Artificial Intelligence and Large Language Model Powered Literature Review Services

Ayman Musleh*[1] and Saif Aldeen Alryalat[2,3]

[1]The University of Jordan, Amman, Jordan
[2]Department of Ophthalmology, The University of Jordan, Amman, Jordan
[3]Department of Ophthalmology, University of Colorado School of Medicine, Sue Anschutz-Rodgers Eye Centre, Aurora, Colorado, USA

Keywords: Large Language Models, LLMs, Natural language processing, Artificial Intelligence

DOI: https://doi.org/10.59707/hymrPSEY7778

Published on: June 1, 2025

Abstract

Large language model (LLM) tools are transforming the way evidence is retrieved by converting natural language prompts into quick, synthesized outputs. These platforms significantly reduce the time required for literature searches, making them more accessible to users unfamiliar with formal search strategies. A close evaluation of four prominent platforms—Undermind.ai, Scite.ai, Consensus.app, and OpenEvidence—highlights both notable advantages and ongoing limitations. Undermind and Consensus utilize the extensive Semantic Scholar database of over 200 million records, Scite enhances results with "Smart Citations" that indicate supportive or opposing references, and OpenEvidence applies a medically-focused LLM trained on licensed sources, including the complete NEJM archive. Despite their benefits, key limitations persist: opaque algorithms, inconsistent responses to identical queries, paywalls or sign-up barriers, and incomplete recall that may compromise systematic reviews. To support critical appraisal, we outline essential information-retrieval metrics—including recall, precision, F1-score, mean average precision, and specificity—and provide open-source code. Until validated, transparent evaluations demonstrate consistently high recall, these tools should be viewed as rapid, first-pass aids rather than replacements for structured database searches required by PRISMA-compliant methodologies.

Introduction

Literature review is the process of extracting relevant articles on a certain topic. These articles are stored within specialized literature databases (e.g., PubMed), which store these articles in the process of "Indexing".(1) Indexing is simply storing newly published articles in an organized way, so that prospective researchers can access them through querying these databases.(2) Developing a research query to retrieve specific articles on a topic from a literature database is the key skill in literature review, which requires knowledge and skills to perform. With the evolution of Large Language Models (LLM), several services emerged that facilitate literature review, one of the essential tasks in healthcare and medical research. The main advantage of such

*Corresponding author: Ayman Musleh ¡aimanmesleh@gmail.com¿

services is their ability to take any user input in natural language without the need for an organized search query that each literature database requires.(3) Such an advantage would provide people with little knowledge about literature review databases and their requirements a way to extract evidence for a topic. However, they still have several disadvantages, including the inconsistency of LLM-generated responses, which might lead to different responses for each trial, unlike searching a literature database directly with a query, which will result in the same output each time. More importantly, and especially when performing a systematic review, LLM-powered literature search services may fail in providing all existing literature on a topic in a systematic way with a well-documented approach, which is required in a systematic review. In this article, we will few examples of existing LLM-powered literature review services and how prospective researchers can evaluate their accuracy in literature review in the medical field.

Currently available AI and LLM powered Literature services

Undermind

Undermind.ai was founded in 2023 as an "AI assistant that condenses weeks of research into minutes" (4). The platform positions itself assisting experts with complex scholarly inquiries. To accomplish this, Undermind queries Semantic Scholar corpus, which offers significantly broader cross-disciplinary coverage compared to specialized databases like PubMed alone. (5)(6) Using an embedded large language model, Undermind first interviews users to clarify their research questions. It then conducts iterative semantic and citation-driven searches, classifies retrieved papers by relevance, and emails users a detailed report within 5 to 8 minutes. This report assesses search completeness, ranks papers based on "topic-match" scores, visualizes their citation networks, and can generate an evidence-based narrative upon request (7). The main limitation in the platform is the lack of systematic search, as well as poor methodological reporting on how the articles were retrieved. It has a rapid turnaround time of around 8 minutes and non-transparent, non-reproducible search methodology limit its applicability for point-of-care queries or formal systematic reviews.(7). Undermind operates a freemium model, limiting free users to five abstract-level searches per month, restricted to analyzing abstracts and metadata. (8)

Scite.ai

Scite.ai is an advanced literature discovery and evaluation platform that employs artificial intelligence to facilitate researchers in finding, analyzing, and comprehending scientific publications. The platform integrates LLMs into a user-friendly research assistant capable of responding to natural-language queries with directly cited evidence from literature sources. (9)A distinctive feature of Scite.ai is "Smart Citations," which extend beyond traditional citation metrics by providing context regarding the intent behind citations—whether they support, contrast, or merely mention previous work. The Scite.ai corpus encompasses articles, book chapters, preprints, and datasets. (10)Its AI-powered Research Assistant can automatically extract structured data such as PICO (Population, Intervention, Comparison, Outcome) terms or lists of biomarkers when set to Table Mode. (11) Additionally, Scite.ai offers innovative tools such as visual citation maps, customizable dashboards, real-time citation alerts, and a Reference Check feature that examines manuscript drafts for retracted or heavily contested sources. The platform suffers from several limitations, including a lack of systematic search, poor methodological reporting on how the articles were retrieved, other than its citation retrieval capabilities. Retrieving non-peer-reviewed content might limit its use for scientific articles (12).

Consensus.app

Consensus.app is a researcher-oriented platform combining large-scale academic searches with LLM-driven synthesis. Launched in 2022, Consensus accesses papers from Semantic Scholar, employing keyword and vector retrieval enhanced by a proprietary scoring algorithm prioritizing recency, citation impact, and study design. Utilizing LLM, Consensus generates concise key takeaways and summaries from top-ranked search results. (13) Notable workflow tools include the Consensus Meter, which quantitatively categorizes the first 20 search results for yes-or-no queries; Study Snapshots that detail population, sample size, methods, and outcomes; and the Pro Analysis (Copilot) for citation-linked follow-up insights. Additional features such as advanced filtering (study design, sample size, publication year, access status) and export capabilities (CSV/RIS, auto-citations) enhance the transparency and efficiency of literature reviews.(14) A subscription-based platform, which also suffer from similar limitations as with other ones . (15)

OpenEvidence

OpenEvidence is a specialized AI-driven medical information platform designed to help healthcare professionals stay current with evidence-based literature. (16) The platform aims to organize and expand global medical knowledge, mitigating information overload for clinicians. (17) OpenEvidence operates as an AI-powered medical literature co-pilot, leveraging a specialized LLM trained exclusively on medical content. Research by OpenEvidence has demonstrated that these clinical-specific models significantly outperform general-purpose LLMs in medical contexts. (18)

OpenEvidence curates information from reputable medical sources rather than general web content. Its comprehensive database covers the extensive history of medical publications, supported by partnerships with premier medical publishers. (19). As with the other examples, it lacks the methodological clarity on how it retrieved the literature.

Evaluating the Accuracy of Literature Review Service

Several metrics can be used to evaluate the accuracy of a literature review done on a topic, regardless of how it was done (20). The key is to check how many relevant articles were retrieved, how many were missed, and how many irrelevant articles were retrieved. After that, several metrics can be calculated accordingly. The most commonly used metrics are: • Recall (most important): The proportion of relevant documents retrieved out of all relevant documents available (true positives / (true positives + false negatives)). • Precision: The proportion of retrieved documents that are relevant (true positives / (true positives + false positives)). • F1 Score: The harmonic means of precision and recall, useful for balancing the two when one metric alone isn't enough. • Mean Average Precision (MAP): Useful if your system ranks results, as it considers the order of relevance across multiple queries. • Specificity: Measures how well the system avoids irrelevant results (true negatives / (true negatives + false positives)), though this is less common in retrieval systems. We developed well-structured codes that can be used to calculate these metrics (Supplementary material). Datasets that can be used to assess these metrics can be obtained through a recently published high quality systematic review, taking into account the need to have the dataset for an up-to-date article. Alternatively, there are several established datasets like TREC Precision Medicine (https://trec.nist.gov/data/precmed.html) or BioASQ (https://www.bioasq.org/about), which provide pre-annotated query-result pairs for medical retrieval tasks. (21,22)

Conclusion

AI-driven platforms provide a promise of compressing hours of manual searching into minutes, democratizing access to evidence. They generally utilize LLMs. However, until transparent, reproducible search logs and consistently high recall are guaranteed, researchers should treat LLM-based outputs as a helpful first pass, followed by conventional database searches to ensure completeness and compliance with PRISMA guidelines.

Conflict of Interest

The authors declare that they have no competing interests.

Acknowledgements

There are no acknowledgements.

Financial Support

There was no funding.

References

[1] Lawrence S, Bollacker K, Giles CL. Indexing and retrieval of scientific literature. In: Proceedings of the eighth international conference on Information and knowledge management. New York, NY, USA: ACM; 1999. p. 139–46.

[2] Badami M, Benatallah B, Baez M. Adaptive search query generation and refinement in systematic literature review. Inf Syst. 2023 Jul;117:102231.

[3] Bramer WM, Rethlefsen ML, Kleijnen J, Franco OH. Optimal database combinations for literature searches in systematic reviews: a prospective exploratory study. Syst Rev. 2017 Dec 6;6(1):245.

[4] Ycombinator - Undermind [Internet]. [cited 2025 May 14]. Available from: https://www.ycombinator.com/companies/undermind

[5] Undermind.AI [Internet]. [cited 2025 May 14]. Available from: https://www.undermind.ai/

[6] Semantic Scholar [Internet]. [cited 2025 May 14]. Available from: https://www.semanticscholar.org/about

[7] Giustini D. Undermind.ai (Product Review). 2025 Mar;

[8] Undermind.ai - Pricing [Internet]. [cited 2025 May 12]. Available from: https://www.undermind.ai/#pricing

[9] Knowledgie - scite vs. perplexity which is better? How does Knowledgie compare? [Internet]. [cited 2025 May 14]. Available from: https://www.knowledgie.com/compare/scite/perplexity

[10] Scite.ai [Internet]. [cited 2025 May 14]. Available from: http://scite.ai/data-and-services

[11] Hong Kong University of Science and Technology (HKUST) - Use Scite.ai like a Pro [Internet]. [cited 2025 May 14]. Available from: http://knowledgie.com/compare/scite/perplexity

[12] Scite.ai - Pricing [Internet]. [cited 2025 May 12]. Available from: https://scite.ai/pricing

[13] Consensus.app [Internet]. [cited 2025 May 14]. Available from: https://consensus.app/

[14] Consensus - Using the Study Snapshot [Internet]. [cited 2025 May 14]. Available from: https://help.consensus.app/en/articles/10065008-using-the-study-snapshot?

[15] Consensus.app - Pricing [Internet]. [cited 2025 May 12]. Available from: https://consensus.app/home/pricing/

[16] 1OpenEvidence - OpenEvidence to Become a Mayo Clinic Platform Accelerate Company [Internet]. [cited 2025 May 14]. Available from: https://www.openevidence.com/announcements/openevidence-to-become-a-mayo-clinic-platform-accelerate-company

[17] OpenEvidence [Internet]. [cited 2025 May 14]. Available from: https://www.openevidence.com/about

[18] Wu V, Casauay J. OpenEvidence. Fam Med. 2025 Mar 5;57(3):232–3.

[19] OpenEvidence - OpenEvidence and NEJM Group, publisher of the New England Journal of Medicine, sign content agreement [Internet]. [cited 2025 May 14]. Available from: https://www.openevidence.com/announcements/openevidence-and-nejm

[20] Bramer WM, Giustini D, Kramer BMR. Comparing the coverage, recall, and precision of searches for 120 systematic reviews in Embase, MEDLINE, and Google Scholar: a prospective study. Syst Rev. 2016 Dec 1;5(1):39.

[21] Krithara A, Nentidis A, Bougiatiotis K, Paliouras G. BioASQ-QA: A manually curated corpus for Biomedical Question Answering. Sci Data. 2023 Mar 27;10(1):170.

[22] Roberts K, Demner-Fushman D, Voorhees EM, Bedrick S, Hersh WR. Overview of the TREC 2020 Precision Medicine Track. Text Retr Conf. 2020 Nov;1266.