Scholar Publishing Group
International Journal of World Medicine
https://doi.org/10.38007/IJWM. 2026.070104.
ISSN 2789-7133 Vol. 7, Issue 1: 28-41

Design and Prototype Validation of an AI-Assisted Evidence Workflow Platform for Early-Stage Biomedical and Materials Research

Zihan Zhang

University of Manitoba, Winnipeg, Manitoba, Canada

Keywords: AI-assisted research platform; drugs and natural products; biomaterials; medical surface materials; domain schema; Evidence Ledger; multi-agent workflow

Abstract: Early-stage work in drugs and natural products, biomaterials, and medical surface materials often has a mixed and uneven literature foundation. The first task in the prototype described here is to organise the evidence, not to automatically discover it. Link the research question to the domain schema, a traceable Evidence Ledger and a staged agent workflow in the system. It can query and modify the information of candidate paper, build a reading queue, save attribute records with timestamps, flag weak claims, retain PDF-access failures as red flags, and keep track of the status of each processing attempt for follow-up investigation. The above is a Feasibility demonstration of the described workflow. It should not be taken to mean that the extraction model has already been validated on a gold-standard set.

1. Introduction

The primary challenge in early-stage research is not finding papers, but determining which parts of each paper are relevant. The other problem was knowing which part of the paper corresponded to what. Research articles in all areas address various topics such as coatings, hydrogels, scaffolds, drug-delivery systems, etc., so the Design of the experiment may also vary. A general summary can obscure the specific experimental conditions and evidence quality that determine whether a finding is applicable. Thus, the proposed system will record the literature statement as an evidence object and also keep track of its source, qualifications, confidence and review status. Antimicrobial hydrogels and functional coatings are cross-domain problems that need to be solved; the research on their preparation, biology and safety has been published in many papers.

The paper has three contributions. First, a common architecture is proposed that can be used for all drugs and natural products, biomaterials, medical surface materials, etc. An Evidence Ledger is created next to act as an audit trail and link the extracted fields with their original sources and reasons for review. Finally, it is a local prototype that can perform real-mode search, keep candidate papers, build a reading queue, generate Evidence Ledger records, map claims to evidence and save run history. The purpose of the prototype is to be a research and development tool, not the subject of study itself.

In the early days, the main questions are generally about practicality: Is the candidate worth learning more about? Is there more than one type of evidence for the route? Are there no safety or

durability data? The background information of drugs is kept in line with the cited literature; for example, one review focuses on the "applications of machine learning in drug discovery and development", and another is "artificial intelligence in drug discovery and development". The aim of the application in this paper is only to identify, screen and predict the safety of targets; it still needs to be verified through experiments and recommended by specialists in the field [1,2]. Judgments of Biomaterials are also relative. Williams thinks that biocompatibility refers to the ability of a material to "have a favourable host response", and scaffolds are assessed based on material composition, processing, structure, mechanics, degradation, cell response, etc[3,4]. Research on surface materials is also related to the design of antibacterial surfaces and the problem of implant infection; therefore, tests for antimicrobial activity, biofilm behaviour, leaching, sterilization and durability are all considered different sections in the evidence system [5,6].

Many of the available tools can only perform one stage of this process, such as searching, answering questions, reading PDFs or citing context. What is missing is the link that connects search results to the domain-specific evidence record, risk rules, manual review steps and stored run history. Therefore, this paper will be presented as an example of the system rather than a general literature search tool. The Design has kept the original language of the cited technical foundations: RAG models "combine pre-trained parametric and non-parametric memory", FAIR stands for Findable, Accessible, Interoperable and Reusable, BM25 is a probabilistic relevance framework, vector retrieval, BERT and SciBERT are used for search and scientific-text modelling, etc [7-12].

2. Research Objectives and Design Principles

A reasonable boundary for the system is that it can offer preparation assistance but will not carry out the experiment itself. For example, it may be that there is no leaching data from the coating study, or the scaffold paper only shows cell viability without degradation in the target condition. It should not be concluded that the material is ready for translation. The result is more of a research file; it shows what is known and what is unknown, which sources support the field, and where a human reviewer still needs to make a judgment.

The system will be used by many different people. A student will need a reading queue and the first project brief. The main reasons for the blockage of the claim are missing validation tests and weak evidence quality. Accordingly, the system records module status, review status, and manual-review decisions within each run log. If the same run is reopened later, another reviewer can see the original question, the selected schema, the source records and the evidence items; otherwise, they would have to start from scratch after an untraceable summary.

To build an artificial intelligence support platform for the first stage of research preparation, this paper intends to achieve this. The platform cannot guarantee that it will work; it does not replace wet-lab experiments and is not an official SOP. It will help the research group gather support materials for the study, find defects in the construction process, make periodic improvements to the project, and so on. Therefore, the prototype output should only be regarded as workflow evidence and not as the results of model performance validation.

The five operation rules are as follows: First, provide the source of the evidence; Second, adjust according to the domain; Third, be auditable in execution; Fourth, automate repetitive steps; and Fifth, have a human review for uncertain or high-risk claims. All essential areas should have a trace back to their origin and a confidence index. The three areas have the same retrieval and storage facilities but are not identical in terms of schema and rules. Agent inputs, outputs, exceptions and versions are written to the run log. Based on the above foundation, the model layer consists of a Transformer model from "Attention Is All You Need", an LLM for medicine, a chain-of-thought prompt which is a "series of intermediate reasoning steps", and ReAct's use of reasoning traces with

task-specific actions in an interleaved workflow[13-16].

3. Overall Platform Architecture

It is a short-term task area and a long-term research memory. The steps in the task run are query planning, API retrieval, ranking, parsing, extraction, review and output generation. The research memory is used to store the log of the run, candidate metadata, parsed chunks, evidence records, review decisions, output paths, etc. Therefore, one query will not be sufficient for the scope of the literature review. New search terms, uploaded papers and reviewer comments may be delayed. A fixed run_id and run folder can be used to reconstruct what happened.

The reason is to keep it neutral. It is used to retrieve and normalise metadata, eliminate duplicates, rank them, parse and extract data, but it will not decide whether a missing field is scientifically reasonable. That judgment is contained in the schema and rule layer. The missing ADMET information is not the same as missing cytotoxicity or missing leaching data. As there are no domain rules in the infrastructure, the same platform can be used for various areas of research without a drop in the demand for evidence.

The layers at the top level of the system are: Input Layer, Shared Core, Domain-Adaptation Layer, Domain Modules, Knowledge Base and Output Layer. The input layer is the research question, the uploaded paper, and the local library record. The common part is responsible for searching, de-duplication, ranking, parsing and extraction. Domain Adaptation selects the Schema and extra fields. There are rules for the Domain Module. The knowledge base stores the history and evidence of the run. The output layer will be the project brief, risk-gap table and validation matrix.

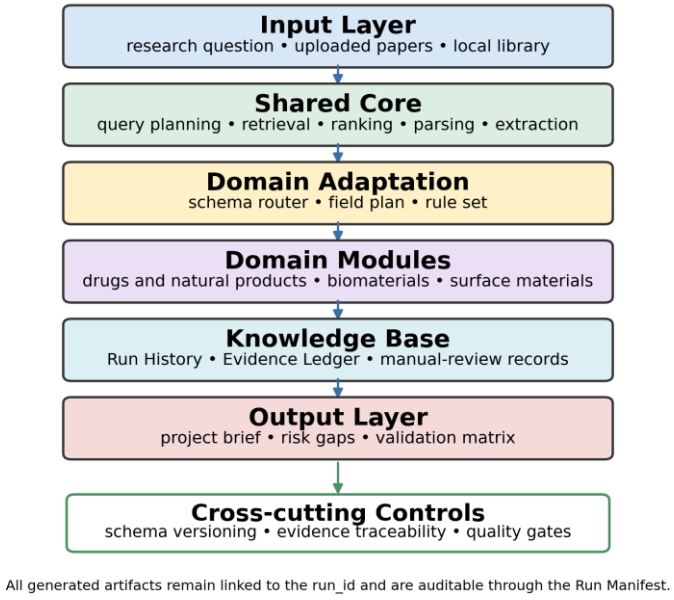

Figure 1: Layered Platform Architecture and Data Flow.

The platform organises each run around two parallel flows. The task flow moves downward: a research question enters the Input Layer, passes through the Shared Core for retrieval, deduplication, ranking, and parsing, then reaches the Domain Adaptation Layer where the Schema Router selects the appropriate domain schema, and finally enters the relevant Domain Module for structured extraction, risk review, and output generation. The knowledge flow moves upward: every run writes its candidate records, parsed chunks, Evidence Ledger entries, manual-review decisions, and Run

Manifest back to the long-term Knowledge Base, making them available for reuse in subsequent runs on related questions. The Shared Core deliberately contains no domain rules; the same retrieval and deduplication pipeline serves all three domains without modification, while domain-specific judgment is entirely confined to the Schema and rule layer above it.

Table 1: Core Platform Layers and Functions.

Layer	Main function	Persistent record
Input layer	Research question, uploaded papers, local library	Original question, file hash, source record
Shared core	Query planning, retrieval, deduplication, ranking, parsing, extraction	Candidate papers, logs, parsed text, PDF status
Domain adaptation	Schema routing, field plan, rule selection	schema_id, rule_version, extension fields
Domain modules	Drug/natural products, biomaterials, medical surface materials	Evidence Ledger, risk flags, review decisions
Knowledge base and output	Run History, project brief, risk gap, validation matrix	Run Manifest and reusable evidence records

4. Domain Modules and Risk Rules

The domain schemas are not to be rigid checklists that all the papers must meet. They set out what kind of information is required before a claim can be made more firmly. A review paper can offer the basis for knowledge; a method paper helps to prepare the route, and an experiment paper is used to demonstrate the effect or safety. In order not to make people think that the general background source directly supports the high-confidence claim, the schema divides the roles.

The logic of routes can also address the cross-domain problem. Although the wound dressing itself is a problem of biomaterials, surface-material fields also need to be addressed to solve the issues of antibacterial properties and elution. A natural-product delivery scaffold can be arranged in the form of a drug/natural-product scheme, and the release and cytocompatibility fields can be taken from the biomaterials module. It will not be a single, too-narrow category for the mixed question.

Therefore, there will be only one infrastructure for the three domain schemas of the platform. The first difference of the domains is that they have different lists of fields and also different rules for incomplete evidence. If there is a missing safety or migration attribute, it will not be included in the weight of the output claim. The rule layer will no longer be regarded as a valid conclusion but as a record of the retrieved facts.

The three domains differ not only in their field lists but in what triggers a confidence downgrade and what the Critical Review Agent does next. For drugs and natural products, a downgrade to "early signal" is triggered when evidence is limited to computational prediction or in vitro activity without ADMET or PK/PD data; the agent then requires the user to supplement toxicity, pharmacokinetic, or in vivo evidence before any development-ready claim is accepted. For biomaterials, a downgrade occurs when performance data exist without structural characterization or cell-based evaluation; the agent flags a characterization gap or biocompatibility gap and blocks the phrase "biocompatible confirmed." For medical surface materials, a downgrade occurs when only short-contact antimicrobial data are available without leaching, durability, or post-sterilization

results; the agent flags an application mismatch and blocks device-ready or non-leaching conclusions. These domain-specific rules are stored as versioned configuration files and referenced by rule_version in each Run Manifest.

Table 2: Condensed-Domain Schemas and Risk Control.

Domain	Core fields	Main downgrading rule
Drugs and natural products	disease/target; candidate modality; efficacy; ADMET; PK/PD; toxicity	Computational or in vitro evidence alone is only an early signal; missing ADMET/PK-PD blocks development-ready claims.
Biomaterials	composition; preparation; characterization; mechanics/degradation; cytotoxicity; application scenario	Performance without characterization or biological evaluation cannot support biocompatibility claims.
Medical surface materials	substrate; modification; antimicrobial assay; organism; leaching/migration; durability; sterilization	Short-term antimicrobial data alone cannot support device-ready or non-leaching claims.

Schema Router decision logic. The Schema Router receives the Intent Profile from the Question Understanding Agent and performs domain matching in three steps. First, if the domain_hint field explicitly specifies drug, biomaterial, or surface_material, the corresponding schema is selected directly. Second, if domain_hint is absent or the semantic matching scores for two or more domains differ by less than 0.10, the Router outputs a ranked list of candidate schemas and pauses execution until the user confirms the primary domain. Third, for cross-domain questions — such as an antimicrobial hydrogel wound dressing that spans both biomaterials and surface materials — the Router applies a composite strategy: one domain is designated as the primary schema, and selected fields from a secondary schema are appended as extension fields. The chosen schema_id, schema_version, any extension_fields, and the user-confirmation status are all written to the Run Manifest to ensure reproducibility.

The search words are varied in different modules. For drugs and natural products, the query plan is a combination of disease or target terms and candidate, efficacy and safety terms. Natural products are managed by an extension that records species, extracts, active compounds, isolation, quality control, etc. The four sections of the search for biomaterials are composition, preparation, characterisation and application. For the medical surface materials module, functional query terms — such as antimicrobial, anti-biofilm, and anti-fouling — are combined in parallel with risk terms including leaching, migration, cytotoxicity, durability, and sterilization. These two query groups run separately rather than being merged into a single long string, so that a paper reporting only antimicrobial activity without any leaching or durability data will still be retrieved and subsequently flagged by the risk rules rather than excluded at the retrieval stage.

5. Model Components and Multi-Agent Workflow

There is a quality gate in the Agent workflow. First, the agent determines the aim and sets a search plan for the question. Retrieval and Parsing Agents obtain the metadata, abstracts, legally distributed PDFs and parsed chunks. Evidence-control agents organise the source material in an orderly way and determine whether this record is founded. The Output and Memory agents write the accepted or review-pending records to the run folder and Run History. As shown in the Design, an API failure, a low-recall search, a missing PDF and a blocked claim are all different states.

The first is that, unlike chat-style summarisation, the agents are not allowed to fill in the blanks

quietly. If there is no PDF, the paper will not be deleted; instead, its DOI and source page will be saved. If only an abstract is available, the record will be marked as low confidence. If an evidence sentence cannot be found in the parsed text, it will be downgraded and added to the list for manual review. Therefore, the purpose of the agent system is not to eliminate the researcher but to make the process of partial automation safer and simpler to monitor.

The parts of the model are introduced based on their functions, not by who supplies them. The retrieval layer can use PubMed, OpenAlex, CrossRef, Europe PMC, local libraries, BM25, etc. Embeddings and Vector Indexes can be used in the Semantic Layer. The Ranking Layer Selects the Papers to be Read. A deep-reading layer can extract organised fields from text, tables and figures if they are present in the source. The review layer is based on model checking and rules, and the output layer produces Word, Excel, brief and matrix files.

In one run, the Question Understanding Agent builds the Intent Profile, the Schema Router selects the main schema, and then the Query Planner turns that schema into search strings. The Retrieval Agent fetches the list of candidate documents, the Ranking Agent orders them into Reading Buckets, and then the Parsing & Evidence Extraction Agents create Parse Chunks and Evidence Records. The main reporter will use either a block or a demotion. The Report Generation Agent writes the output and updates Run History.

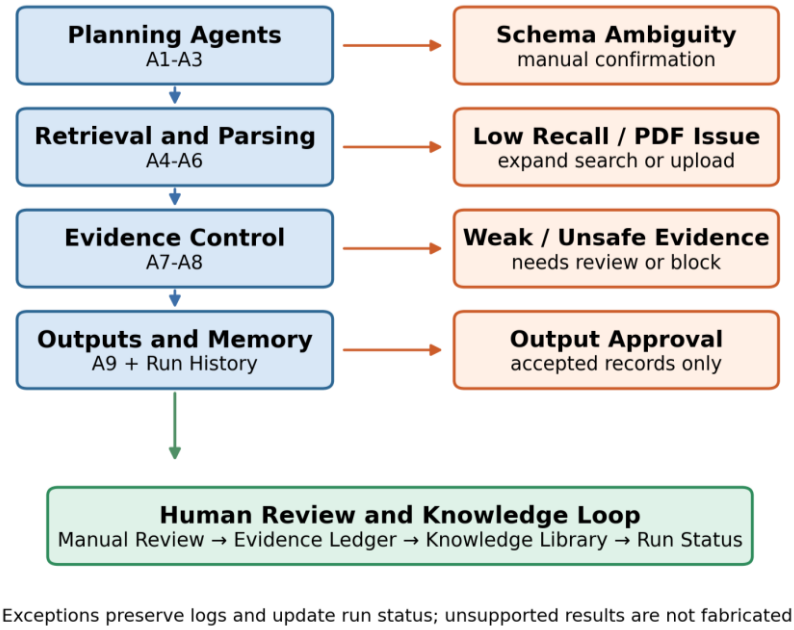

Figure 2. Agent Workflow, Exception Branches and Manual Review Mechanism

Expanded search is limited to a maximum of two to three iterations. If the number of core-relevant papers remains below the threshold after the final iteration, the workflow exits with a Literature Insufficient status, writes the current candidate set to the run folder for manual review, and does not proceed to extraction. This prevents indefinite looping and makes the retrieval boundary explicit in the run record.

Automation is conditional. If there are problems with the schema, no results will be returned or there will be no grounds for evidence collection; if there is a model output conflict, it will also be stopped or rerouted for manual review through a block rule. This is an intentional safety measure; if there is no data, then no answer should be returned, not an imputed one.

Table 3. Agent Workflow and Exception Handling.

Stage	Normal output	Exception branch
Question understanding and routing	Intent Profile and selected schema	Ambiguous schema -> manual confirmation
Search planning and retrieval	Search plan and candidate records	Low recall -> expanded search; zero records -> Literature Insufficient
Ranking, parsing, and PDF access	Ranked bucket, parsed chunks, PDF status	PDF unavailable -> DOI fallback and manual upload
Evidence extraction	Evidence Ledger with qualifiers	Low confidence or ungrounded evidence -> Manual Review
Critical review and output	Claim map, risk review, outputs, Run History	Unsupported or unsafe claim -> blocked or downgraded

6. Evidence Ledger and Quality Control

Table 4: Evidence Ledger and Quality Control Fields.

Component	Required fields	Quality rule
Evidence Record	field_name; value; evidence_text; DOI/PMID/source_id; page_or_section; confidence; risk_flag; review_status	Evidence without traceable source is downgraded and sent to Manual Review.
Confidence scoring	source completeness; evidence level; field completeness; model agreement; rule status	Scores use High=2, Medium=1, Low=0; any critical Low or blocking rule forces Low confidence.
Span grounding	evidence_span_id; grounding_status; match_score; source_locator	Exact or normalized fuzzy matching is required; unverified evidence cannot enter final conclusions.
Condition binding	Unit; method; organism/model; contact_time/dose; control group; sterilization or durability status	Measurements are compared only under recorded conditions.
Numerical evidence	Value_source_type; unit_check_status; numeric_range_check	Table_cell and figure_derived evidence require unit checks; figure-derived data default to manual review.

Quality of materials and surface-material work is often expressed in terms of adjectives. A statement that there will be a 99% reduction in bacteria is not sufficient unless we know the type of bacteria, the time of contact, the assay method, substrate, coating conditions, sterilization status, etc. The Evidence Ledger saves the above qualifiers along with the extracted value. It is easier to see when two papers are similar in form but have different experimental environments.

Grounding is needed, but it is not the same as scientific acceptance. A grounded record is a piece of text that can be linked back to a specific part, table cell or abstract entry. It does not confirm that the interpretation is right or that the study design is good. Therefore, confidence is the combination of source completeness, level of evidence, field completeness, model agreement and rule status. A record can be rooted, and it will still need to be reviewed if a key qualifier is missing or if the claim is stronger than the evidence.

The first form of audit is the Evidence Ledger. Each record stores the field name, extracted value, evidence text, source identifier, page or section, confidence, risk flag, review status and model version. Text-derived, table-cell-derived and figure-derived values are not all the same. When the source locator of evidence cannot be verified reliably, it will be excluded from the final conclusion and sent for manual review.

Confidence aggregation. The five dimensions in Table 4 are scored numerically: High = 2, Medium = 1, Low = 0. An aggregate score is computed as Σ(dimension scores) / 10. Records with a score ≥ 0.85 and no triggered blocking rule are assigned High confidence; scores between 0.60 and 0.85 with no blocking rule receive Medium; scores below 0.60, any single critical dimension rated Low, or any triggered blocking rule force the overall confidence to Low and automatically route the record to the Manual Review Queue. This hard-floor rule ensures that a record with strong source completeness but missing ADMET data, for example, cannot be promoted to High confidence by averaging.

Each Evidence Record corresponds to one extraction instance of a single schema field from a single paper (a field × paper unit). For the N-halamine prototype run, 30 candidate papers processed through the surface_material_schema generated 241 records, averaging approximately eight records per paper, which reflects the number of schema fields with extractable values in each source.

Antimicrobial activity, cytotoxicity, degradation, release and leaching do not have values on their own. Therefore, the platform attaches corresponding units, ways of doing so, organisms or models, contact periods or doses, control groups, states of samples, sterilisation or durability conditions, etc., to measurements. It will not be a context-free comparison of numbers from different experiments.

7. Data Architecture and Outputs

The research group has been using the Library for a long time to study the targets and coating strategies of the research. If there is no run-level storage, a necessary query, source decision or reviewer comment will be lost after a short time has passed during export. Therefore, the data architecture will contain the search plan, retrieval log, candidate, parsed chunk, evidence ledger, review queue, output file, etc. Later on, one will know whether it is based on the full text, abstract or other reasons for the decision.

Table 5. Data Objects and Outputs.

Object	Purpose	Update strategy
Run Manifest	Records question, parameters, schema/rule versions, prompt_version, temperature, seed, outputs, and errors	Generated per run and never overwritten.
Paper Metadata	Stores DOI, title, year, journal, abstract, source, and PDF status	New DOI records are added; duplicates are merged.
Evidence Ledger	Stores field-level evidence, confidence, risk, and review status	Versioned and manually revisable.
Manual Review Queue	Stores reviewer decision, reviewer_id, timestamp, and rationale	Review decisions write back to the Evidence Ledger.
Outputs	Core paper list, risk gap table, route comparison, validation plan, project brief	Generated only from reviewed or clearly labelled records.

Deduplication is also in the research records. First check the DOI; if not found, then use title +

first author + year to look up information for that record. If a paper can be obtained from more than one database, the source tags will be combined. Thus, it will not increase the number of candidates and will be easier to understand.

Each run will create a new folder, and inside, you can find the original question, search plan, candidate records, parsed chunks, Evidence Ledger, risk flags, manual-review queue and final output. The library of literature has compiled reusable metadata and supports PDFs or snippets; it can parse text, figures and tables, and also extract earlier extraction results based on DOI or file hash. If the identifier supports it, preprints can be linked to the official publication.

8. Prototype Implementation and Run Data

The prototype run in this paper should be considered a workflow test. It can be seen that the system has started a real-mode run, saved candidate papers, created a reading queue, generated Evidence Ledger records, mapped claims to evidence, and stored the run in Run History. It has the problem that access to PDFs is limited by open-access conditions, publisher permissions and other reasons. If the local PDF cannot be obtained, only its metadata and the access link will be saved in the system; a fake file will not be created.

Table 6. Main Prototype Run and Module Status.

Item	Value / status	Interpretation boundary
Research question	How can N-halamine polymer coatings achieve rechargeable antimicrobial activity with reduced leaching risk?	Used as one Real Mode prototype case.
Schema and sources	surface_material_schema; OpenAlex and PubMed	Demonstrates external retrieval entry points.
Candidate / ranked / reading queue	30 / 30 / 30	Shows candidate preservation and reading-queue generation, not validated recall.
Evidence / grounded records	241 / 241	Shows field-level evidence creation and grounding to parsed sources.
Manual review / blocked claims	241 / 0	All evidence remained review-gated; no final scientific claim is automatic.
PDF available / missing	0 / 3	Shows DOI/source fallback rather than full-text availability.
Final status	needs_review	Prototype output is a feasibility demonstration.

The same run also shows that manual review is still necessary. It produced 241 evidence records; although they were marked as grounded, all 241 were still in the manual-review category. It is not a problem with the ground. It indicates that the two are not the same decision. There may be a source span, and the qualifier, claim wording or application boundary still need to be determined by a person.

A local front-end and back-end system has been constructed according to the prototype at present. The Front-end is the AI Agent R&D Workflow Control Console. The backend is connected to the literature sources and model services. Real Mode is for external search and real run record. Mock Mode is for the purpose of interface testing and should not be considered scientific evidence.

Automatic PDF download is not to be done, or the paper will be useless. If the PDF returns a 403 error, only provides a publisher's landing page, or cannot be downloaded legally, then the platform will save the DOI, source URL, open-access status, abstract and failure reason. If a person is authorised to view the entire text legally, then manual PDF upload will be conducted.

The main case does not have a human gold standard yet, there is no manual audit of the representative evidence records, no complete raw-to-ranked retrieval logs, and no formal baseline comparison. Therefore, it should be used to demonstrate that the workflow can run and maintain auditable objects, rather than to show that recall, precision or extraction accuracy have already been realised.

9. Governance, Compliance, and Validation Plan

Governance should be added to the workflow, and the system will save public metadata, licensed full-text, parsed snippets, internal notes and review decisions. They are of different kinds of data. Public metadata can be spread more freely than licensed publisher text or unpublished group notes. Therefore, the system should have data classification, append-only logs, reviewer identity fields and license-aware PDF handling.

The Validation Plan must be properly prepared. 30 candidate papers or 241 evidence records are typical counts of objects in the pipeline. They do not show that the objects are correct. Correctness needs human annotation and evidence, manual inspection of typical cases, comparison with a manual workflow, etc., of RAG and other research-assistance tools. Retrieval and extraction should be followed by a lack of evidence; it should also not block legitimate claims, be low-burden to review, and have a good origin.

As the knowledge base stores retrieval records, parsed snippets, Evidence Ledger entries, manual-review decisions and outputs, it should separate public literature metadata, licensed full text, internal notes and unpublished experimental information. Do not skip the paywalled literature. If there is no legal basis or institutional licence, only store the metadata and allowed snippets in the system; do not store the full text.

Table 7. Governance, Standards and Validation Design.

Area	Design choice	Reason
Data governance	Separate public metadata, licensed full text, internal notes, and unpublished data	Protects intellectual property and preserves auditability.
Copyright boundary	No bypassing of publisher permissions; DOI/source fallback and manual upload only within legal access	Prevents treating download failure as evidence failure while respecting access rights.
Regulatory anchors	ICH M3(R2), ICH S series, OECD GLP, 21 CFR Part 58, ISO 10993, ISO 14971, FDA guidance, ASTM degradation tests	Links risk rules to recognized external standards.
Validation design	Compare manual workflow, bare RAG, existing tools, and the proposed platform	Evaluates recall, extraction, grounding, risk-gap identification, cost, and latency.
Gold standard	5-10 papers per domain annotated by human reviewers	Needed before claiming precision, recall, F1, or kappa.

10. Discussion

It is a general AI literature assistant and a particular drug discovery system. It is not a way to produce molecules or carry out virtual screening. It starts earlier; before a research group selects what to validate, it organises all sorts of evidence. The main value is the combination of domain schemas, Evidence Ledger, risk rules, manual review and Run History, especially when a question spans the boundaries of drugs, biomaterials and device surfaces.

It is currently only local; external APIs need to be obtained, there is open-access coverage, model-provider restrictions, and manual PDF uploads. The quality of the output will only increase with the accumulation of more retrieval logs, audited evidence samples and reviewer decisions. Therefore, the next stage will be controlled validation and not full automation.

The comparison should be with actual tools, not just with a simplified bare-RAG baseline. I have put the tool descriptions next to their original references: Elicit is only to be used for systematic reviews; Consensus is an AI research assistant; Scite is a citation-context index that categorises reasons for citation; and SciSpace is a tool to extract and read PDF data[17-20]. The new platform is different from the previous one in that it has three domain schemas, condition-bound Evidence Records, hard risk downgrading, cross-context extrapolation control, review queues and Run History write-back.

Table 8. Differentiated Positioning from the Background.

Tool / baseline	Relevant capability	Differentiation of this platform
Elicit	Systematic review support, screening, and data extraction	Adds domain schemas, risk downgrading, and long-term Evidence Ledger write-back.
Consensus	AI search and evidence analysis for peer-reviewed literature	Transforms evidence into validation needs and risk-gap records.
Scite	Citation context and citation-intent classification	Focuses on original experimental fields, spans, qualifiers, and blocking rules.
SciSpace	PDF extraction, table/statistics recognition, Q&A, and export	Adds domain-specific review rules and run-level auditability.
Bare RAG	Retrieval-enhanced answer generation	Treated as a component rather than a sufficient research workflow.

There are also bounds. The platform shall not have the power to conduct tests for the suitability of a certain person, develop its own regular operating rules independently, or act as an administrative supervision department. The output is the preparation materials that a group of people need for an investigation and planning.

11. Conclusion

A platform for AI-assisted evidence workflow in the first stage of work for drugs and natural products, biomaterials and medical surface materials is proposed in this paper. The Design is composed of domain schemas, an Evidence Ledger, agent orchestration, risk-review rules, manual review and Run History. The purpose is to organize the scattered literature evidence and review decisions of the research preparation object in an organized manner.

It is Architecture, Evidence and Practicality. The two parts of the architecture are separate from one another. At the level of evidence, the extracted statement should keep the source, qualifier,

confidence and review status. Practically speaking, search, ranking, parsing, extraction, critical review, output generation and knowledge write-back can all be reopened and audited in a process.

Based on the above workflow, the prototype shows that it is possible to realise the workflow in a local real mode with retrieval entry points, candidate preservation, reading-queue generation, evidence ledger records, claim-to-evidence mapping, manual-review routing, PDF fallback and run history. The above are the reasons. They have not been able to make a reliable claim about the accuracy of extraction because there is no human gold standard, systematic evidence audit or head-to-head baseline test.

Table 9. Summary of the Main Content and Support Measures.

Contribution Area	Platform Mechanism	Research Value
Multi-domain architecture	Shared core plus domain schemas for drugs/natural products, biomaterials, and medical surface materials	Supports related but different research directions without reducing them to one generic workflow.
Evidence traceability	Evidence Ledger with source, DOI, qualifiers, confidence, grounding, and review status	Makes extracted evidence auditable and reduces unsupported summarization.
Risk-aware reasoning	Hard downgrade rules, blocked claims, and manual-review routing	Prevents weak evidence from being written as strong scientific conclusions.
Research memory	Run Manifest, Run History, Knowledge Library, and review write-back	Turns one-time searches into reusable research-group knowledge.

Table 10. Practical Research Scenarios and Expected Outputs.

Use Scenario	Typical User Question	Expected Platform Output
Topic scoping	What is known about a candidate material, coating, or natural product?	Core paper list, reading queue, and brief evidence summary.
Route comparison	Which material or surface strategy is better supported by evidence?	Route comparison matrix with methods, performance, limitations, and risks.
Risk review	Which claims are unsafe or under-supported?	Claim-to-evidence map, blocked claims, and missing-evidence list.
Validation planning	What should be tested next before stronger claims are made?	Validation gap matrix and literature-derived planning inputs.
Project continuity	What did previous searches find and how were they reviewed?	Run History, Run Manifest, Evidence Ledger versions, and manual-review decisions.

The main function of the platform in terms of research is to provide a place for preparation and revision. It can show which papers are worth paying attention to, which areas have traceable support, which claims should be reduced, and which validation tests are missing. Finally, all the scientific claims need to be approved by experts and verified through experiments.

Table 11. Remaining Limitations and Future Work.

Current Limitation	Why It Matters	Planned Improvement
Incomplete gold-standard validation	Prototype counts do not prove extraction accuracy or retrieval recall.	Annotate 5-10 representative papers per domain and calculate precision, recall, F1, and grounding rate.
PDF access depends on OA and permissions	Unavailable PDFs limit full-text extraction and page-level grounding.	Keep DOI/source fallback, support legal manual upload, and record download failure reasons.
Limited figure/table numerical grounding	Some important results appear only in tables, spectra, or plots.	Add table-cell binding, unit checking, and manual review for figure-derived values.
Local prototype deployment	Performance, persistence, and multi-user governance require further testing.	Move toward stable database storage, role-based access control, and audit logging.
No formal baseline comparison yet	The added value over existing tools must be measured, not only argued.	Compare against manual review, bare RAG, and tools such as Elicit, Consensus, Scite, and SciSpace.

These are also the situations where they can be used. The platform will provide the study material; it will not be the final protocol or a regulatory filing. It can show the foundation of the evidence before a group of people has to invest their time and money in the next round of research.

First, in the future, I will perform validation work. A small but representative evaluation set needs to be set up before the platform can be regarded as accurate, efficient and generalisable. The corresponding evidence files are to be manually verified, and the process will be compared with that of manual inspection, bare RAG, etc.

The above system is usually referred to as a research-memory and evidence-workflow platform. It will not replace scientists; rather, it will reduce the tedious process of collecting evidence and increase the transparency of uncertainty at the beginning of planning. After confirmation of the extraction and strengthening of governance and attention to copyright management, it can be used as actual support for research groups to work together on projects related to drugs, biomaterials and medical surface materials.

References

[1] Vamathevan J, Clark D, Czodrowski P, et al. Applications of machine learning in drug discovery and development. Nature Reviews Drug Discovery. 2019;18(6):463-477.

[2] Paul D, Sanap G, Shenoy S, et al. Artificial intelligence in drug discovery and development. Drug Discovery Today. 2021;26(1):80-93.

[3] Williams DF. On the mechanisms of biocompatibility. Biomaterials. 2008;29(20):2941-2953.

[4] Place ES, George JH, Williams CK, et al. Synthetic polymer scaffolds for tissue engineering. Chemical Society Reviews. 2009;38(4):1139-1151.

[5] Li W, Thian ES, Wang M, et al. Surface design for antibacterial materials: From fundamentals to advanced strategies. Advanced Science. 2021;8(19):2100368.

[6] Arciola CR, Campoccia D, Montanaro L. Implant infections: Adhesion, biofilm formation and immune evasion. Nature Reviews Microbiology. 2018;16(7):397-409.

[7] Lewis P, Perez E, Piktus A, et al. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In: Advances in Neural Information Processing Systems. 2020;33:9459-9474.
[8] Wilkinson MD, Dumontier M, Aalbersberg IJJ, et al. The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data. 2016;3:160018.
[9] Robertson S, Zaragoza H. The Probabilistic Relevance Framework: BM25 and Beyond. Foundations and Trends in Information Retrieval. 2009;3(4):333-389.
[10] Johnson J, Douze M, Jegou H. Billion-scale similarity search with GPUs. IEEE Transactions on Big Data. 2019;7(3):535-547.
[11] Devlin J, Chang MW, Lee K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of NAACL-HLT. 2019:4171-4186.
[12] Beltagy I, Lo K, Cohan A. SciBERT: A pretrained language model for scientific text. In: Proceedings of EMNLP-IJCNLP. 2019:3615-3620.
[13] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. In: Advances in Neural Information Processing Systems. 2017;30:5998-6008.
[14] Thirunavukarasu AJ, Ting DSJ, Elangovan K, et al. Large language models in medicine. Nature Medicine. 2023;29:1930-1940.
[15] Wei J, Wang X, Schuurmans D, et al. Chain-of-thought prompting elicits reasoning in large language models. In: Advances in Neural Information Processing Systems. 2022;35:24824-24837.
[16] Yao S, Zhao J, Yu D, et al. ReAct: Synergizing reasoning and acting in language models. In: The Eleventh International Conference on Learning Representations (ICLR). 2023.
[17] Elicit. Systematic literature reviews: AI for evidence synthesis. 2026 [cited 2026-06-15]. Available from: https://elicit.com/solutions/systematic-review.
[18] Consensus. Consensus: AI for research. 2026 [cited 2026-06-15]. Available from: https://consensus.app/.
[19] Nicholson JM, Mordaunt M, Lopez P, et al. scite: A smart citation index that displays the context of citations and classifies their intent using deep learning. Quantitative Science Studies. 2021;2(3):882-898.
[20] SciSpace. AI-powered data extraction from research PDFs. 2026 [cited 2026-06-15]. Available from: https://scispace.com/extract-data.