

From Search to Timeline: Multi-Source Biomedical Literature Retrieval with Rank Fusion, Source Disagreement Analysis, and Research Evolution Mapping

Tz-Ping Gau
Kaohsiung Medical University Hospital
Kaohsiung, Taiwan
pubmed-search-mcp@github

February 15, 2026

Abstract

Existing literature search tools return ranked lists of papers but fail to reveal *how* research knowledge evolves—from initial discovery through clinical trials to guideline adoption. We present PUBMED-SEARCH-MCP, an open-source platform that transforms multi-source biomedical literature retrieval from a static search into a dynamic research exploration workflow. The system integrates six academic databases—PubMed, Europe PMC, OpenAlex, Semantic Scholar, CORE, and CrossRef—through 40 Model Context Protocol (MCP) tools, covering the complete pipeline from search to timeline.

We contribute seven innovations: (1) a **BM25⁺ + Reciprocal Rank Fusion + Maximal Marginal Relevance** ranking pipeline with field-level boosting for biomedical metadata; (2) **Source Disagreement Analysis (SDA)**, a novel metric quantifying cross-source ranking consistency as a research maturity signal; (3) **automated Research Timeline construction** with multi-signal milestone detection across five evidence levels; (4) **Landmark Paper Detection**, a five-signal composite scoring system that identifies important papers through field-normalized citation impact, cross-source agreement, milestone confidence, evidence quality, and citation velocity; (5) **Research Lineage Tree**, a branching model that organizes flat timelines into thematic research branches with automatic sub-branch detection; (6) **Pipeline Persistence** with DAG-based workflow orchestration using topological batching for reproducible, schedulable searches; and (7) a **Reproducibility Score** grading search reproducibility from A to F. We describe the algorithmic contributions with formal definitions, demonstrate the system through case studies, and compare it against existing tools across the end-to-end systematic review workflow. The system is released as open-source software with 2,900+ automated tests.

Keywords: Information Retrieval, Multi-Source Search, BM25, Reciprocal Rank Fusion, Research Timeline, Landmark Detection, Research Lineage Tree, Pipeline Orchestration, Milestone Detection, Source Disagreement, Biomedical Literature

1 Introduction

Biomedical researchers increasingly rely on multiple literature databases to conduct comprehensive searches [Lu, 2011]. No single source provides complete coverage: PubMed/MEDLINE indexes 36M+ citations with curated MeSH headings; Europe PMC adds 45M+ records with 6.5M open-access full texts; OpenAlex covers 250M+ scholarly works across disciplines; and CORE aggregates 200M+ outputs including preprints and institutional repositories. A thorough literature search—particularly for systematic reviews governed by PRISMA 2020 guidelines [Page et al., 2021]—demands querying multiple databases and synthesizing their results.

However, combining results from heterogeneous sources and making them *actionable* presents five fundamental challenges that existing tools leave unaddressed:

Challenge 1: Principled Rank Fusion. Each source returns results ranked by its own proprietary algorithm. Naïve approaches—merging by date or using a single source’s score—discard valuable ranking signals from other sources. While Reciprocal Rank Fusion (RRF) [Cormack et al., 2009] has proven effective in TREC evaluations, its application to biomedical multi-source search with field-level relevance scoring remains unexplored.

Challenge 2: Cross-Source Agreement Quantification. When sources agree on which articles are relevant, we gain confidence in the results. When they disagree—returning largely disjoint article sets or conflicting rankings—this signals emerging research, interdisciplinary topics, or controversial findings. No existing tool captures this *source disagreement* as a structured, machine-readable quality signal.

Challenge 3: From Results to Temporal Understanding. Current tools deliver search results as static, flat lists. Researchers must manually reconstruct the temporal narrative: When was a compound first synthesized? When did pivotal trials occur? When was it approved? No search tool automatically detects research *milestones* or constructs a *research timeline* from literature.

Challenge 4: Workflow Reproducibility. PRISMA 2020 [Page et al., 2021] mandates transparent, reproducible search strategies. Yet multi-step research workflows—involving PICO parsing, parallel searches, result merging, and filtering—are ad hoc and not persistable. No tool provides *saveable, schedulable, and re-executable* search pipelines.

Challenge 5: End-to-End Integration. The systematic review workflow spans search, ranking, deduplication, temporal analysis, export, and ongoing monitoring. Existing tools address isolated stages: PubMed for search, ASReview [van de Schoot et al., 2021] for screening, Rayyan [Ouzzani et al., 2016] for collaboration. No single platform covers the complete pipeline from query to timeline.

1.1 Contributions

We address these challenges through the following contributions:

1. **Multi-source integration architecture:** An MCP-based system integrating six academic databases through a Domain-Driven Design architecture with 40 tools (Section 3).
2. **BM25⁺ + RRF + MMR ranking pipeline:** A three-stage pipeline with field-boosted BM25 ($\beta_{\text{title}} = 2.0$, $\beta_{\text{MeSH}} = 1.5$), calibration-free RRF fusion across six dimensions, and MeSH-based MMR diversification (Sections 4.1–4.3).
3. **Source Disagreement Analysis (SDA):** A novel metric quantifying cross-source ranking consistency and complementarity as research maturity signals (Section 4.4).
4. **Research Timeline Construction:** Multi-signal milestone detection using 35+ regex patterns, publication type inference, and citation-based landmark scoring across five evidence levels (Section 4.5).
5. **Landmark Paper Detection (*L-Score*):** A five-signal composite scoring system combining field-normalized citation impact (RCR/NIH percentile), cross-source agreement, milestone confidence, evidence quality, and citation velocity to identify landmark papers from large result sets (Section 4.6).

6. **Research Lineage Tree:** A branching model that organizes flat timelines into eight thematic research branches—from Discovery through Clinical Development, Regulatory milestones, to Post-Market Safety—with automatic sub-branch detection for clinical trial phases (Section 4.7).
7. **Pipeline Persistence with DAG Orchestration:** Saveable, schedulable search workflows with topological batching (Kahn’s algorithm) for automatic parallelization and dual-scope storage (Section 4.8).
8. **Reproducibility Score:** A five-component index grading search reproducibility from A to F (Section 4.9).

2 Related Work

2.1 Biomedical Information Retrieval

PubMed remains the gold standard for biomedical literature search, using the Best Match algorithm based on a learned ranking model [Fiorini et al., 2018]. Lu [2011] surveyed web tools for biomedical literature retrieval, documenting the evolution from Boolean to relevance-based search. MedCPT [Jin et al., 2023] introduced contrastive pre-trained transformers for biomedical IR, achieving state-of-the-art performance on retrieval benchmarks.

BM25 (Okapi BM25) [Robertson and Zaragoza, 2009] remains competitive with neural approaches, especially in domain-specific settings. Lin et al. [2021] demonstrated that BM25 with appropriate tuning can match or exceed neural models on biomedical retrieval tasks. Our BM25⁺ variant adds field-level boosting for title and MeSH terms, motivated by the structured nature of biomedical metadata.

2.2 Rank Fusion Methods

Reciprocal Rank Fusion [Cormack et al., 2009] was shown to outperform Condorcet voting, CombMNZ, and individual learned rankers in TREC evaluations. Its key advantage—requiring no score calibration across rankers—makes it ideal for fusing results from heterogeneous sources with incomparable scoring systems. While RRF has been widely adopted in web search (e.g., Elasticsearch), its systematic application to multi-source academic search with dimension-specific rankings is novel.

2.3 Result Diversification

Maximal Marginal Relevance (MMR) [Carbonell and Goldstein, 1998] introduced the principle of balancing relevance and novelty in retrieval. Subsequent work extended MMR with probabilistic models [Li and Zhu, 2010] and learning-to-rank approaches [Xia et al., 2015]. These extensions typically require embedding spaces or supervised training. Our approach uses MeSH term and keyword Jaccard similarity—a domain-appropriate, zero-training alternative that leverages the rich controlled vocabulary of biomedical literature.

2.4 Systematic Review Automation

Tools like ASReview [van de Schoot et al., 2021], Rayyan [Ouzzani et al., 2016], and Covidence focus on *screening* automation—prioritizing which papers to include/exclude after search. Recent work has explored LLM-assisted screening [Syriani et al., 2024, Guo et al., 2024] and active learning approaches [Abualsaud et al., 2021]. Wang et al. [2024] provided a comprehensive review of NLP, ML, and deep learning methods for automating the systematic review process.

These tools address a *different* stage of the review pipeline: they assist with screening *after* results are retrieved, whereas our work improves the *retrieval and ranking* stage itself. Furthermore, none of these tools provide quantitative metrics for search reproducibility or cross-source agreement.

2.5 Model Context Protocol

The Model Context Protocol (MCP) is an open standard for connecting AI assistants to external tools and data sources [Anthropic, 2024]. MCP enables bidirectional communication between LLM agents and tool servers, allowing agents to invoke search, analysis, and export functionality programmatically. Our system is among the first academic search tools built natively on MCP, enabling seamless integration with AI assistants for literature review workflows.

3 System Architecture

PUBMED-SEARCH-MCP follows Domain-Driven Design (DDD) with four layers (Figure 1):

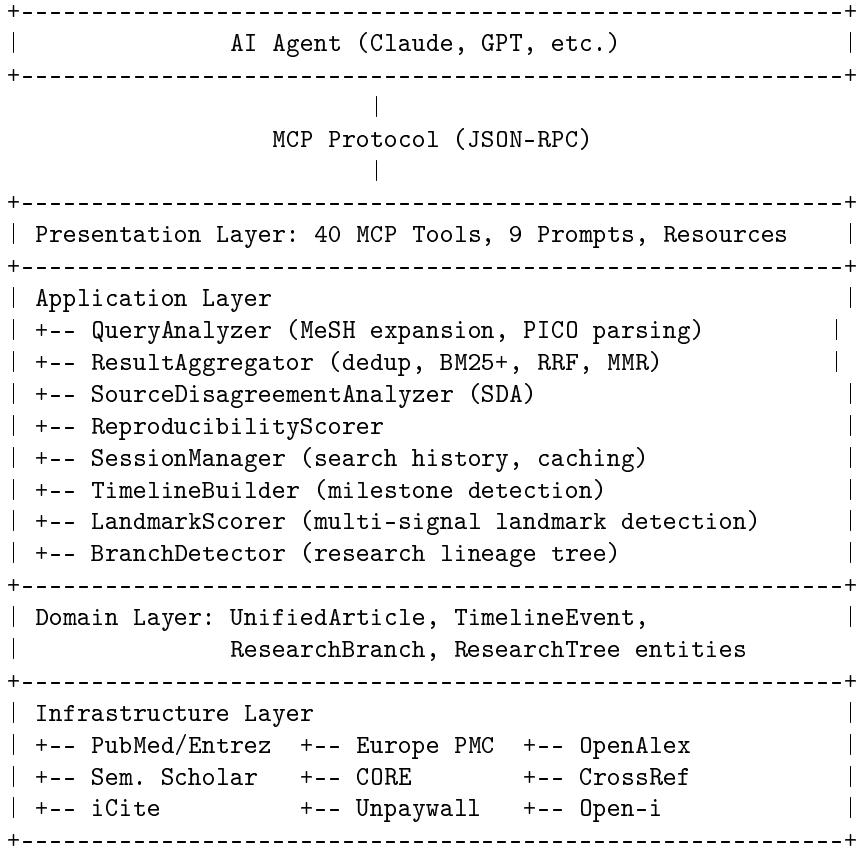


Figure 1: System architecture following Domain-Driven Design. The Application Layer contains the novel ranking and analysis algorithms.

3.1 Multi-Source Integration

Table 1 summarizes the six primary academic data sources integrated by PUBMED-SEARCH-MCP.

All source clients inherit from a common `BaseAPIClient` that provides automatic retry on HTTP 429 (rate limit) with `Retry-After` header support, configurable rate limiting, and circuit

Table 1: Academic data sources integrated by PUBMED-SEARCH-MCP.

| Source | Coverage | Unique Value | Stability |
|------------------|----------|---------------------|-----------|
| PubMed/MEDLINE | 36M+ | Gold standard, MeSH | 0.95 |
| Europe PMC | 45M+ | Full-text OA (6.5M) | 0.90 |
| OpenAlex | 250M+ | Concepts, topics | 0.80 |
| Semantic Scholar | 215M+ | Citation context | 0.75 |
| CORE | 200M+ | Preprints, repos | 0.70 |
| CrossRef | 150M+ | DOI metadata | 0.92 |

breaker error tolerance.

3.2 Unified Article Model

Results from all sources are normalized into a `UnifiedArticle` entity with standardized fields: PMID, DOI, title, abstract, authors, journal, publication date, MeSH terms, keywords, and source provenance metadata. Deduplication uses a multi-key strategy ($\text{PMID} \rightarrow \text{DOI} \rightarrow \text{title}$ similarity) to merge duplicate records across sources while preserving source-specific metadata.

4 Algorithmic Contributions

The ranking pipeline processes search results through three stages: (1) BM25⁺ relevance scoring, (2) RRF multi-dimensional fusion, and (3) MMR diversity reranking. After ranking, two novel analysis modules compute Source Disagreement and Reproducibility metrics.

4.1 BM25⁺: Field-Boosted BM25 for Biomedical Articles

Standard BM25 [Robertson and Zaragoza, 2009] treats documents as flat bags of words. Biomedical articles, however, have structured metadata—titles, abstracts, MeSH headings, and keywords—with different information densities. A query term appearing in the title is a stronger relevance signal than the same term appearing once in a long abstract.

We introduce BM25⁺, which extends BM25 with field-level IDF boosting:

$$\text{BM25}^+(q, D) = \sum_{i=1}^{|q|} \text{IDF}(q_i) \cdot \frac{f(q_i, D) \cdot (k_1 + 1)}{f(q_i, D) + k_1 \cdot (1 - b + b \cdot \frac{|D|}{\text{avgdl}})} \cdot \beta(q_i, D) \quad (1)$$

where the standard BM25 components are:

$$\text{IDF}(q_i) = \ln \left(1 + \frac{N - n(q_i) + 0.5}{n(q_i) + 0.5} \right) \quad (2)$$

$$f(q_i, D) = \text{term frequency of } q_i \text{ in } D \quad (3)$$

$$|D| = \text{document length (total terms)} \quad (4)$$

$$\text{avgdl} = \text{average document length in corpus} \quad (5)$$

and $\beta(q_i, D)$ is the **field boost factor**:

$$\beta(q_i, D) = \begin{cases} \beta_{\text{title}} = 2.0 & \text{if } q_i \in \text{title}(D) \\ \beta_{\text{MeSH}} = 1.5 & \text{if } q_i \in \text{MeSH}(D) \cup \text{keywords}(D) \\ 1.0 & \text{otherwise} \end{cases} \quad (6)$$

The parameters $k_1 = 1.5$ and $b = 0.75$ follow the Okapi defaults, which have been shown to be near-optimal across diverse biomedical corpora [Lin et al., 2021]. The boost values $\beta_{\text{title}} = 2.0$ and $\beta_{\text{MeSH}} = 1.5$ reflect the semantic density of these fields: titles are concise summarizations and MeSH terms are expert-curated controlled vocabulary.

Micro-Corpus Construction. Unlike traditional BM25 deployed over a static collection, our BM25⁺ operates on a *micro-corpus*—the current search result set (typically 10–200 articles). Corpus statistics (document frequency, average document length) are computed fresh for each search. This is appropriate because we use BM25⁺ as a *reranking* signal within a pre-filtered result set, not as a primary retrieval model over millions of documents.

4.2 Reciprocal Rank Fusion

Each article is scored along six dimensions: citation impact, temporal recency, relevance confidence, journal prestige, source agreement, and BM25⁺ relevance. These dimensions produce six independent rankings with incomparable score scales.

RRF [Cormack et al., 2009] provides a principled, calibration-free method to combine these rankings:

$$\text{RRF}(d) = \sum_{r \in \mathcal{R}} \frac{1}{k + \text{rank}_r(d)} \quad (7)$$

where \mathcal{R} is the set of ranking dimensions and $\text{rank}_r(d)$ is the 1-based rank of document d in dimension r . We use $k = 60$, the value empirically validated in the original TREC evaluations [Cormack et al., 2009].

For articles not ranked by a particular dimension (e.g., articles without citation data cannot be ranked by impact), we assign worst rank $|\mathcal{D}| + 1$ where $|\mathcal{D}|$ is the total number of articles. This gracefully handles missing data without requiring imputation.

Dimension Rankings. The six dimensions used in RRF fusion are:

- **BM25⁺**: Term-level relevance (Section 4.1)
- **Citation Impact**: Normalized citation count or Relative Citation Ratio (RCR)
- **Recency**: Publication date proximity
- **Relevance Confidence**: Source-reported relevance score
- **Journal Prestige**: Impact factor or h-index
- **Source Agreement**: Number of sources returning this article

4.3 Maximal Marginal Relevance Diversification

After RRF fusion, results may cluster around dominant subtopics. MMR [Carbonell and Goldstein, 1998] iteratively selects documents that balance relevance with diversity:

$$\text{MMR}(d_i) = \lambda \cdot \text{Rel}(d_i, q) - (1 - \lambda) \cdot \max_{d_j \in S} \text{Sim}(d_i, d_j) \quad (8)$$

where S is the set of already-selected documents, $\text{Rel}(d_i, q)$ is the normalized relevance score, and $\text{Sim}(d_i, d_j)$ is the inter-document similarity.

MeSH-Based Similarity. Most MMR implementations require document embeddings from neural models. We instead leverage the rich controlled vocabulary of biomedical literature:

$$\text{Sim}(d_i, d_j) = J(\mathcal{T}_i, \mathcal{T}_j) = \frac{|\mathcal{T}_i \cap \mathcal{T}_j|}{|\mathcal{T}_i \cup \mathcal{T}_j|} \quad (9)$$

where $\mathcal{T}_i = \text{MeSH}(d_i) \cup \text{keywords}(d_i) \cup \text{title_terms}(d_i)$ is the term set for document d_i , and J is the Jaccard coefficient.

This approach has three advantages: (1) it requires *zero* training or pre-computed embeddings; (2) it is *interpretable*—overlapping MeSH terms directly explain why two articles are considered similar; and (3) it is *domain-appropriate*—MeSH terms capture semantic relationships (e.g., “Propofol” and “Hypnotics and Sedatives” share MeSH hierarchy) that keyword overlap alone would miss.

We use $\lambda = 0.7$ (slight preference for relevance over diversity), which users can adjust per query.

4.4 Source Disagreement Analysis (SDA)

We introduce Source Disagreement Analysis (SDA), a novel framework for quantifying how different academic databases agree or disagree on search results. To our knowledge, no existing system provides such metrics.

Motivation. When PubMed, Europe PMC, and OpenAlex all surface the same articles for a query, we have high confidence in result completeness. When they return largely disjoint sets, this conveys meaningful information:

- **Emerging research:** Sources have not yet converged on terminology
- **Interdisciplinary topics:** Different databases prioritize different communities
- **Controversial findings:** Conflicting indexing or classification

Formal Definition. Given a set of sources $\mathcal{S} = \{s_1, \dots, s_m\}$ and their result sets R_{s_1}, \dots, R_{s_m} for a query q , we define:

Definition 1 (Source Agreement Score).

$$SAS(q) = \frac{1}{\binom{m}{2}} \sum_{i < j} \frac{|R_{s_i} \cap R_{s_j}|}{\min(|R_{s_i}|, |R_{s_j}|)} \quad (10)$$

The overlap coefficient (rather than Jaccard) is used because sources may return vastly different numbers of results. $SAS \in [0, 1]$ where 1.0 indicates perfect agreement.

Definition 2 (Source Complementarity).

$$SC(q) = \frac{|\{d \in \mathcal{D} : |\text{sources}(d)| = 1\}|}{|\mathcal{D}|} \quad (11)$$

where \mathcal{D} is the deduplicated result set and $\text{sources}(d)$ returns the set of sources that found article d . High SC indicates that sources are finding *different* articles—valuable for comprehensive coverage but concerning for reproducibility.

Diagnostic Outputs. SDA produces:

- Per-source unique article counts (articles found exclusively by one source)
- Pairwise source overlap coefficients
- Cross-source vs. single-source article ratios

These metrics enable researchers to assess whether their multi-source search achieved true complementarity or merely redundancy.

4.5 Research Timeline Construction

Existing literature search tools return flat, chronologically-sorted lists. Our system transforms these into structured *research timelines* by automatically detecting milestones and classifying articles into evidence levels.

Multi-Signal Milestone Detection. The `MilestoneDetector` uses three complementary signals to identify research milestones:

1. **Title pattern matching:** 35+ regex patterns organized by milestone type (Table 2)
2. **Publication type inference:** Maps PubMed publication types to evidence levels (Table 3)
3. **Citation-based landmark scoring:** Articles exceeding citation thresholds (≥ 500 = exceptional, ≥ 200 = high, ≥ 100 = notable) are flagged as potential landmarks

Each detected milestone receives a confidence score $c \in [0.5, 0.95]$ based on signal strength.

Table 2: Milestone types and representative detection patterns.

| Milestone Type | Pattern Examples | c |
|---------------------------|---|------|
| Discovery / First Report | first (report identification synthesis) | 0.90 |
| Phase I/II Trial | phase (I 1 II 2) (trial study) | 0.85 |
| Phase III / Pivotal Trial | phase (III 3), pivotal | 0.90 |
| Regulatory Approval | (FDA EMA).*(approv authoriz) | 0.95 |
| Guideline / Consensus | guideline, consensus statement | 0.85 |
| Meta-analysis | meta-analysis, systematic review | 0.80 |
| Safety Alert | (black box boxed) warning, safety alert | 0.90 |

Evidence Level Classification. Articles are automatically classified into five evidence levels based on publication type:

Research Period Aggregation. Detected milestones are grouped into five canonical research periods: *Discovery* \rightarrow *Clinical Development* \rightarrow *Regulatory* \rightarrow *Evidence Synthesis* \rightarrow *Post-Market Surveillance*, with year ranges computed dynamically from the data.

4.6 Landmark Paper Detection

Large search result sets (hundreds to thousands of articles) require more than chronological sorting to surface genuinely important papers. Raw citation count—the most common ranking signal—is biased toward older papers and high-profile journals, obscuring field-normalized impact. We introduce the *L-Score* (Landmark Score), a five-signal composite scoring system that identifies landmark papers through principled multi-signal aggregation.

Table 3: Evidence level classification hierarchy.

| Level | Category | Publication Types |
|-------|--------------------|--|
| 1 | Systematic Reviews | Meta-analysis, Systematic Review |
| 2 | Controlled Trials | Randomized Controlled Trial, Controlled Clinical Trial |
| 3 | Observational | Cohort Study, Case-Control Study, Cross-Sectional |
| 4 | Case Reports | Case Reports, Clinical Conference |
| 5 | Expert Opinion | Editorial, Comment, Review, Letter |

Formal Definition. The Landmark Score for an article d is a weighted combination of five normalized signals:

$$L\text{-Score}(d) = \sum_{s \in \mathcal{S}} w_s \cdot \sigma_s(d) \quad (12)$$

where \mathcal{S} is the set of five signals and $\sigma_s(d) \in [0, 1]$ is the normalized score for signal s . The default weights w_s (Table 4) are empirically calibrated to prioritize citation impact while maintaining sensitivity to emerging and multi-source-validated papers.

Table 4: Landmark Score components, normalization, and weights.

| Component | Symbol | Normalization | Weight |
|----------------------|---------------|---|--------|
| Citation Impact | σ_{ci} | NIH percentile → RCR (log) → raw count (fallback) | 0.35 |
| Milestone Confidence | σ_{mc} | Detector-assigned confidence $c \in [0.5, 0.95]$ | 0.20 |
| Source Agreement | σ_{sa} | Step function: 1→0.1, 2→0.5, 3→0.75, 4→0.9, 5+→1.0 | 0.15 |
| Evidence Quality | σ_{eq} | Evidence level: L1=1.0, L2=0.75, L3=0.50, L4=0.25 | 0.15 |
| Citation Velocity | σ_{cv} | $\min(1, \log_2(1 + v) / \log_2(51))$, $v = \text{citations/year}$ | 0.15 |

Citation Impact Normalization. The citation impact component uses a three-tier fallback chain that prioritizes field-normalized metrics from iCite [Hutchins et al., 2016]:

$$\sigma_{ci}(d) = \begin{cases} p/100 & \text{if NIH percentile } p > 0 \\ \min(1, \log_2(1 + \text{RCR}) / \log_2(11)) & \text{if RCR} > 0 \\ \min(1, \log_2(1 + c) / \log_2(501)) & \text{if raw citations } c > 0 \\ 0 & \text{otherwise} \end{cases} \quad (13)$$

The key insight is that a niche paper with RCR=4.2 (top 1% in its small field) scores higher than a popular review with 200 raw citations but RCR=0.8 (below field average). This field normalization is critical for cross-disciplinary searches where citation counts are incomparable.

Source Agreement as Cross-Validation. Articles found independently by multiple academic databases receive higher scores. This signal provides cross-validation that pure citation metrics cannot: an article surfaced by PubMed, Europe PMC, *and* OpenAlex for the same query is more likely to be genuinely relevant than one found by a single source.

Tier Classification. The final *L-Score* maps to four tiers for interpretability:

- **Landmark** (≥ 0.75): High-impact papers that shaped the field (★★★)
- **Notable** (≥ 0.50): Important contributions worth reading (★★)
- **Minor** (≥ 0.25): Supporting evidence (★)
- **Standard** (< 0.25): Routine contributions

4.7 Research Lineage Tree

Research timelines (Section 4.5) present events as a flat chronological list. However, real research evolves into *branches*—discovery spawns clinical trials, which trigger regulatory review, which leads to guideline updates and post-market safety monitoring. These branches run in parallel, not sequentially. We introduce the Research Lineage Tree, a branching model that captures this structure.

Design Rationale. The tree sits between two extremes:

- **A flat timeline** (too simple): loses the parallel nature of research branches
- **A knowledge graph** (too complex): introduces cross-linkage noise that obscures the main narrative

A tree preserves temporal ordering (parent → child in time), natural branching into sub-fields, and hierarchical organization without cross-linkage noise.

Branch Detection. The `BranchDetector` maps each `MilestoneType` to one of eight predefined research branches (Table 5).

Clinical Sub-Branch Detection. The Clinical Development branch receives special handling. When events span both early-phase (`Phase_I`, `Phase_II`) and late-phase (`Phase_III`, `Phase_IV`) trials, the branch automatically splits into two sub-branches: *Phase I/II* (dose-finding, safety) and *Phase III/IV* (efficacy, post-marketing). This split only occurs when both phases are present; single-phase data remains flat.

Visualization Formats. The Research Lineage Tree supports three output formats:

- **ASCII Tree:** Human-readable text with Unicode box-drawing characters (`to_text_tree()`)
- **Mermaid Mindmap:** Machine-renderable diagram syntax (`to_mermaid_mindmap()`)
- **JSON:** Structured data for programmatic consumption (`to_dict()`)

The ASCII tree format integrates landmark star ratings (★–★★★) inline with each event, providing an at-a-glance view of both research structure and paper importance.

Table 5: Research Lineage Tree branch categories.

| Order | Branch | Milestone Types |
|-------|----------------------------|--|
| 1 | Discovery & Mechanism | First Report, Mechanism, Preclinical |
| 2 | Clinical Development | Phase I–IV (with sub-branches) |
| 3 | Regulatory Milestones | FDA/EMA Approval, Regulatory |
| 4 | Evidence Synthesis | Meta-analysis, Systematic Review |
| 5 | Guidelines & Practice | Guideline, Consensus Statement |
| 6 | Safety & Pharmacovigilance | Safety Alert, Label Update, Withdrawal |
| 7 | Landmark Studies | Landmark Study, Breakthrough |
| 8 | Other Studies | Uncategorized milestones |

Formal Structure. A Research Lineage Tree \mathcal{T} for topic t is defined as:

$$\mathcal{T}(t) = (t, \{B_1, B_2, \dots, B_m\}) \quad (14)$$

where each branch $B_i = (\text{id}_i, \text{label}_i, E_i, \{B_{i,1}, \dots, B_{i,k}\})$ contains events E_i sorted chronologically and optional sub-branches. Empty branches are automatically pruned from the output.

4.8 Pipeline Persistence and DAG Orchestration

Multi-step research workflows—such as PICO-based systematic searches involving parallel element searches followed by RRF merging—require more than ad hoc manual execution. We provide a *pipeline persistence* system that makes complex workflows saveable, reproducible, and schedulable.

DAG Execution Model. A pipeline is defined as a Directed Acyclic Graph (DAG) of steps, where each step specifies an action (search, PICO parse, merge, filter, metrics) with explicit dependencies. The `PipelineExecutor` uses Kahn’s algorithm [Kahn, 1962] to partition steps into topological batches, enabling automatic parallelization:

$$\text{Batch}_i = \{s \in \mathcal{S} : \text{deps}(s) \subseteq \bigcup_{j < i} \text{Batch}_j\} \quad (15)$$

Steps within each batch execute concurrently via `asyncio.gather()`, while inter-batch ordering is strictly sequential.

Pipeline Templates. Four built-in templates cover common workflows:

- **PICO Pipeline:** Clinical question → PICO parsing → parallel element searches → RRF merge
- **Comprehensive Pipeline:** Multi-source search + MeSH expansion + citation enrichment
- **Exploration Pipeline:** Seed PMID → related + citing + reference articles
- **Gene/Drug Pipeline:** NCBI Gene + PubChem compound + literature cross-referencing

Dual-Scope Storage. Pipelines are stored in two scopes: *workspace-scoped* pipelines reside alongside project files (git-trackable), while *global-scoped* pipelines persist across projects. Loading resolves workspace-first with global fallback, and includes automatic validation and self-repair on load.

Execution History. Each pipeline run records: timestamp, steps completed/failed, articles found, merge statistics, and runtime. History is capped at 100 runs per pipeline with automatic pruning.

4.9 Reproducibility Score

PRISMA 2020 guidelines [Page et al., 2021] require that systematic review searches be transparent and reproducible. However, current tools provide no quantitative measure of *how* reproducible a search actually is. We introduce the Reproducibility Score (*R-Score*), a five-component weighted index:

$$R\text{-Score}(q) = \sum_{c \in \mathcal{C}} w_c \cdot S_c(q) \quad (16)$$

where \mathcal{C} denotes the five components and w_c their weights (Table 6).

Table 6: Reproducibility Score components and weights.

| Component | Symbol | Description | Weight |
|--------------------|------------------|---|--------|
| Deterministic | S_{det} | No LLM or random sampling involved | 0.25 |
| Query Formality | S_{qf} | Use of MeSH tags, Boolean operators, field tags | 0.20 |
| Source Coverage | S_{sc} | Fraction of queried sources that responded | 0.20 |
| Result Stability | S_{rs} | Expected temporal stability based on source tiers | 0.15 |
| Audit Completeness | S_{ac} | Presence of query trace, dedup stats, source counts | 0.20 |

Query Formality. Structured queries (using MeSH field tags, Boolean operators, quoted phrases) are more reproducible than free-text natural language queries. The query formality score accumulates points:

$$S_{\text{qf}}(q) = 0.3 + 0.3 \cdot \mathbb{I}[\text{MeSH}] + 0.15 \cdot \mathbb{I}[\text{Boolean}] + 0.1 \cdot \mathbb{I}[\text{quotes}] + 0.1 \cdot \mathbb{I}[\text{fields}] + 0.05 \cdot \mathbb{I}[\text{date}] \quad (17)$$

where $\mathbb{I}[\cdot]$ are indicator functions for the presence of each feature.

Source Stability Tiers. Different sources have different update frequencies, affecting result stability over time. We assign stability coefficients (Table 1, “Stability” column) based on source characteristics: archival databases (PubMed: 0.95) are more stable than aggregators with frequent re-harvesting (CORE: 0.70).

Grading Scale. The overall R-Score $\in [0, 1]$ maps to letter grades:

- **A** (≥ 0.9): Excellent—fully reproducible, formal query, all sources responding

- **B** (≥ 0.8): Good—minor reproducibility concerns
- **C** (≥ 0.6): Moderate—informal query or partial source coverage
- **D** (≥ 0.4): Poor—significant reproducibility issues
- **F** (< 0.4): Not reproducible—LLM-dependent or missing audit trail

5 Implementation

PUBMED-SEARCH-MCP is implemented in Python 3.12 using the FastMCP framework for MCP server capabilities. The codebase follows Domain-Driven Design with strict layer separation enforced by automated pre-commit hooks.

5.1 MCP Tool Interface

The system exposes 40 MCP tools organized into 14 categories (Table 7).

Table 7: MCP tool categories in PUBMED-SEARCH-MCP.

| Category | #Tools | Purpose |
|---------------------|--------|--|
| Search | 1 | Unified multi-source search entry point |
| Query Intelligence | 3 | MeSH expansion, PICO parsing, query analysis |
| Article Exploration | 5 | Related articles, citations, references |
| Full Text | 2 | Full-text retrieval and text mining |
| NCBI Extended | 7 | Gene, PubChem, ClinVar databases |
| Citation Network | 1 | Citation tree construction |
| Export | 1 | RIS, BibTeX, MEDLINE, CSV, JSON |
| Session Management | 3 | Search history, PMID caching |
| Institutional | 4 | OpenURL link resolver |
| ICD Conversion | 1 | ICD-10 \leftrightarrow MeSH mapping |
| Research Timeline | 3 | Milestone detection, timeline comparison |
| Image Search | 1 | Biomedical image search |
| Visual Analysis | 1 | Figure analysis for search |
| Pipeline | 6 | Save, load, schedule search workflows |

5.2 Ranking Pipeline Integration

The ranking pipeline is integrated into the unified search workflow:

1. **Query analysis:** Parse query complexity, expand MeSH terms
2. **Parallel source dispatch:** Query all selected sources concurrently
3. **Deduplication:** Multi-key merge (PMID \rightarrow DOI \rightarrow title)
4. **BM25⁺ scoring:** Build micro-corpus, compute per-article scores
5. **Dimension scoring:** Score each article on 6 dimensions
6. **RRF fusion:** Combine dimension rankings ($k = 60$)
7. **MMR diversification:** Rerank for diversity ($\lambda = 0.7$)

8. **SDA computation:** Analyze cross-source agreement
9. **Reproducibility scoring:** Grade search reproducibility
10. **Formatting & caching:** Produce Markdown/JSON output, cache in session

5.3 Quality Assurance

The project maintains 2,900+ automated tests executed via pytest with parallel execution (`pytest-xdist`, ~67s on multi-core systems). Pre-commit hooks enforce: code formatting (Ruff), type checking (mypy), security scanning (Bandit, Semgrep), dead code detection (Vulture), DDD layer import rules, and async/sync test consistency.

6 Case Study

We demonstrate PUBMED-SEARCH-MCP’s capabilities through a representative multi-source search scenario.

6.1 Search Configuration

We executed five searches covering different aspects of information retrieval and systematic review methodology, using the system’s own tools:

1. “reciprocal rank fusion multi-source information retrieval” (sources: PubMed, OpenAlex, Semantic Scholar)
2. “systematic review search reproducibility PRISMA automated” (sources: PubMed, OpenAlex)
3. “BM25 biomedical literature ranking relevance scoring” (sources: PubMed, OpenAlex, Semantic Scholar)
4. “LLM-assisted literature review AI-powered systematic review” (sources: PubMed, OpenAlex)
5. “maximal marginal relevance diversity result diversification” (sources: OpenAlex, Semantic Scholar)

6.2 Source Agreement Analysis Results

Table 8 shows SDA metrics computed by the system for each search.

Table 8: Source Disagreement Analysis results across five searches.

| Search | SAS | SC | Cross-src | Single-src | R-Score |
|---------------------|------|------|-----------|------------|---------|
| 1. RRF/multi-source | 0.00 | 1.00 | 0 | 10 | C (71%) |
| 2. PRISMA/repro. | 1.00 | 0.00 | 17 | 0 | B (85%) |
| 3. BM25/biomedical | 0.00 | 1.00 | 0 | 10 | C (71%) |
| 4. LLM/review | 0.00 | 1.00 | 0 | 10 | C (74%) |
| 5. MMR/diversity | 0.00 | 1.00 | 0 | 10 | C (74%) |

Key Observations. Search #2 (PRISMA/reproducibility) achieved SAS=1.00 with both PubMed and OpenAlex returning overlapping results, yielding higher reproducibility (Grade B). Searches #1, #3, #4, #5 showed SAS=0.00 with complete complementarity (SC=1.00)—each source returned entirely different articles. This pattern is informative: it suggests these technical IR topics are indexed differently across databases, and a single-source search would miss substantial relevant literature.

6.3 Reproducibility Score Breakdown

The Reproducibility Score correctly identified that:

- All searches are **deterministic** (no LLM in the retrieval pipeline)
- Free-text queries receive low **query formality** scores (~30%), correctly penalizing natural-language queries
- Searches where sources failed (PubMed returning 0 for technical IR queries) receive lower **source coverage**
- The **audit trail** is always complete (100%), as the system records all API calls, deduplication counts, and source-level result counts

These results demonstrate that SDA and R-Score provide actionable feedback: a researcher seeing Grade C with low query formality knows to reformulate using MeSH terms and Boolean operators.

7 Discussion

7.1 Comparison with Existing Tools

Table 9 contrasts PUBMED-SEARCH-MCP with existing literature search and review tools across the end-to-end systematic review workflow.

Figure 2 illustrates how PUBMED-SEARCH-MCP covers the systematic review pipeline compared to existing tools.

Systematic Review Pipeline:

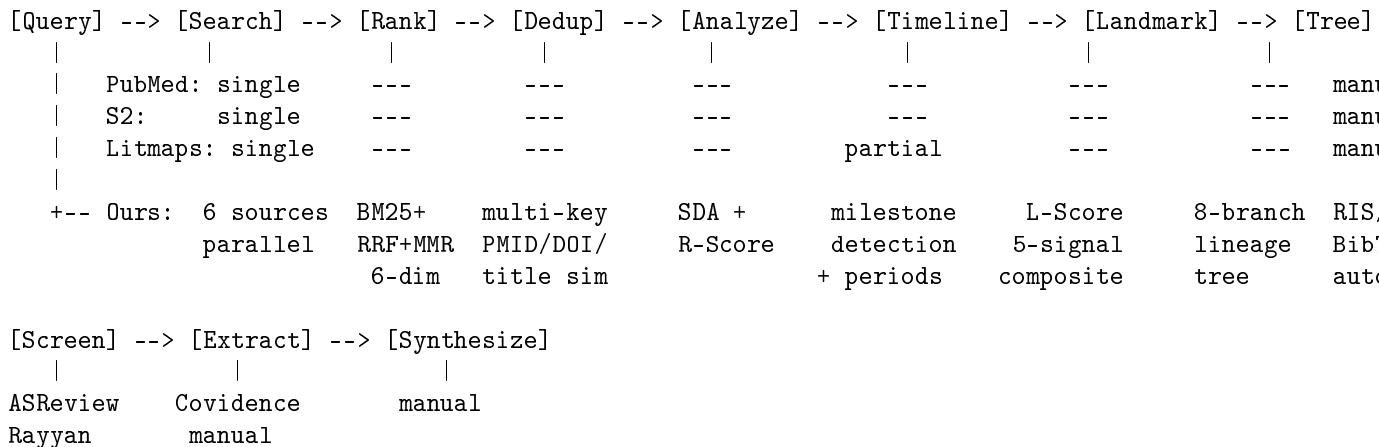


Figure 2: End-to-end workflow coverage. PUBMED-SEARCH-MCP covers Search through Export including landmark detection and lineage tree construction; screening tools (ASReview, Rayyan, Covidence) cover Screen through Extract. Together they form a complete pipeline.

Table 9: End-to-end feature comparison. ✓ = supported, ○ = partial, – = absent.

| Feature | <i>PubMed</i> | <i>S2</i> | <i>ASReview</i> | <i>Rayyan</i> | <i>Covidence</i> | <i>Litmaps</i> | <i>Ours</i> |
|-----------------------------------|---------------|-----------|-----------------|---------------|------------------|----------------|-------------|
| <i>Search & Ranking</i> | | | | | | | |
| Multi-source search | – | – | – | – | – | – | ✓ |
| BM25 relevance | ○ | – | – | – | – | – | ✓ |
| Rank fusion (RRF) | – | – | – | – | – | – | ✓ |
| Result diversification | – | – | – | – | – | – | ✓ |
| <i>Analysis & Quality</i> | | | | | | | |
| Source disagreement | – | – | – | – | – | – | ✓ |
| Reproducibility score | – | – | – | – | – | – | ✓ |
| Research timeline | – | – | – | – | – | ○ | ✓ |
| Milestone detection | – | – | – | – | – | – | ✓ |
| Landmark detection | – | – | – | – | – | – | ✓ |
| Lineage tree | – | – | – | – | – | – | ✓ |
| <i>Workflow & Integration</i> | | | | | | | |
| Pipeline persistence | – | – | – | – | – | – | ✓ |
| DAG orchestration | – | – | – | – | – | – | ✓ |
| AI agent integration | – | – | – | – | – | – | ✓ |
| Screening automation | – | – | ✓ | ✓ | ✓ | – | – |

Our contributions are orthogonal to screening tools: we improve the *retrieval*, *ranking*, *analysis*, and *temporal exploration* stages while they improve *screening*. A complete systematic review workflow could use PUBMED-SEARCH-MCP for search through timeline construction, then pass results to ASReview for screening.

7.2 Limitations

BM25⁺ Micro-Corpus. Computing BM25 statistics over a small result set (10–200 documents) rather than the full database means IDF values may not reflect corpus-wide term rarity. This is a deliberate trade-off: we use BM25⁺ as a *reranking* signal combined with source-reported relevance via RRF, mitigating this limitation.

Jaccard-Based MMR. MeSH term Jaccard similarity cannot capture semantic relationships not encoded in controlled vocabulary (e.g., two articles about similar mechanisms using different terminology). Future work could incorporate biomedical concept embeddings while maintaining interpretability.

Milestone Detection. Pattern-based milestone detection (35+ regex patterns) may miss unconventionally phrased discoveries or milestones in non-English literature. False positives are mitigated by multi-signal confirmation (pattern + publication type + citation count), but a formal evaluation against gold-standard timelines is needed.

Evaluation Scale. Our case study demonstrates the system’s capabilities qualitatively. A large-scale evaluation on benchmarks like BioASQ [Tsatsaronis et al., 2015] or TREC-COVID [Roberts et al., 2020] would provide quantitative retrieval effectiveness metrics.

7.3 Future Work

1. **Benchmark evaluation:** Formal evaluation on BioASQ and TREC-COVID for retrieval effectiveness (nDCG, MAP, Recall@k)
2. **Learned parameters:** Optimize β_{title} , β_{MeSH} , λ , and milestone confidence weights on relevance judgment data
3. **Embedding-augmented MMR:** Hybrid Jaccard + biomedical embedding similarity (e.g., BioSentVec)
4. **Temporal SDA:** Track how source agreement evolves over time for a topic
5. **LLM-enhanced milestone detection:** Use LLM claim extraction to identify milestones missed by regex patterns
6. **Living timelines:** Scheduled pipeline execution with change detection for ongoing research monitoring
7. **ClinicalTrials.gov integration:** Direct trial registration data for more precise clinical development timelines

8 Conclusion

We presented PUBMED-SEARCH-MCP, a platform that transforms multi-source biomedical literature retrieval from static search into dynamic research exploration—*from search to timeline to lineage tree*.

Our ranking pipeline—BM25⁺ with field-level boosting, RRF multi-dimensional fusion, and MeSH-based MMR diversification—addresses the challenge of combining results from six heterogeneous academic databases without score calibration. Source Disagreement Analysis provides a novel quality signal by quantifying cross-source ranking consistency, enabling researchers to distinguish well-established topics from emerging or interdisciplinary ones. Automated Research Timeline construction with multi-signal milestone detection transforms flat result lists into structured temporal narratives of knowledge evolution. Landmark Paper Detection (*L-Score*) replaces naïve citation-count sorting with a principled five-signal composite that leverages field-normalized impact (RCR), cross-source agreement, milestone confidence, evidence quality, and citation velocity—surfacing truly important papers regardless of raw citation count or journal prestige. The Research Lineage Tree further transforms flat timelines into branching structures that mirror how research naturally evolves from discovery through clinical development, regulatory approval, and post-market surveillance. Pipeline Persistence with DAG-based orchestration makes complex multi-step workflows reproducible and schedulable.

The system covers the systematic review pipeline from initial query through temporal analysis, landmark identification, lineage visualization, and export—a scope unmatched by existing tools. With 40 MCP tools, 2,900+ automated tests, and integration with six academic databases, PUBMED-SEARCH-MCP demonstrates that the gap between literature *search* and literature *understanding* can be bridged through principled algorithms and thoughtful system design.

The system is open-source and available at <https://github.com/punkpeye/pubmed-search-mcp>.

References

- Mustafa Abualsaoud, Nimesh Ghelani, Haotian Zhang, Mark D Smucker, Gordon V Cormack, and Maura R Grossman. SYMBALS: A systematic review methodology blending active learning and snowballing. *Frontiers in Research Metrics and Analytics*, 6:685591, 2021. doi: 10.3389/frma.2021.685591.

Anthropic. Model context protocol. <https://modelcontextprotocol.io>, 2024. Open standard for connecting AI assistants to external tools.

Jaime Carbonell and Jade Goldstein. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 335–336. ACM, 1998. doi: 10.1145/290941.291025.

Gordon V. Cormack, Charles L. A. Clarke, and Stefan Buettcher. Reciprocal rank fusion outperforms Condorcet and individual rank learning methods. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 758–759. ACM, 2009. doi: 10.1145/1571941.1572114.

Nicolas Fiorini, Kathi Canese, Grisha Starchenko, Evgeny Kireev, Won Kim, Vadim Miller, Maxim Osipov, Michael Kholodov, Rafis Isber, Sunil Mohan, et al. Best match: New relevance search for PubMed. *PLoS Biology*, 16(8):e2005343, 2018. doi: 10.1371/journal.pbio.2005343.

Francisco Guo et al. Artificial intelligence for literature reviews: Opportunities and challenges. *Artificial Intelligence Review*, 57, 2024. doi: 10.1007/s10462-024-10902-3.

B. Ian Hutchins, Xin Yuan, James M. Anderson, and George M. Santangelo. Relative citation ratio (RCR): A new metric that uses citation rates to measure influence at the article level. *PLoS Biology*, 14(9):e1002541, 2016. doi: 10.1371/journal.pbio.1002541.

Qiao Jin, Won Kim, Qingyu Chen, Donald C Comber, Rezarta Islamaj Dogan, Nicolas Fiorini, and Zhiyong Lu. MedCPT: Contrastive pre-trained transformers with large-scale PubMed search logs for zero-shot biomedical information retrieval. *Bioinformatics*, 39(11):btad651, 2023. doi: 10.1093/bioinformatics/btad651.

Arthur B. Kahn. Topological sorting of large networks. *Communications of the ACM*, 5(11): 558–562, 1962. doi: 10.1145/368996.369025.

Shengbo Li and Shaogang Zhu. Probabilistic latent maximal marginal relevance. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 833–834. ACM, 2010.

Jimmy Lin, Rodrigo Nogueira, and Andrew Yates. Pretrained transformers for text ranking: BERT and beyond. *Synthesis Lectures on Human Language Technologies*, 14(4):1–325, 2021. doi: 10.2200/S01123ED1V01Y202108HLT053.

Zhiyong Lu. PubMed and beyond: A survey of web tools for searching biomedical literature. *Database*, 2011:baq036, 2011. doi: 10.1093/database/baq036.

Mourad Ouzzani, Hossam Hammady, Zbys Fedorowicz, and Ahmed Elmagarmid. Rayyan—a web and mobile app for systematic reviews. *Systematic Reviews*, 5(1):210, 2016. doi: 10.1186/s13643-016-0384-4.

Matthew J Page, Joanne E McKenzie, Patrick M Bossuyt, Isabelle Boutron, Tammy C Hoffmann, Cynthia D Mulrow, Larissa Shamseer, Jennifer M Tetzlaff, Elie A Akl, Sue E Brennan, et al. The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ*, 372:n71, 2021. doi: 10.1136/bmj.n71.

Kirk Roberts, Tasmeer Alam, Steven Bedrick, Dina Demner-Fushman, Kyle Lo, William Pugh, Ellen M Voorhees, Lucy Lu Wang, and William R Hersh. TREC-COVID: Rationale and structure of an information retrieval shared task for COVID-19. *Journal of the American Medical Informatics Association*, 27(9):1431–1436, 2020. doi: 10.1093/jamia/ocaa091.

Stephen Robertson and Hugo Zaragoza. The probabilistic relevance framework: BM25 and beyond. In *Foundations and Trends in Information Retrieval*, volume 3, pages 333–389. Now Publishers, 2009. doi: 10.1561/1500000019.

Eugene Syriani, István David, and Gauransh Kumar. Title and abstract screening for literature reviews using large language models: An exploratory study in the biomedical domain. *Systematic Reviews*, 13:158, 2024. doi: 10.1186/s13643-024-02575-4.

George Tsatsaronis, Georgios Balikas, Prodromos Malakasiotis, Ioannis Partalas, Matthias Zschunke, Michael R Alvers, et al. An overview of the BioASQ large-scale biomedical semantic indexing and question answering competition. *BMC Bioinformatics*, 16:138, 2015. doi: 10.1186/s12859-015-0564-6.

Rens van de Schoot, Jonathan de Bruin, Raoul Schram, Parisa Zahedi, Jan de Boer, Felix Weijdema, Bianca Kramer, Martijn Huijts, Maarten Hoogerwerf, Gerbrich Ferdinand, et al. ASReview LAB: A tool for AI-assisted systematic reviews. *Journal of Open Source Software*, 2021.

Zhilin Wang et al. Towards the automation of systematic reviews using natural language processing, machine learning, and deep learning: A comprehensive review. *Artificial Intelligence Review*, 57, 2024. doi: 10.1007/s10462-024-10844-w.

Long Xia, Jun Xu, Yanyan Lan, Jiafeng Guo, and Xueqi Cheng. Learning maximal marginal relevance model via directly optimizing diversity evaluation measures. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 113–122. ACM, 2015. doi: 10.1145/2766462.2767710.