An AI Pipeline for Scientific Literacy and Discovery: a Demonstration of Perspicacité-AI Integration with Knowledge Graphs

Lucas Pradi[1], Tao Jiang[1,2], Matthieu Feraud[1,2], Madina Bekbergenova[1,2], Yousouf Taghzouti[2,3] and Louis-Felix Nothias[1,2]

[1]*Univ. Côte d'Azur, CNRS, ICN, France*
[2]*Interdisciplinary Institute for Artificial Intelligence (3iA) Côte d'Azur, France*
[3]*Univ. Côte d'Azur, Inria, I3S, France*

Abstract

Keeping up with the rapid pace of publishing is becoming an increasingly challenging task. Moreover, the interdisciplinary nature of research poses significant challenges for both students and academics. In this demo, we present the Perspicacité-AI Expanded Pipeline: an agentic workflow powered by LLMs that leverages bibliographic knowledge graphs, as well as local and web-based scientific literature searches. Our approach lowers the entry barrier for new researchers while accelerating literature discovery and reference curation. We demonstrate the system in a live prototype demo, showing how it generates domain-specific bibliographies and delivers well-sourced, curated responses drawn from the most relevant literature in the user's field of inquiry in real time.

Keywords

Large Language Models, Scientific Literature, Retrieval-Augmented Generation

1. Introduction

Science is evolving quickly, with the number of scientific publications increasing exponentially, while becoming ever more interdisciplinary, creating knowledge integration barriers and challenges to both students and researchers [1, 2]. Large language models (LLMs) exhibit remarkable capabilities in synthesizing complex information, identifying connections across disciplines, and generating comprehensive summaries, capabilities that have been already applied in a diverse set of areas, including for scientific research and education [3, 4], although with some drawbacks, such as outdateness, hallucination and inability of sourcing [4, 5, 6, 7, 8, 9]. Retrieval Augmented Generation (RAG) can mitigate some of those limitations of standalone LLMs by grounding responses in relevant texts.

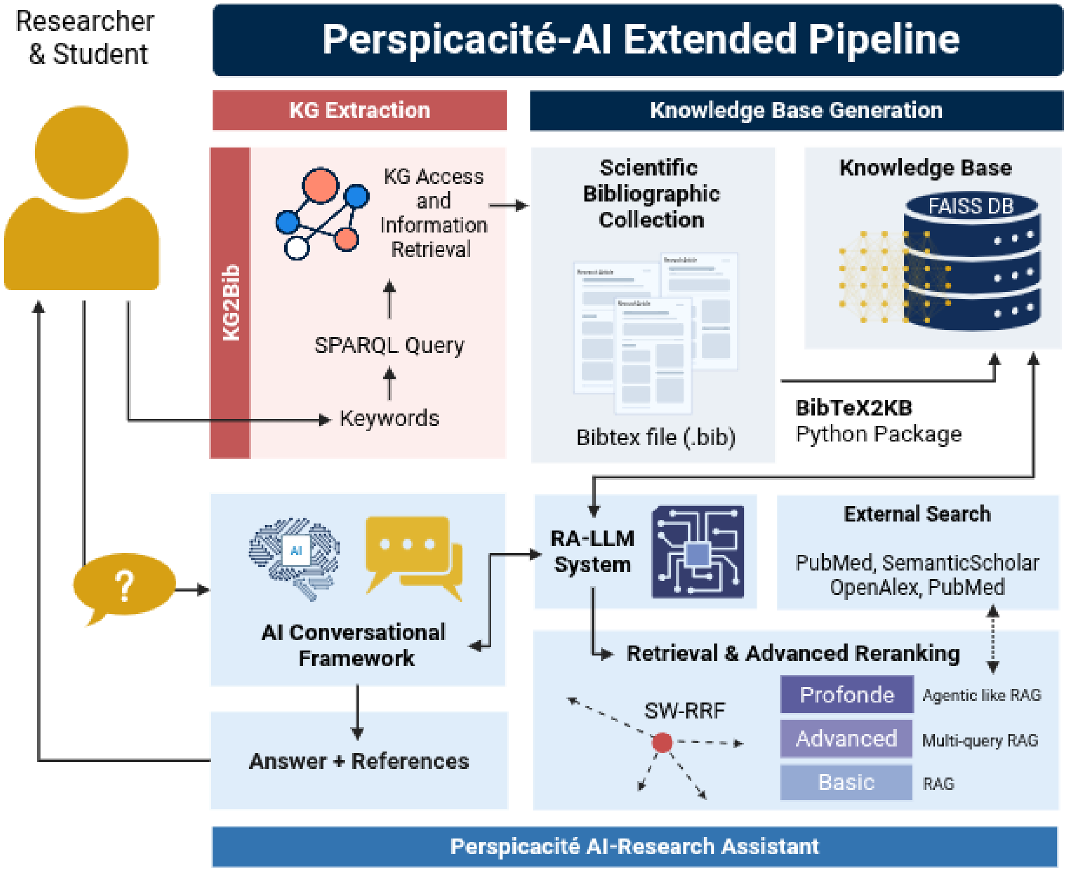
Perspicacité-AI is a flexible, open-source and free-to-use agentic pipeline that leverages RAG by providing support for multiple LLMs and advanced retrieval modes while empowering the use of both local and web-searched scientific literature. One of the main advantages of this pipeline is the ability to programmatically generate FAISS Vectorstores from BibTeX (.bib) files to be used in RAG frameworks. Our system adopts a hybrid retrieval strategy that combines a dense vector retriever with a sparse retriever (BM25), which improves the retrieval of domain-specific terms.

Even though the usage of user-provided .bib files can be an advantage, it poses significant challenges for newcomers in scientific areas. Knowledge graphs (KGs), such as the DBLP KG [10], provide a structured, semantic bibliography that integrates bibliographic data, e.g. authors, papers, and venues,

ISWC 2025 Companion Volume, November 2–6, 2025, Nara, Japan

Lucas.PRADI@univ-cotedazur.fr (L. Pradi); tao.jiang@cnrs.fr (T. Jiang); matthieuferaud31@gmail.com (M. Feraud); Madina.BEKBERGENOVA@univ-cotedazur.fr (M. Bekbergenova); yousouf.taghzouti@univ-cotedazur.fr (Y. Taghzouti); louis-felix.nothias@cnrs.fr (L. Nothias)
https://orcid.org/0009-0003-8641-6268 (L. Pradi); https://orcid.org/0000-0002-5293-3916 (T. Jiang); https://orcid.org/0009-0001-5496-6690 (M. Feraud); https://orcid.org/0009-0007-6356-9754 (M. Bekbergenova); https://youctagh.github.io/ (Y. Taghzouti); https://lfnothias.github.io/ (L. Nothias)
0009-0003-8641-6268 (L. Pradi); 0000-0002-5293-3916 (T. Jiang); 0009-0001-5496-6690 (M. Feraud); 0009-0007-6356-9754 (M. Bekbergenova); 0000-0003-4509-9537 (Y. Taghzouti); 0000-0001-6711-6719 (L. Nothias)

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073



Researcher & Student
Perspicacité-AI Extended Pipeline
KG Extraction
Knowledge Base Generation
KG2Bib
KG Access and Information Retrieval
SPARQL Query
Keywords
Scientific Bibliographic Collection
Bibtex file (.bib)
Knowledge Base
FAISS DB
BibTeX2KB
Python Package
?
AI
AI Conversational Framework
RA-LLM System
External Search
PubMed, SemanticScholar OpenAlex, PubMed
Retrieval & Advanced Reranking
SW-RRF
Profonde
Agentic like RAG
Advanced
Multi-query RAG
Basic
RAG
Answer + References
Perspicacité AI-Research Assistant

Figure 1: Perspicacité-AI Expanded Pipeline

in a machine-readable format. These KGs have enriched semantic relationships and can easily be integrated with other KGs, e.g. Wikidata. In this demo, we showcase the usage of Knowledge Graphs (KGs) for knowledge gathering and .bib file generation to be further used in Perspicacité-AI.

2. System Overview: RAG Automation, From KG to Online Deployment

As shown in Figure 1, our pipeline comprises multiple layers, starting with keywords and ending with a deployable scientific QA assistant. This is achieved through a sequence of independent packages that can be composed into an end-to-end framework. Below, we describe each part of component separately.

2.1. KG2Bib - KG Harvester:

Users specify keywords for their research area of interest and a limit of papers to extract using a CLI script. These inputs are transformed into a focused SPARQL query executed against a public bibliographic KG endpoint (in our case, DBLP). Listing 1 illustrates a sample query. The query identifies publications whose titles contain at least one of the specified keywords through case-insensitive matching. It assigns a relevance score based on the number of matching keywords (one point per keyword) and counts citations for each publication. Results are returned according to the user-defined limit (or 10 by default), ordered first by relevance score and then by citation count, ensuring that the most relevant and highly-cited works are prioritized. The results are subsequently transformed into a single .bib file for use in subsequent steps.

Listing 1: A SPARQL query to retrieve top-cited papers containing keywords: LLM, scientific literacy, or knowledge graphs in their titles

```
PREFIX dblp: <https://dblp.org/rdf/schema#>
PREFIX cito: <http://purl.org/spar/cito/>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>

SELECT ?publ ?label (COUNT(?citation) AS ?cites) ?score WHERE {
 ?publ rdf:type dblp:Publication .
 ?publ dblp:title ?title .
 BIND(LCASE(STR(?title)) AS ?lowerTitle)

 # Compute how many of the target words are in the title
 BIND(
  IF(CONTAINS(?lowerTitle, ’llm’), 1, 0) +
  IF(CONTAINS(?lowerTitle, ’scientific literacy’), 1, 0) +
  IF(CONTAINS(?lowerTitle, ’knowledge graphs’), 1, 0)
  AS ?score
 )

 # Only keep titles that contain at least one of the target words
 FILTER(?score > 0)

 ?publ dblp:omid ?omid .
 ?publ rdfs:label ?label .
 ?citation rdf:type cito:Citation .
 ?citation cito:hasCitedEntity ?omid .
}
GROUP BY ?publ ?label ?score

ORDER BY DESC(?score) DESC(?cites)
LIMIT 55
```

2.2. BibTeX2KB - FAISS Vectorstore Generation Pipeline:

This package parses each BibTeX entry and retrieves full-text content through a hierarchical protocol: (1) local files linked in reference managers, (2) Unpaywall-verified open-access content, (3) publisher APIs, and (4) abstract fallback for paywalled articles. It also supports diverse content types: GitHub README files and code scripts, YouTube video transcripts via API integration, and direct incorporation of proprietary PDFs without third-party transmission. This multi-modal approach encompasses both formal and informal sources of scientific information, supporting comprehensive learning and research.

The pipeline normalizes, chunks, and embeds all texts, storing them in a FAISS Vectorstore alongside rich metadata (DOI, source type, chunk-id, etc). BibTeX2KB supports merging multiple knowledge bases, enabling researchers in interdisciplinary fields to perform multiple keyword searches with KG2Bib and combine the results into a unified knowledge base.

2.3. Retrieval Core:

A router chooses among three modes:

- **Basic RAG:** A standard Retrieval-Augmented Generation pipeline that performs a single-pass similarity search over a FAISS-based vector store. When a user submits a question, the system embeds the question and retrieves the most semantically similar documents directly from the knowledge base. These documents are then passed to the language model to generate a response.
- **Advanced RAG:** An enhanced RAG system where an LLM generates multiple diverse subqueries by rephrasing the original question. Each subquery is independently used to retrieve relevant documents from the vector store. The retrieved results are then merged and ranked using our novel Sigmoid-Weighted Reciprocal Rank Fusion (SW-RRF) function, which combines the rankings across subqueries. SW-RRF modifies the traditional Reciprocal Rank Fusion (RRF) by introducing a sigmoid-based weighting mechanism that dynamically adjusts document weights based on their similarity scores, helping to reduce the impact of marginally related content in multi-query

retrieval scenarios. The top-N documents from this fused ranking are then passed to the LLM to generate a more comprehensive and contextually grounded answer.

- **Perspicacité-Profonde:** A cutting-edge agent-based system that employs a plan–search–reason loop. The agent autonomously formulates a multi-step strategy, identifies intermediate information needs, and issues targeted searches on specialized scientific platforms (e.g., PubMed, Semantic Scholar). It iteratively refines its understanding and synthesizes insights across diverse sources to answer complex, multi-layered scientific questions. A dedicated technical article detailing the agent-based mode, including its routing strategies and experimental evaluations, is in preparation.

2.4. Entity-aware Retrieval:

Identifier queries (e.g., short protein codes) are particularly challenging, since embedding models rarely encode domain-specific tokens properly due to their underrepresentation in training corpora. In our system, we have explored a hybrid retrieval strategy that combines a dense vector retriever with BM25 [11], a well-known sparse retriever. Dense search provides descriptive context, while sparse matching preserves exact term retrieval. This combination improves recall on exact matches and enhances retrieval of domain-specific terms.

2.5. Conversational Layer:

The final layer is the web interface that connects the retrieval core with the end user. Users first create an account on the platform to manage multi-user sessions. All users have access to their previous chat history.

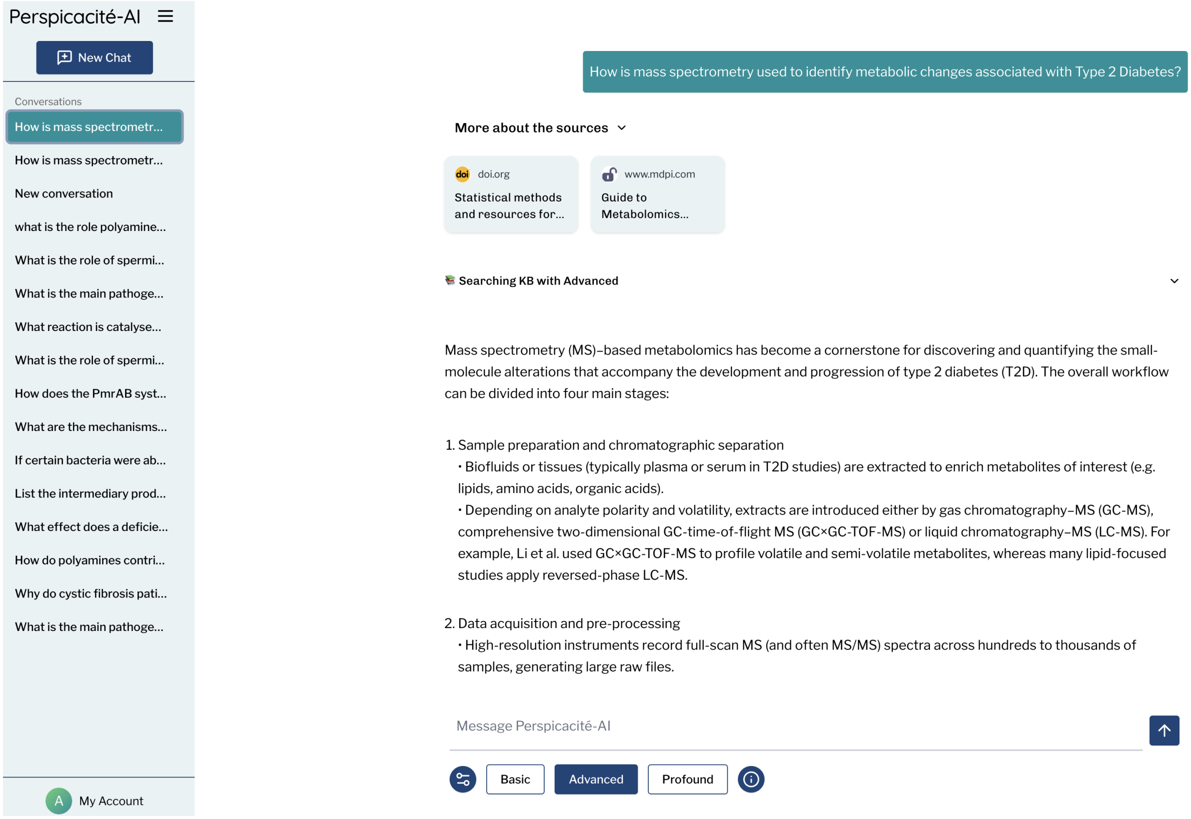
A user starts by typing a question that will be answered using the pre-built knowledge base that has been prepared in advance. Then, they select one of the answering strategies. Each mode can be further parameterized to fit the user's needs, e.g. the number of documents to retrieve and the number of query rephrases.

The results are then streamed with inline citations that link back to the source article, PDF, website, code or YouTube video, ensuring verifiable, up-to-date answers. Figure 2 shows a screenshot of the results of asking a question in the profonde mode in Perspicacité-AI. The online prototype uses a curated computational mass spectrometry knowledge base and is available at: https://metaboguide.holobiomicslab.eu/. The code is open source and can be found on GitHub[1].

3. Conclusion and Future Work

This demo illustrates a KG-driven workflow that can streamline the creation of scientific question-answering systems, providing an end-to-end solution. By using keyword-based SPARQL queries in DBLP, KG2Bib assembles a focused BibTeX corpus that integrates directly into the Perspicacité-AI RAG pipeline, giving users rapid access to well-sourced literature without manual curation. Because each component adheres to standard data formats (BibTeX, FAISS, and JSON), the workflow is both seamless and implementation-agnostic. Future research will involve generalizing this workflow to additional KGs to extend coverage across disciplines and explore more complex KG query logic beyond keyword matching. Moreover, we plan to extend KG2Bib beyond keyword restrictions by incorporating semantic expansion and more advanced ranking functions to improve retrieval precision in interdisciplinary queries. We also plan to develop adaptive ingestion strategies for heterogeneous data sources, moving beyond the current standardized pipeline to optimize vectorization and downstream retrieval quality.

[1] https://github.com/HolobiomicsLab/Perspicacite-AI


Perspicacité-AI
New Chat
Conversations
How is mass spectrometr...
How is mass spectrometr...
New conversation
what is the role polyamine...
What is the role of spermi...
What is the main pathoge...
What reaction is catalyse...
What is the role of spermi...
How does the PmrAB syst...
What are the mechanisms...
If certain bacteria were ab...
List the intermediary prod...
What effect does a deficie...
How do polyamines contri...
Why do cystic fibrosis pati...
What is the main pathoge...
My Account
More about the sources
doi.org
Statistical methods and resources for...
www.mdpi.com
Guide to Metabolomics...
Searching KB with Advanced
Message Perspicacité-AI
Basic
Advanced
Profound

How is mass spectrometry used to identify metabolic changes associated with Type 2 Diabetes?

Mass spectrometry (MS)–based metabolomics has become a cornerstone for discovering and quantifying the small-molecule alterations that accompany the development and progression of type 2 diabetes (T2D). The overall workflow can be divided into four main stages:

1. Sample preparation and chromatographic separation
   - Biofluids or tissues (typically plasma or serum in T2D studies) are extracted to enrich metabolites of interest (e.g. lipids, amino acids, organic acids).
   - Depending on analyte polarity and volatility, extracts are introduced either by gas chromatography–MS (GC-MS), comprehensive two-dimensional GC-time-of-flight MS (GC×GC-TOF-MS) or liquid chromatography–MS (LC-MS). For example, Li et al. used GC×GC-TOF-MS to profile volatile and semi-volatile metabolites, whereas many lipid-focused studies apply reversed-phase LC-MS.
2. Data acquisition and pre-processing
   - High-resolution instruments record full-scan MS (and often MS/MS) spectra across hundreds to thousands of samples, generating large raw files.

Figure 2: A screenshot of the results of asking a question in the "Advanced" mode in Perspicacité-AI using a mass spectrometry curated knowledge base.

Acknowledgments

This work was supported by the French government through the France 2030 investment plan managed by the National Research Agency (ANR), as part of the Initiative of Excellence Université Côte d'Azur under reference number ANR-15-IDEX-01. Lucas Pradi is supported by the Fonds National de la Recherche, Luxembourg, project 17994255. Additional support came from the French Government's France 2030 investment plan (ANR-22-CPJ2-0048-01), as well as through the MetaboLinkAI bilateral project (ANR-24-CE93-0012-01 and SNSF 10002786).

Declaration on Generative AI

During the preparation of this work, the author(s) used ChatGPT/Claude and DeepL for the following: Grammar and spelling checks. After using these tools/services, the author(s) reviewed and edited the content as needed, taking full responsibility for the publication's content.

References

[1] Y. Russell, Three problems of interdisciplinarity 13 (2022) 1–19.

[2] E. Pain, How to keep up with the scientific literature, Science Careers 30 (2016).

[3] W. X. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong, Y. Du, C. Yang, Y. Chen, Z. Chen, J. Jiang, R. Ren, Y. Li, X. Tang, Z. Liu, P. Liu, J.-Y. Nie, J.-R. Wen, A survey of large language models, 2025. URL: https://arxiv.org/abs/2303.18223. arXiv:2303.18223.

[4] J. G. Borger, A. P. Ng, H. Anderton, G. W. Ashdown, M. Auld, M. E. Blewitt, D. V. Brown, M. J. Call, P. Collins, S. Freytag, et al., Artificial intelligence takes center stage: exploring the capabilities

and implications of chatgpt and other ai-assisted technologies in scientific research and education, Immunology and cell biology 101 (2023) 923–935.

[5] I. Mirzadeh, K. Alizadeh, H. Shahrokhi, O. Tuzel, S. Bengio, M. Farajtabar, Gsm-symbolic: Understanding the limitations of mathematical reasoning in large language models, 2024. URL: https://arxiv.org/abs/2410.05229. arXiv:2410.05229.

[6] B. Jiang, Y. Xie, Z. Hao, X. Wang, T. Mallick, W. J. Su, C. J. Taylor, D. Roth, A peek into token bias: Large language models are not yet genuine reasoners, 2024. URL: https://arxiv.org/abs/2406.11050. arXiv:2406.11050.

[7] S. Farquhar, J. Kossen, L. Kuhn, Y. Gal, Detecting hallucinations in large language models using semantic entropy, Nature 630 (2024) 625–630.

[8] L. Huang, W. Yu, W. Ma, W. Zhong, Z. Feng, H. Wang, Q. Chen, W. Peng, X. Feng, B. Qin, T. Liu, A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions, ACM Transactions on Information Systems 43 (2025) 1–55. URL: http://dx.doi.org/10.1145/3703155. doi:10.1145/3703155.

[9] Z. Xu, S. Jain, M. Kankanhalli, Hallucination is inevitable: An innate limitation of large language models, 2025. URL: https://arxiv.org/abs/2401.11817. arXiv:2401.11817.

[10] M. R. Ackermann, H. Bast, B. M. Beckermann, J. Kalmbach, P. Neises, S. Ollinger, The dblp Knowledge Graph and SPARQL Endpoint, Transactions on Graph Data and Knowledge 2 (2024) 3:1–3:23. URL: https://drops.dagstuhl.de/entities/document/10.4230/TGDK.2.2.3. doi:10.4230/TGDK.2.2.3.

[11] S. Robertson, H. Zaragoza, The probabilistic relevance framework: Bm25 and beyond, Foundations and Trends® in Information Retrieval 3 (2009) 333–389. URL: http://dx.doi.org/10.1561/1500000019. doi:10.1561/1500000019.