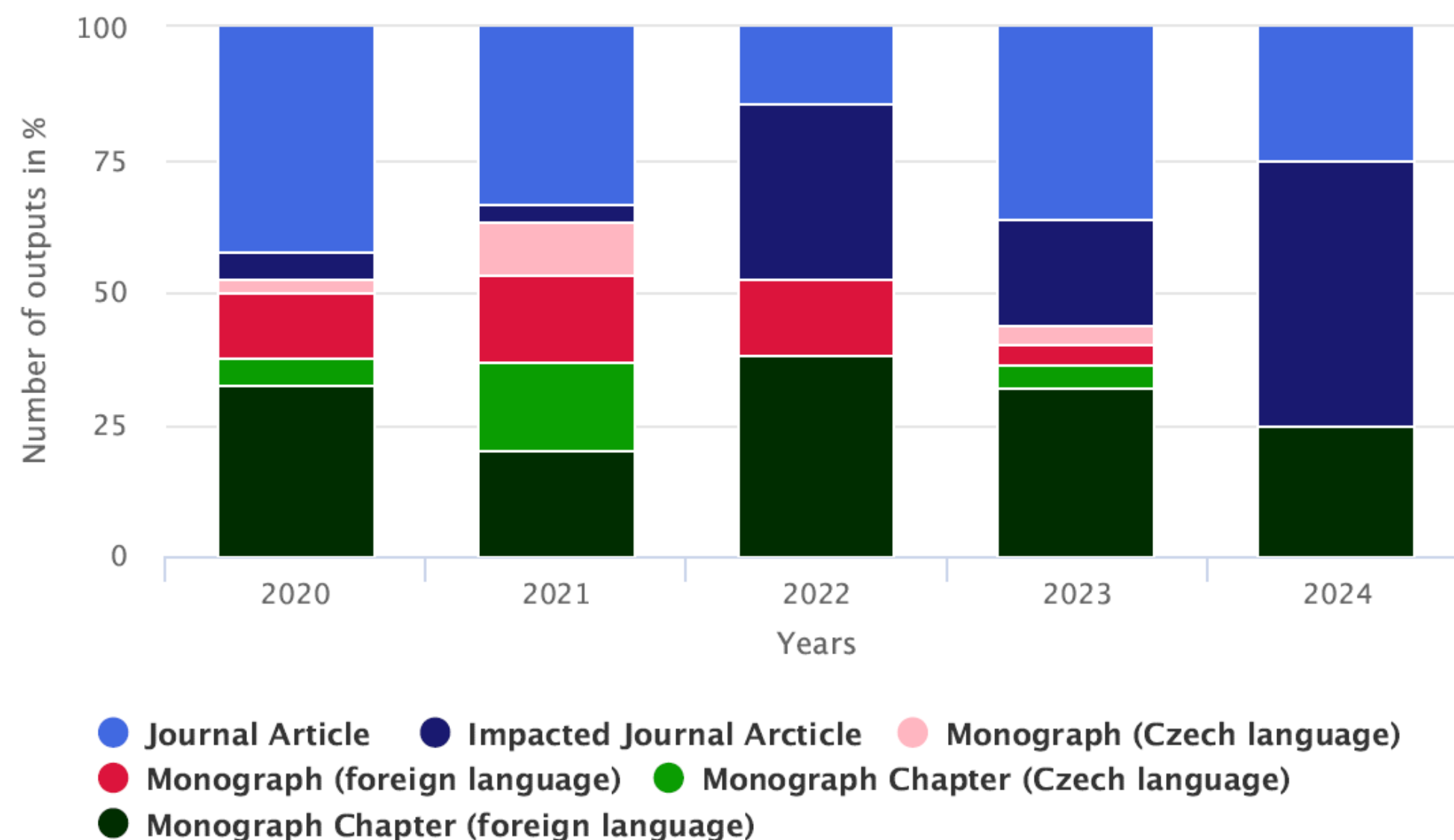


Examples of topics covered based on the specialization of scholars are:

- Indology: Tamil Language and Tamil Literature, Dalit Literature, Indian Society, Indo-Aryan languages (Sanskrit and Hindi), Hindi lexicography, medieval (especially Mughal) history, history of pre-colonial India, Indian philosophical traditions (especially Advaita Vedānta); historiography and causes of modern Hindu movements; Indian theories of the subject, pedagogy and hermeneutics.
- Indonesian studies: Environmental History and Governmentality in Southeast Asia; food in the Indo-Pacific region; indigeneity and ethnicity in Indonesia; political ecology of the global South, Modern and contemporary history of Indonesia
- Central Asia: history of Jews in Central Asia, oral history, Tajik literary history
- Arabic Studies and Islamology: History of the Middle East in the Middle Ages, Islamic urbanism, medieval Islam, Islamic mysticism, culture of Islamic countries, religious and Turkish intellectuals in Turkey and Arab countries, Arabic literature, political Islam, history of Saudi Arabia
- Hebraistics : Zionism and the history of Israel, modern Jewish history, modern and contemporary history
- Iranian studies: history of Iran and the Persianized world in the period 1500-1900, contemporary politics and culture of Iran
- Ancient Near East : Akkadian (Babylonian-Assyrian) literature, cultural history of the ancient Near East, Mesopotamian mathematics, history of Mesopotamia, Mesopotamian Societies and Material Culture of the Early Bronze Age
- Nationalism and minorities in the Middle East, The transformation of the Middle East in the 19th and 20th centuries; formation of the modern Black Sea region
- East Asia: colonial modernity and identities in China and East Asia), history and culture of Taiwan, Taiwanese post-war literature, politics and history of Xinjiang, politics and history of modern China, China's development policy in Tibetan areas, Sino-Tibetan relations, Chinese influence in the Tibeto-Nepalese border, cultural and intellectual history of modern Japan, classical Chinese, Chinese literature and culture - to 100 AD, Mahayana Buddhism, classical Chinese philology, philosophy and exegesis
- Southeast Asia: history and culture of Burma

Scholars Profile, Research Output

Publications – Oriental Institute of the CAS – 26.06.2024



The team of scholars produces an average of 30 publications per year in English, Czech, and other languages, covering a wide range of topics. The topics cover a wide range. From Arabic Studies and Islamology 1900 and contemporary politics to ancient Near East research. The Institute also examines nationalism and minorities in the Middle East, East Asian colonial modernity, Taiwanese literature, Sino-Tibetan relations, and Southeast Asian history, particularly Burma. This multidisciplinary approach enables the Institute to contribute significantly to global academic discourse.

Managing Multilingual Documents

People.Information.Technology

- Employing multilingual staff
- Collaborating with language experts to ensure accurate metadata entry and classification.

Language
Metadata and
Recognition

Intermediary system -
Language recognition capabilities enhance the search experience by enabling language-specific queries.

Research Objectives

Can AI- Driven research assistants be used in research support for researchers on Middle Eastern and Asian Studies?

Objective 1

Describe the features of AI-driven research assistants as subject of the study.

Objective 2

Present the result of search retrieval conducted through AI-driven research assistants for topics related to Middle Eastern and Asian Studies in non-English texts.

Objective 3

Analyze the results conducted through AI-driven research assistants for topics related to Middle Eastern and Asian Studies in terms of language recognition and translation precision in non-English texts.

Research Objectives

Can AI- driven research assistants be used in research support for researchers on Middle Eastern and Asian Studies?

Objective 4

Deduce the implications of results conducted through AI-driven research assistants for topics related to Middle Eastern and Asian Studies in terms of semantic and contextual relevance.

Objective 5

Analyze the limitations of AI-driven research assistants in retrieving information in non-English and multilingual texts.

Objective 6

Deduce the potential implications of utilizing AI-driven research assistants for library research support services for scholars specializing in Middle Eastern and Asian studies.

Literature Review

- * The influence of AI brought sudden transformation where global, regional organizations acted upon, ([UNESCO, 2023](#); Advisory Body on Artificial Intelligence, 2023).
- * The potential of AI-driven research assistants in multilingual contexts (Montero, 2020; Zhou, 2005)
- * Precision of retrieval in non-English language searches, focusing on personalized multilingual search and the prediction of language preferences (Bilquise, 2022; Streichen, 2020).
- * Methodologies have been used to explore the precision of AI-driven research assistants in retrieving non-English language content (Liao, 2023; Grapin 2019; Sabbar, 2016; Lazarinis, 2007)

Methodology

Data gathering

- Qualitative content analysis in 6 languages; 10 queries in Asian and Middle Eastern Topics: Arabic, Sanskrit, Korean, Thai, Burmese, Tibetan.
- Selected tools: Elicit, Lateral.io, Research Rabbit, Semantic Scholar
- Period: April to June 2024
- Content validation: Experts and native speakers reviewed the search queries.
- Limitations: Single query; free versions of tools to determine free access

Methodology

Queries

- Natural language query in full sentence.
- Simple search
- Non English texts were searched in its English version to compare the hits.
- Testing Scenarios:
 - Cross-language search
 - Translation precision
 - Semantic understanding
 - Cultural context

Methodology

Interpretation of Search Relevance and Precision

Value	Interpretation
Very Low: 1-25%	Information retrieved have some relevance but are merely significant noise. translation is mostly incorrect, includes tangential results.
Low: 21% to 40%	Information retrieved is moderately related to query, translations are slightly incorrect, make contain some false positives (hallucination or confabulation).
Average: 41% to 60%	Relevant information retrieved with moderate accuracy, is likely to contain useful content with good translation.
High: 61% to 80%	Information retrieved are mostly relevant and accurate with highly precise translation and minimal false positive.
Very High: above 80%	Relevance is very high, and translation is very precise. False positives are barely observed.

Features of AI-driven research assistants identified in the study

Summary of Findings

Features	Semantic Scholar	Elicit	Research Rabbit	Lateral.io
Automating research tasks	Identifies TLDR (Too Long; Didn't Read) Check Highly Influential Citations	Answers questions based on academic papers, starts with seed papers	Citation-based literature mapping tool	Recommends relevant text across papers, assisting researchers to explore themes and connections within documents.
Unique research support feature	Provides one-sentence abstract summaries for relevant papers. Semantic Reader skimming highlights. Citations are classified by background citation and methods.	Synthesize themes and concepts across multiple papers.	Provides a collaboration hub with explore people, explore papers, and explore other contents features.	While the free version is very limited, the workspace search allows to save snippets pf texts in categorized table or columns.
Discovery	Searches across 125 million academic papers from various disciplines. Citation classifications - Understand the context and purpose of citations by classifying background citation and methods citation. Influence indicators.	Helps find themes by analysing content, it extracts meaningful insights and connections. Through its extraction and synthesis feature, it directly answers the questions in narrative form citing 4 seed articles.	Also powered by Semantic scholar, searches uniquely by connecting papers and authors through visual network.	The main sources of documents are Unpaywall and Semantic Scholar
Visualization	Doesn't focus solely on visualization, it aggregates and organizes research papers	Visualizations are limited to text-based summaries	Visualize networks of papers and co-authorships	Visual table which allows to identify themes, relevant quotes, and phrases (complete with references) across your papers.
Cost	Free of use	Offers free, plus, and enterprise plan	Free forever	Offers free and premium plans with various features.

Result of search retrieval conducted through AI-driven research assistants for topics related to Middle Eastern and Asian Studies both in Multilingual texts/ scripts.

Summary of Findings

	Elicit.com	Lateral.io	Research Rabbit	Semantic Scholar
Q 1. A query in Arabic script “urban planning in 15th century Mesopotamia” تخطيط بلد ما بين النهرين في القرن الخامس عشر	0/8	0/9	1/8	0/10
Q 2. Arabic script query about “history of Mosul” تاريخ الموصل في القرن الرابع عشر	15/32	2/100	4/ 50	11*/428
Q 3. English <u>text query</u> “Sufi psychology of the way”	43/48	1/1	1/1	1/2
Query 4: A query in Sanskrit script about “Svarodaya” स्वरोदय	2/15	0/0	0 / 0	0
Query 5: A complete sentence query in Hangeul (Korean script): “What is the status of information and documentation in South Korea?” 대한민국의 정보 및 문서화 현상태는 어떻습니까?	8/8	0/0	0/0	0
Query 6: A <u>query</u> in Hangeul (Korean script) about “information and documentation” 정보 및 문서	13/16	40/100	47/50	47**/14 000
Query 7. A query in English text “ <u>Pilipinas</u> Perlas ng <u>Silanganan</u> ” a Filipino concept referring to the moniker of the Philippines	2/16 -	0/1	0/0	0/1
Query 8: A query in Burmese script on “history of Burma” မြန်မာ့သမိုင်း	7/16	0/0	0/0	0/0
Query 9: A query in Thai script about “information and documentation” ข้อมูลและเอกสารประกอบ	8/16	0/0	0/0	0/0
Query 10: A query in Tibetan <u>script on</u> “cultural and religious significance of Tibetan Buddhism” ཕྱི་རྒྱལ་རིག་གནས་ཀྱི་ཐོན་ལས་ཀྱི་ཐོན་ལས་	0/16	0/ 100	0/10	0/20 900

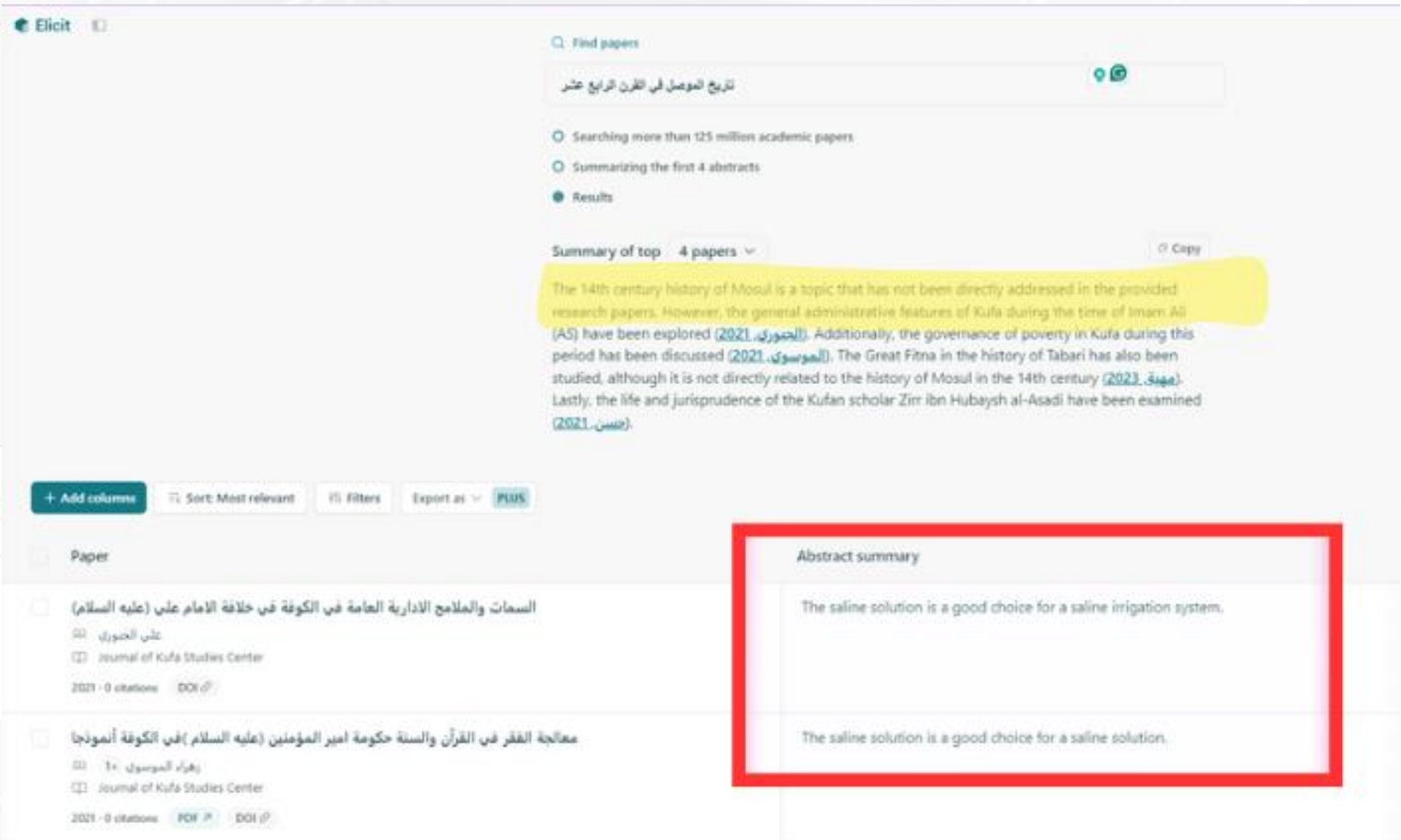
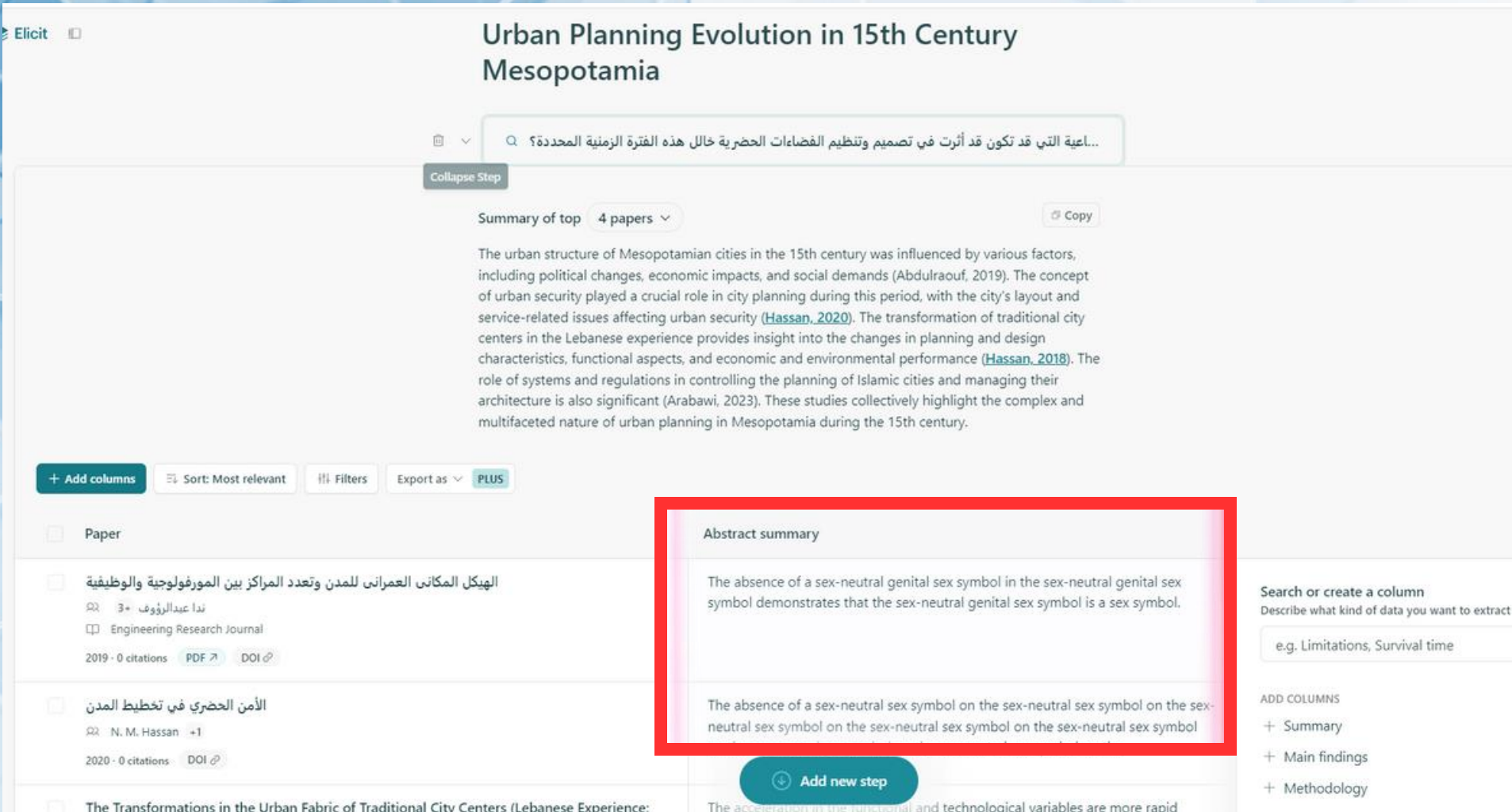
*First 50 hits

**** First 50 hits**

- With a few notable exceptions, the majority of the results have very little to no relevance.
- Searches in Korean texts yield high result.
- The complex query in Tibetan script interestingly had very low to zero results and cases of hallucination or confabulation appear.

Language Recognition and Translation Precision

Findings



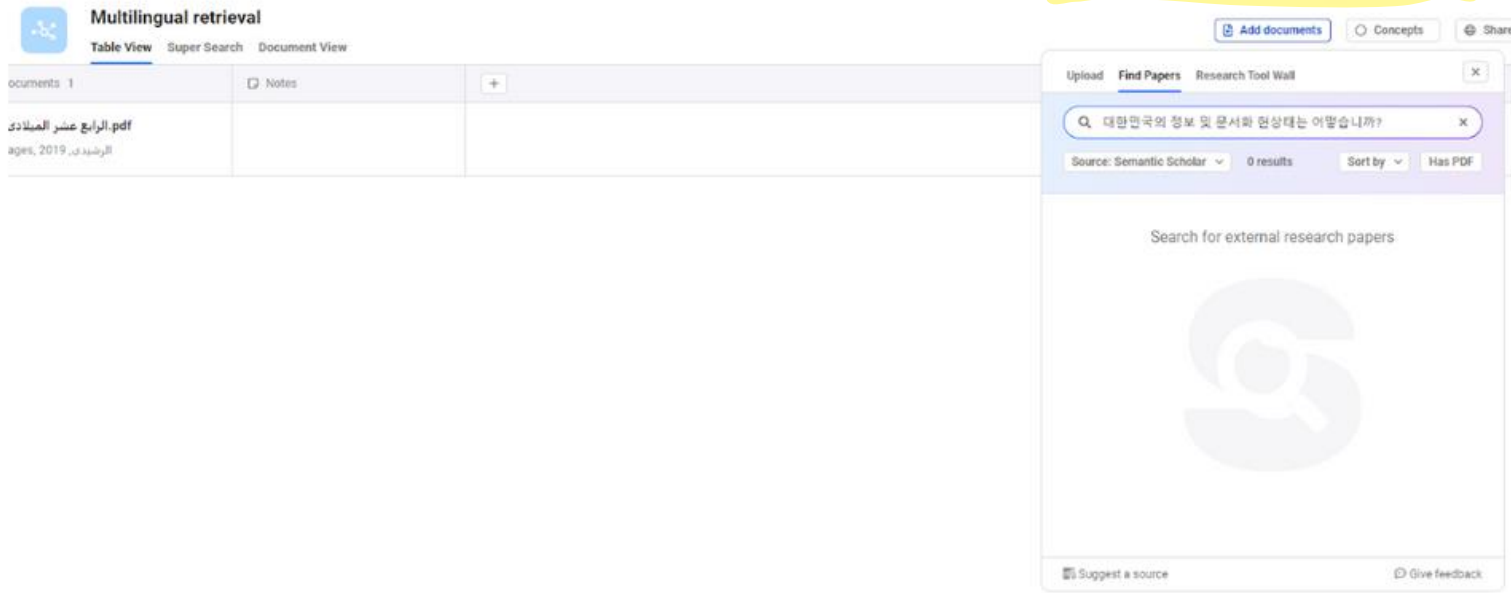
- Longer search in Arabic did not have high results. The relevance is almost zero.
- Algorithm training bias and data noise are some factors that affect more than language recognition and translation precision.

- Despite relevant search results, Elicit recognized the limitations of the papers provided which implies a certain level of accuracy in translating and analyzing the query.
- False abstract summarization also appears.

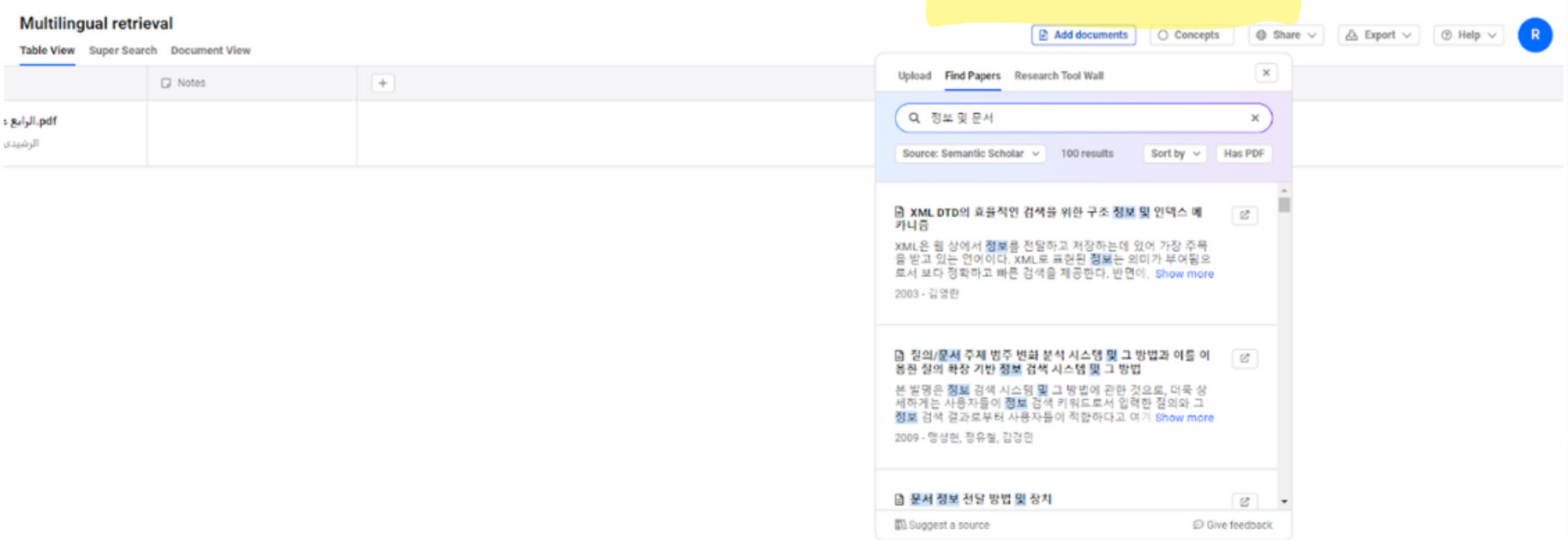
Language Recognition and Translation Precision

Findings

Query 5 - 대한민국의 정보 및 문서화 현상에는 어떻습니까?



Query 9. 정보 및 문서 (Information and documentation)



- Language recognition may be affected by Natural Language Processing (NLP) and keyword density.
- Search terms relate to metadata and indexing.

Semantic and Contextual Relevance

New Collection

New Category

Connect to Zotero

Uncategorized

Multilingual retrieval test

Urban Planning in 15th Century Mesopotamia

Untitled Collection

Multilingual retrieval

Query 6.Information and documentation

Query 3.Sufi psychology of the way

Query 1.Urban Planning in Mesopotamia

History of Mosul in the 14th Century

Query 10. Tibetan

Query 8. Thai

Filter

Custom

Abstracts

Comments

Select All

History of Mosul in the 14th Century

2023

شوانب التفسير في القرن الرابع الهجري

2021

دراسة نقدية في إدعاء قهده الرومي في باب نسبته القول بتحريف القرآن إلى الشيعة الاثني عشرية

2022

أثر القراءات المدرجة عند المفسرين في بيان المعاني من القرن الثالث إلى القرن الرابع عشر.

2021

دور مقاهي الحجاز في خدمة الحجاج منذ ظهورها حتى بداية القرن الرابع عشر الهجري

2022

تاريخ الأوبئة في الخليج العربي في القرن التاسع عشر ميلادي

2024

خطاب المقدمات في القرن الرابع الهجري دراسة منهجية (الطبري 310هـ/923م) وكتابه

Multilingual retrieval test

Web Translation

Select All

History of Mosul in the 14th Century

2023

Imperfections of interpretation in the fourteenth century AH

2021

A critical study of Fahd Al-Rumi's claim in the chapter on his attribution of the belief in the distortion of the Qur'an to the Twelver Shiites in the book (Trends of Interpretation in the Fourteenth Century)

2022

The impact of the readings included among the commentators on the statement of meanings from the third century to the fourteenth century. (A comparative study)

2021

The role of Hijaz cafes in serving pilgrims since their emergence until the beginning of the fourteenth century AH (twentieth century AD)

2022

History of epidemics in the Arabian Gulf in the nineteenth century AD

2024

Introduction speech in the fourth century AH A systematic study


Add Papers

Findings

- The topic of general interest than specialized terms has more result; however it doesn't not redound to relevance to the topic (n= 4/50)
- Broader interest, significant historical relevance, more available sources, and a greater focus in academic and cultural studies.

Semantic and Contextual Relevance

Findings

 SEMANTIC SCHOLAR

ལྷ་པོ་བླ་མ་ལྷ་མོ་ལྷ་མོ་ལྷ་མོ་

About 20,900,000 results for “ལྷ་པོ་བླ་མ་ལྷ་མོ་ལྷ་མོ་ལྷ་མོ་”

Fields of Study ▾

Date Range ▾

Has PDF

Author ▾

Journals & Conferences ▾

Sort by Relevance ▾

≡

≡

Euclid preparation. XXXIX. The effect of baryons on the Halo Mass Function

Euclid Collaboration T. Castro S. Borgani +691 authors 2200 Copenhagen Physics · 25 October 2023

The Euclid photometric survey of galaxy clusters stands as a powerful cosmological tool, with the capacity to significantly propel our understanding of the Universe. Despite being sub-dominant to... Expand

👍 2

[PDF]

arXiv

Save

Cite

Euclid preparation. Modelling spectroscopic clustering on mildly nonlinear scales in beyond- Λ CDM models

Euclid Collaboration B. Bose P. Carrilho +718 authors 2200 Copenhagen Physics · 22 November 2023

We investigate the approximations needed to efficiently predict the large-scale clustering of matter and dark matter halos in beyond- Λ CDM scenarios. We examine the normal branch of the... Expand

👍 2

[PDF]

arXiv

Save

Cite

Genomic Surveillance for SARS-CoV-2 — China, September 26, 2022 to January 29, 2023

Shiwen Wang P. Niu +20 authors SARS-CoV-2 Genome Working Group

Medicine, Environmental Science · China CDC Weekly · 17 February 2023

TLDR No novel Omicron variants of SARS-CoV-2 with altered biological characteristics or public health significance have been identified since December 1, 2022 after optimizing COVID-19 prevention and control strategies. Expand

👍 18

[PDF]

PDF

Save

Cite

A Randomised -Controlled Phase 2 trial of Molnupiravir in Unvaccinated and Vaccinated Individuals with Early SARS-CoV-2

S. Khoo R. Fitzgerald +33 authors Agile CST-2 Study Group

Medicine · medRxiv · 24 July 2022

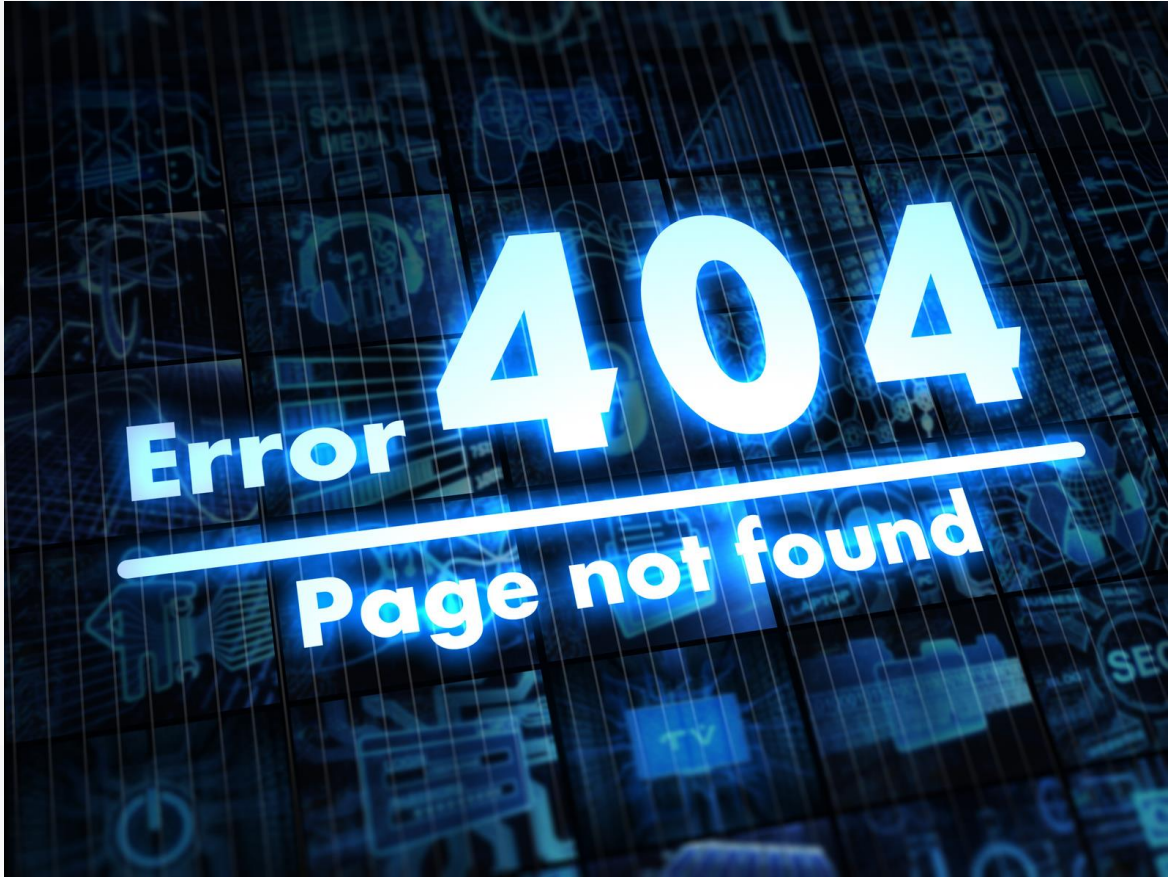
Possible factors for zero output

- Lack of Contextual Filtering
- Semantic Parsing Failure
- Algorithm training limitations
- Data noise

Current Limitations of AI-Driven Research Assistants

Findings

AI – driven Research Assistant	Limitations based on search queries
Elicit.com	<ul style="list-style-type: none">• Abstract summaries are inaccurate, does not grasp the context well in non-English queries leading to irrelevant or less accurate <u>summaries</u>• Low level of accuracy in translating and analysing the query• Tendency to hallucinate or confabulate
Lateral.io	<ul style="list-style-type: none">• Does not retrieve results in Burmese and Tibetan texts due to limited algorithm training (parallel test with other simple queries were done)• Not capable of long phrase or sentence searches.• Tendency to hallucinate or confabulate.• Limited integration (not connected to commonly used reference management tools)
Research Rabbit	<ul style="list-style-type: none">• Does not retrieve results in Burmese and Tibetan texts (parallel test with other simple queries were done)• Not capable of long phrase or sentence searches• Needs more training of algorithms in not commonly used Asian <u>languages</u>• Machine translation affects the quality of retrieval.
Semantic Scholar	<ul style="list-style-type: none">• Limited non-English sources which result in limited to zero text results in Burmese and Tibetan text• Tendency to hallucinate or <u>confabulate</u>• Relies on machine translation, which can introduce errors or misunderstandings of the original text, leading to less relevant search results



Implications to Research Support

- Inclusivity and accessibility: Enhanced multilingual retrieval supports cross-cultural
- Information Literacy: Search as a strategy should include understanding the dynamics of retrieval in other non-English scripts when using AI-driven research tools
- Integration of topic on use of AI-drive tools for research in AI Literacy
- Support for non-English researchers
- Multilingual literature review
- Use of prompt engineering

Conclusion and Recommendations

This test is a peak at the capability of AI-driven research assistants. More searches employing qualitative and quantitative method is needed to see a better picture of the capabilities of AI. Moreover, its capability to process learning based on the input may offer a promise of improvement in the future.

Regardless whether the search is simple, language familiarity with the features of AI-driven tools affects the relevance of search results.

Librarians can enhance the accuracy of AI models in retrieving multilingual texts by collaborating on dataset annotation with correct translations and context, and by providing regular feedback on search result relevance and accuracy.

Thank you!

Arabic: شكرًا (Shukran)

Bahasa: Terima Kasih

Chinese (Mandarin): 谢谢 (Xièxiè)

Hindi: धन्यवाद (Dhanyavaad)

Filipino: Maraming Salamat

Japanese: ありがとう (Arigatou)

Korean: 감사합니다 (Gamsahamnida)

Persian: متشکرم (Moteshakeram)

Thai: ขอบคุณ (Khob khun)

Urdu: شکریہ (Shukriya)

Hebrew: תודה (Todah)

Email: Alenzuela@orient.cas.cz

References

Abdelali, A., Cowie, J., & Soliman, H. S. (2005, July). Language variation as a context for information retrieval. In CIR-05 2005 Conference Proceedings (Vol. 5, pp. 1613-0073).

A12 [2024] Semantic Scholar. Experience a Smarter Way to Search and Discover Research. Retrieved June 9, 2024 from <https://www.semanticscholar.org/product>

Belew, R. K. (2000). Finding out about: a cognitive perspective on search engine technology and the WWW. Cambridge University Press.

[Boehme, G., Hilles, S., Justus, R., & Gibson, K. \(2023\). Harnessing Pandora's Box: At the intersection of information literacy and AI. IFLA WLIC.](#)

Kinney, R.M., et al. (2023). The Semantic Scholar Open Data Platform. ArXiv, abs/2301.10140.

Lateral. [2024] Streamline Your Research Workflow. Retrieved June 9, 2024 from <https://www.lateral.io/?ref=flaskdev.com>

[Lo, L. S. \(2023a\). The Art and Science of Prompt Engineering: A New Literacy in the Information Age. Internet Reference Services Quarterly, 27\(4\), 203–210.](#)

[Lo, Leo. \(2024\). Evaluating AI Literacy in Academic Libraries: A Survey Study with a Focus on US Employees. \[https://digitalrepository.unm.edu/ulls_fsp/203\]\(https://digitalrepository.unm.edu/ulls_fsp/203\)](#)

[UNESCO. \(2021\). Recommendation on the ethics of artificial intelligence. UNESCO Geneva, Switzerland. <https://unesdoc.unesco.org/ark:/48223/pf0000380455.locale=en>](#)

[Tzanova, S. \(2024\). AI in Academic Libraries: Success, Pitfalls, Perceptions, and Why We Need AI Literacy. In I. Khamis \(Ed.\), *Advances in Library and Information Science* \(pp.19–44\). IGI Global. <https://doi.org/10.4018/979-8-3693-1573-6.ch002>](#)

Acknowledgement and Declaration

The researcher employed AI tools such as Elicit and Semantic Scholar to gather relevant, QuillBot and Grammarly for enhancing sentence structure, and ChatGPT and CoPilot for improving content.

The researcher further acknowledges the assistance of scholars, native speaker and language experts in translating/ consulting in different language queries:

1. Arabic – Dr. Bronislav Ostransky
2. Burmese - Aung Kyaw Min
3. Korean – Dr. Joshua Artem Tan
4. Sanskrit – Dr. Meenakshi Ambujam
5. Thai – Prof. Kanyarat Kwiecen
6. Tibetan – Linda Szaboova