Retrieval-Augmented Phytopharmacology: Connecting Literature Evidence, Molecular Databases, Pathway Knowledge, and Expert Review for Drug Discovery

Noah Williams[1]*, Grace Collins[1], Ethan Brooks[2]

[1]Department of Artificial Intelligence for CYP450 Metabolism of Phytochemicals, Faculty of Pharmacy, University of Sydney, Sydney, Australia.
[2]Department of Cheminformatics for Blood-Brain Barrier Penetration, Faculty of Pharmacy, University of Queensland, Brisbane, Australia.

ABSTRACT

Phytopharmacology depends on heterogeneous and frequently fragmented evidence spanning botanical identity, phytochemical composition, chemical structures, bioactivity measurements, molecular targets, biological pathways, safety findings, and expert interpretation. Although retrieval-augmented approaches may improve access to this evidence, retrieval alone cannot establish molecular identity, biological mechanism, therapeutic relevance, safety, or candidate validity. This article proposes a conceptual retrieval-augmented phytopharmacology framework for evidence-grounded drug discovery reasoning. The framework begins with explicit research-question formulation and retrieves relevant peer-reviewed literature, natural-product and phytochemical records, molecular structures, bioactivity observations, target annotations, pathway knowledge, and ADMET or toxicity evidence. Retrieved information is then subjected to source grounding, citation traceability, database-provenance assessment, identity resolution, duplicate detection, conflict identification, and evidence-level classification. Pathway mappings and knowledge relationships are interpreted as contextual support for mechanistic hypotheses rather than proof of mechanism. Natural-product chemists, pharmacologists, toxicologists, and other relevant experts remain responsible for evaluating plausibility, resolving conflicts, identifying evidence gaps, and defining appropriate validation studies. Drug discovery decision logic is therefore limited to transparent hypothesis prioritisation, uncertainty communication, validation planning, and auditable research support. The principal contribution is an integrated framework that connects retrieval, molecular database integration, pathway interpretation, expert review, and decision boundaries while explicitly controlling hallucination, provenance loss, evidence overstatement, and unsupported clinical or candidate-level claims.

Key Words: *Retrieval-augmented generation, Phytopharmacology, Natural products, Molecular databases, Pathway knowledge, Expert review*

eIJPPR 2026; 16(2):70-84

HOW TO CITE THIS ARTICLE: Williams N, Collins G, Brooks E. Retrieval-Augmented Phytopharmacology: Connecting Literature Evidence, Molecular Databases, Pathway Knowledge, and Expert Review for Drug Discovery. Int J Pharm Phytopharmacol Res. 2026;16(2):70-84. https://doi.org/10.51847/tXL4jnck8i

INTRODUCTION

Phytopharmacology is an evidence-intensive field in which conclusions may depend on the alignment of botanical taxonomy, plant-part provenance, extraction conditions, phytochemical identity, structural characterization, assay design, molecular targets, biological pathways, exposure conditions, toxicity findings, and disease context. These evidence types are distributed across primary studies, reviews, molecular resources, pathway databases,

Corresponding author: Noah Williams
Address: Department of Artificial Intelligence for CYP450 Metabolism of Phytochemicals, Faculty of Pharmacy, University of Sydney, Sydney, Australia.
E-mail: ✉ noah.williams@sydney.edu.au
Received: 19 January 2026; **Revised:** 29 March 2026; **Accepted:** 02 April 2026

toxicology collections, and expert-curated knowledge sources, making comprehensive retrieval and interpretation difficult. Retrieval-augmented methods may improve access to external biomedical evidence and reduce dependence on information encoded only within a language model, but their reliability remains contingent on retrieval design, corpus quality, evaluation procedures, and safeguards against unsupported generation [1]. A phytopharmacology workflow must therefore treat retrieval as an evidence-access mechanism rather than as independent verification of chemical or biological truth.

Retrieval-augmented generation commonly combines source indexing, query-based retrieval, optional ranking or reranking, context augmentation, and response generation. In a phytopharmacological setting, each stage introduces domain-specific requirements: queries must distinguish compounds from extracts, synonyms must be reconciled, botanical sources must be normalized, database versions must be recorded, and retrieved passages must remain linked to their original sources. Healthcare-oriented retrieval-augmented systems illustrate how external knowledge can improve grounding while leaving unresolved risks involving inaccurate retrieval, biased corpora, incomplete context, privacy, transparency, and inappropriate downstream use [2]. Consequently, a retrieval-augmented phytopharmacology framework must evaluate not only whether an answer appears plausible, but also whether its supporting records are relevant, traceable, current, internally consistent, and appropriately interpreted.

The drug-discovery implications of retrieval augmentation require particular caution. Artificial intelligence can organize chemical and biological information, identify patterns, and assist hypothesis generation, yet apparent computational performance may have limited translational value when tasks are poorly defined, datasets are unrepresentative, labels are unreliable, or biological validation is absent [3]. In phytopharmacology, these risks are compounded by mixtures, stereochemical ambiguity, taxonomic inconsistencies, heterogeneous assay systems, concentration-dependent effects, uncertain bioavailability, and the frequent reuse of database records derived from the same underlying publication. A retrieved association between a phytochemical and a target, pathway, phenotype, or disease therefore cannot be converted automatically into a mechanistic conclusion or candidate recommendation.

Natural products remain important sources of chemical diversity and pharmacological hypotheses, but their contribution to drug discovery depends on rigorous characterization, reproducible biological evaluation, safety assessment, and staged translational investigation [4]. Retrieval augmentation may assist this process by connecting literature evidence, molecular records, pathway knowledge, and expert review within a traceable reasoning workflow. The purpose of this article is to propose a retrieval-augmented phytopharmacology framework that integrates source-grounded evidence retrieval, molecular database provenance, identity resolution, pathway-context mapping, evidence triangulation, multidisciplinary review, uncertainty management, decision boundaries, and validation planning. The proposed framework is conceptual and is intended to support research reasoning and hypothesis prioritisation rather than autonomous candidate selection, mechanistic confirmation, clinical decision-making, or replacement of professional expertise.

Retrieval-augmented phytopharmacology

Retrieval-augmented phytopharmacology can be defined as a source-grounded research approach in which phytochemical, pharmacological, molecular, pathway, safety, and contextual evidence is retrieved from explicitly defined sources and organized to support expert-led drug discovery reasoning. It differs from generic retrieval-augmented generation because the retrieval target is not merely textually relevant information; it is evidence that must be chemically, biologically, and methodologically compatible with the question. It also differs from unsourced chatbot reasoning, which may produce fluent explanations without traceable support, and from static database querying, which retrieves records without necessarily reconciling their meanings or evidentiary status. Retrieval-and-reflection approaches indicate that selective use of external information and subsequent checking may improve evidence-oriented reasoning, although success remains dependent on retrieval relevance, task design, and generalizability [5]. Within the proposed framework, retrieval-augmented reasoning therefore involves query refinement, evidence categorization, contradiction detection, and explicit acknowledgement of unresolved questions.

Source-corpus definition is central to this distinction. A phytopharmacology corpus may include peer-reviewed primary studies, review literature where methodologically appropriate, curated natural-product resources, phytochemical collections, bioactivity and target databases, pathway resources, ADMET and toxicity sources, supported pharmacovigilance records, and knowledge graphs. Inclusion in the corpus does not imply equal evidentiary value. Each source must be classified according to its purpose, curation process, version, coverage, update history, underlying evidence, and limitations. Guideline-grounded retrieval studies demonstrate the importance of restricting retrieval to an identifiable evidence corpus and evaluating outputs with domain experts rather than assuming that retrieved context

guarantees accurate interpretation [6]. The same principle applies to phytopharmacology: a response should remain bounded by the evidence actually retrieved, and any extrapolation should be labelled as interpretation or hypothesis.

Retrieval quality also depends on how domain entities are represented. Queries based only on surface text may fail when a phytochemical has multiple synonyms, when a botanical species has historical taxonomic names, when a target has several identifiers or isoforms, or when a database record describes an extract while another describes an isolated constituent. Entity-aware retrieval methods show that structured recognition of clinically meaningful concepts can improve context selection compared with undifferentiated retrieval strategies [7]. In the proposed framework, this principle is extended to phytochemical names, molecular identifiers, botanical taxa, plant parts, target accessions, disease terms, pathway identifiers, assay types, toxicity endpoints, and evidence categories. This adaptation is a proposed design element and requires specialized vocabularies, synonym mapping, structure-based matching, and review of unresolved entity conflicts.

Retrieval-augmented phytopharmacology must ultimately remain a human-supervised research process. Source-grounded generation may organize evidence into a coherent answer, but it cannot independently determine whether a compound was correctly identified, whether a measured concentration is pharmacologically relevant, whether a pathway interpretation is causal, or whether a safety signal is clinically meaningful. Professional decision-support frameworks based on retrieval have been designed to supplement specialist workflows rather than replace the responsible expert [8]. Accordingly, the proposed framework places citation traceability, uncertainty communication, expert correction, and no-clinical-claim boundaries alongside retrieval and generation. Its intended outputs are evidence maps, qualified hypotheses, conflict summaries, research priorities, and validation plans—not validated mechanisms, confirmed safety conclusions, autonomous candidate rankings, or therapeutic recommendations.

Literature evidence and molecular databases

Peer-reviewed literature provides the principal context for determining what was investigated, how an experiment was conducted, which material was tested, what outcome was measured, and which limitations were reported. Effective literature retrieval should therefore operate below the article-title level and connect specific passages with normalized biomedical entities, study designs, experimental systems, interventions, assays, and outcomes. Automated biomedical concept annotation can link full-text passages to standardized chemical, gene, disease, species, and mutation concepts, thereby supporting entity-oriented search and evidence navigation [9]. Nevertheless, automated annotations must be verified because specialized natural-product terminology, uncommon compound names, ambiguous abbreviations, and plant-specific entities may be missed or incorrectly normalized. Review literature can assist orientation and source discovery, but compound-specific claims should be traced back to the most relevant primary evidence whenever possible.

Semantic indexing can extend this process by identifying biomedical entities and relations that are not captured reliably through exact keyword matching. Domain-adapted language representations have demonstrated utility in biomedical named-entity recognition, relation extraction, and question-answering tasks [10]. In retrieval-augmented phytopharmacology, comparable text-mining functions may help identify compound–target statements, plant–constituent relationships, assay outcomes, pathway references, safety observations, and disease-context associations. Such extraction should not be treated as evidence confirmation. Every extracted relationship must retain its source passage, DOI, study type, experimental model, evidence category, and interpretive status so that a reviewer can determine whether the statement represents a measured finding, a prediction, a review-level summary, a traditional-use report, or an author hypothesis.

Molecular databases provide complementary structure and identity information, but they differ substantially in coverage, accessibility, curation, standardization, biological annotation, and provenance documentation [11]. Cross-database retrieval may reveal multiple records for the same natural product, inconsistent synonyms, unresolved stereochemistry, alternative structures, incompatible taxonomic sources, or biological annotations inherited from different evidence types. Aggregated natural-product resources illustrate the value of bringing distributed structures into a common collection while also demonstrating the need for normalization, duplicate handling, and preservation of source-database provenance [12]. The proposed framework therefore separates record retrieval from identity resolution. A database hit becomes a candidate identity record that must be reconciled using molecular structure, stereochemical information, identifiers, formula, mass, synonym history, botanical source, analytical characterization, and supporting literature.

Bioactivity, target, pathway, ADMET, toxicity, pharmacovigilance, and knowledge-graph sources add further layers of interpretation. Natural-product activity databases can connect compounds with organism sources, measured biological activities, assays, and molecular

targets, but heterogeneity in endpoint definitions, units, experimental conditions, and evidence depth must be retained rather than collapsed into a single activity claim [13]. Curated phytochemical resources can connect medicinal plants, reported constituents, and therapeutic or traditional-use contexts, yet these categories must remain distinct because botanical occurrence or historical use does not establish pharmacological efficacy [14]. The proposed framework consequently requires source provenance, database versioning, evidence-type labels, measured-versus-predicted status, duplicate-source detection, conflict preservation, and expert adjudication before information is used in downstream reasoning. **Table 1** summarises evidence sources and molecular database components required for retrieval-augmented phytopharmacology.

Table 1. Literature Evidence and Molecular Database Components for Retrieval-Augmented Phytopharmacology: Source Grounding, Phytochemical Identity, Chemical Structure Records, Bioactivity Evidence, Target Annotations, Pathway Data, ADMET and Toxicity Evidence, Provenance, and Validation Needs

Evidence source	Core retrieval question	Required metadata	Potential retrieved output	Quality or provenance concern	Validation requirement	Failure risk	Framework role
Peer-reviewed primary literature	What direct experimental evidence addresses the compound, extract, botanical material, target, phenotype, pathway, or safety question?	DOI; full citation; article type; material identity; species; plant part; extraction or isolation method; experimental system; assay; comparator; outcome; limitations	Source passages, study descriptors, measured findings, experimental conditions, and evidence-level labels	Abstract-only interpretation, citation mismatch, inadequate material characterization, corrected or retracted work, and loss of experimental context	Verify the complete article, supporting passage, methods, study design, material identity, and distinction between results and interpretation	A generated answer may attribute a claim to a study that did not test the stated compound, mechanism, or outcome	Primary evidence layer for claim-level grounding and citation traceability
Review or synthesis literature where appropriate	What broader conclusions, disagreements, and evidence gaps have been reported?	Review type; search scope; eligibility criteria; included evidence; synthesis approach; date of search; DOI; limitations	Evidence overview, recurring findings, conflicting interpretations, research gaps, and candidate primary sources	Selective coverage, outdated conclusions, heterogeneous included studies, and overgeneralization from review-level statements	Examine review methods and retrieve decisive primary studies before making compound-specific or mechanism-specific claims	A review conclusion may be presented as direct experimental confirmation	Orientation and evidence-mapping layer rather than a substitute for primary verification
Natural-product databases	Which records describe the natural product, its structure, synonyms, organism source, and related biological information?	Stable record identifier; database name and version; structure representation; stereochemistry; synonyms; organism;	Candidate identities, chemical structures, synonyms, source organisms, related records, and provenance links	Duplicate records, inconsistent structures, missing stereochemistry, uncertain taxonomy, and incomplete original provenance	Cross-check identity and structure across curated resources and supporting primary literature	Different compounds may be merged, or one compound may be fragmented across multiple records	Initial molecular-record retrieval and identity-resolution layer

ISSN (Online) 2249-6084 (Print) 2250-1029

		source publication; update date					
Phytochemical databases	Which plant–compound relationships are reported, and what evidence supports the association?	Accepted botanical name; taxonomic authority; synonym history; plant part; extraction context; analytical method; compound identifier; source reference	Plant–phytochemical associations, reported constituents, plant-part records, chemical classes, and contextual annotations	Taxonomic ambiguity, unverified constituent lists, regional naming variation, and confusion between detected constituents and assumed composition	Resolve botanical taxonomy and inspect the analytical evidence supporting constituent identification	A reported plant–compound association may be presented as universally established or quantitatively characterized	Botanical-source and phytochemical-identity layer
Chemical structure collections	Which structure record corresponds to the retrieved compound name or synonym?	Canonical and isomeric structure strings; registry identifiers; formula; mass; stereochemistry; charge state; salt form; tautomer status; provenance	Standardized structure candidates, matched identifiers, synonym sets, and unresolved alternatives	Name–structure mismatch, missing stereochemistry, salt or tautomer inconsistency, and structurally duplicated records	Compare structure representations, identifiers, source articles, and analytical characterization	An incorrect structure may propagate into target, pathway, and ADMET retrieval	Molecular identity normalization and structure-matching layer
Bioactivity databases	What measured activity has been reported against a target, phenotype, organism, or assay endpoint?	Compound and target identifiers; assay type; activity value; units; relation operator; experimental system; species; conditions; publication; evidence status	Binding, inhibition, activation, phenotypic-response, potency, and assay records	Mixed endpoints, inconsistent units, censored values, assay interference, duplicated measurements, and absent experimental conditions	Normalize units and endpoint types, inspect assay design, preserve conflicting values, and verify the primary source	Incomparable measurements may be combined into a false consensus or unsupported potency claim	Quantitative bioactivity evidence layer
Target databases	Which targets are experimentally associated with the compound, and which associations	Target accession; organism; isoform; assay category; activity measurement; mechanism	Experimental target records, predicted associations, binding measurements, and	Predicted and measured targets may be mixed; protein families, homologues, or isoforms may be conflated	Separate experimental from predicted evidence and verify target identity, assay	A predicted association or textual co-occurrence may be reported as confirmed	Target-annotation and downstream pathway-query layer

	are predicted or indirect?	type; prediction status; source publication	functional effects		relevance, and species	target engagement	
Pathway databases	Which pathways contain the supported target or biological process, and in what organism or disease context?	Pathway identifier and version; organism; constituent entities; relationship type; disease context; evidence source; update date	Pathway membership, reactions, molecular interactions, disease pathways, and contextual maps	Curated pathway boundaries are incomplete and version-dependent; membership does not prove compound-induced modulation	Verify compound–target evidence first and seek direct pathway measurements in relevant experimental systems	A target–pathway relationship may be transformed into an unsupported mechanism claim	Biological-context and mechanistic-hypothesis layer
ADMET or toxicity resources	What measured or predicted pharmacokinetic, physicochemical, medicinal-chemistry, or toxicity evidence is available?	Endpoint; measured-versus-predicted status; model or assay; units; species; route; dose; exposure; applicability domain; confidence; source	Property profiles, experimental measurements, predicted endpoints, structural alerts, and evidence gaps	Predictions may fall outside their applicability domain; measured and predicted values may be combined without distinction	Label evidence status, inspect applicability, compare orthogonal sources, and define experimental follow-up	A computational estimate may be presented as demonstrated safety, exposure, or developability	Provisional developability and safety-information layer
Pharmacovigilance sources where supported	Are adverse-event or interaction signals associated with the compound, extract, product, or co-administered medicine?	Product composition; ingredient identity; event terminology; seriousness; timing; reporter type; co-medications; duplicate status; source	Reported safety signals, interaction patterns, event narratives, and uncertainty statements	Underreporting, duplicate reports, uncertain product composition, confounding, stimulated reporting, and missing denominators	Confirm exposure identity, remove duplicates, evaluate temporality and alternatives, and triangulate with toxicology and clinical evidence	A spontaneous association may be reported as causal toxicity or incidence	Conditional post-use safety-signal layer
Knowledge graphs where supported	Which traceable relationships connect compounds, targets, genes, pathways, phenotypes, diseases, and safety outcomes?	Node identifiers; relation types; directionality; source database; source publication; evidence code; graph version;	Multi-hop evidence paths, alternative hypotheses, relationship networks, conflicts, and missing links	Graph topology may amplify popular entities, duplicated sources, inferred edges, or ontology-mapping errors	Preserve edge-level provenance, distinguish observed from inferred relations, and review biologically meaningful paths	A short or highly connected graph path may be misrepresented as causal evidence	Evidence-integration and hypothesis-triangulation layer

		inferred-versus-observed status					
Expert-curated sources	Which elements have undergone domain curation, and what aspect of each record was curated?	Curating organization; curation scope; evidence code; source publication; update date; correction policy; database version	Standardized entities, controlled terminology, curated relationships, and confidence-qualified records	"Curated" may describe formatting or ontology mapping rather than experimental verification	Review the curation policy and verify high-impact claims against their underlying sources	Curated status may be treated as proof of biological or clinical validity	Priority retrieval source that still requires claim-level verification
Conflicting evidence records	Where do sources disagree about identity, occurrence, activity, target, pathway, safety, or interpretation?	Competing values; source identifiers; publication dates; methods; assay conditions; species; database versions; correction status	Conflict sets, method-stratified evidence, unresolved alternatives, and uncertainty flags	Automated harmonization may suppress minority evidence or privilege frequently repeated records	Retain competing records, compare primary methods, document adjudication, and escalate unresolved conflicts to experts	Contradictory findings may be silently averaged, omitted, or converted into false consensus	Conflict-management and uncertainty-preservation layer

Pathway knowledge and expert review

Pathway-context retrieval begins only after the relevant chemical and target entities have been sufficiently resolved. Curated pathway resources can connect genes, proteins, compounds, reactions, diseases, and cellular processes, thereby providing a structured context in which retrieved target evidence may be interpreted [15]. Such mappings can help identify whether a proposed compound–target relationship intersects with a disease-relevant process, whether multiple targets converge on related biological functions, and whether alternative pathways may explain an observed phenotype. Heterogeneous biomedical knowledge graphs may extend this analysis by integrating drugs, genes, diseases, phenotypes, and other entities into traceable relationship networks that support hypothesis prioritisation [16]. Nevertheless, neither pathway membership nor a graph path establishes that a phytochemical modulates the pathway under biologically relevant conditions.

Knowledge-graph integration should preserve the provenance and evidentiary status of every node and edge. A relationship derived from a measured binding assay is not equivalent to one inferred from textual co-occurrence, database mapping, structural similarity, or computational prediction. Biomedical knowledge resources demonstrate how multiple entity classes and relation types can be harmonized while retaining source information and standardized identifiers [17]. Within phytopharmacology, the proposed graph layer would connect botanical sources, phytochemicals, structures, targets, pathways, disease contexts, phenotypes, safety observations, and source publications. Structure-based pharmacokinetic and medicinal-chemistry estimates may be added as provisional context, but such predictions assist early triage rather than confirm exposure, developability, or biological relevance [18]. Graph integration must therefore preserve the distinction between observed, curated, predicted, inferred, and expert-proposed relationships.

Safety interpretation requires comparable discipline. Structure-based models may estimate absorption, metabolism, transport, toxicity, or related properties, but performance can vary when compounds lie outside the chemical space represented in model-development data [19]. Natural products may create additional challenges because of complex stereochemistry, unusual scaffolds, reactive functionalities, metabolites, mixtures, and uncertain exposure. Predicted ADMET or toxicity outputs should therefore be tagged with their model source, endpoint definition, applicability information, and computational status. Experimental toxicology,

pharmacokinetic measurements, reported interactions, curated chemical–gene relationships, and pharmacovigilance signals should remain separate evidence categories. The absence of retrieved toxicity information must be interpreted as a data gap rather than evidence of safety, while an isolated predicted alert should not be represented as demonstrated toxicity.

Expert review is the interpretive control that connects retrieved information with defensible research reasoning. Natural-product chemists are needed to examine botanical attribution, structural identity, stereochemistry, analytical characterization, purity, and mixture composition. Pharmacologists must evaluate potency, selectivity, exposure, dose–response relationships, assay relevance, target context, and alternative mechanisms. Toxicologists must distinguish predictive alerts, experimental findings, exposure-dependent hazards, and uncertain safety signals. Human-centred explanation should enable experts to interrogate why evidence was retrieved, how records were connected, which assumptions were made, and where uncertainty remains, rather than merely presenting technically interpretable model output [20]. Expert corrections should feed back into query refinement, synonym resolution, corpus curation, conflict annotation, and future retrieval while remaining documented in an auditable record.

Drug discovery decision logic

Retrieval-augmented drug discovery decision logic should begin with a narrowly framed question rather than an open-ended request to identify a promising drug. The query should specify the botanical material or phytochemical, disease or phenotype context, evidence type, target or pathway hypothesis, safety concern, and intended research decision. Artificial intelligence may assist target identification, compound screening, molecular characterization, and other discovery stages, but advancement between stages continues to require fit-for-purpose experimental and development validation [21]. The proposed logic therefore converts a broad discovery problem into a sequence of answerable evidence questions, each associated with source requirements, admissible interpretations, uncertainty labels, and explicit stopping conditions.

The first decision gate concerns molecular and evidentiary identity. Candidate records should not proceed to comparative interpretation until botanical taxonomy, plant part, extraction or isolation context, compound name, structure, stereochemistry, and synonym relationships have been reconciled sufficiently for the intended question. The next gate evaluates whether retrieved bioactivity represents measured binding, functional modulation, phenotypic response, prediction, review-level interpretation, or textual association. Drug-discovery AI literature supports the integration of chemical, biological, target, pathway, and development information across iterative discovery stages [22]. Within the proposed framework, however, integration does not mean numerical aggregation by default. Each evidence type remains visible so that conflicting assays, incompatible units, different organisms, or shared source provenance cannot be concealed within a composite score.

The third gate examines biological relevance and conflict. Target annotations should be evaluated for species, isoform, assay design, selectivity, experimental system, and relationship to achievable exposure. Pathway mappings should be interpreted in disease or phenotype context and checked for directionality, tissue relevance, alternative explanations, and evidence of actual pathway perturbation. Chemical and biological datasets can contain heterogeneous labels, assay conditions, measurement errors, inactive-value conventions, duplicate records, and context-dependent outcomes that materially alter the interpretation of computational results [23]. Consequently, conflicting evidence should be retained as a structured conflict set rather than removed automatically. Resolution may involve prioritising primary over derivative records, comparing methods, separating contexts, requesting expert adjudication, or marking the question as unresolved.

The final decision gate determines whether the available evidence justifies further investigation and which validation activities are required. A candidate may be considered suitable for additional research when its identity is sufficiently resolved, relevant evidence is traceable, bioactivity is interpretable, mechanistic hypotheses are plausible, major conflicts are documented, and safety or developability uncertainties are explicit. Reliable machine-learning support in drug discovery requires well-defined questions, appropriate data, interpretability, repeatability, and prospective validation rather than dependence on retrospective performance alone [24]. Retrieval-augmented decision logic should therefore produce an auditable rationale, an uncertainty profile, a list of competing explanations, and a validation plan. It must not designate a validated drug candidate, infer clinical efficacy, recommend treatment, determine dosing, or cross the no-clinical-claim boundary.

Proposed framework

The proposed retrieval-augmented phytopharmacology framework begins with research-question definition and evidence-scope specification. The user or expert defines the compound, extract, botanical source, disease or phenotype context, target or pathway hypothesis, safety question, and intended drug discovery decision. The framework then converts this definition into structured

retrieval queries for literature, molecular databases, bioactivity resources, target records, pathway knowledge, and safety information. Multiparameter ADMET platforms may provide provisional physicochemical, pharmacokinetic, medicinal-chemistry, and toxicity predictions, but every retrieved result must be labelled as measured, curated, predicted, inferred, or unverified [25]. Query expansion may use synonyms, taxonomic names, molecular identifiers, and ontology terms, while query restriction prevents evidence from unrelated compounds, species, extracts, or disease contexts from entering the reasoning chain.

The second stage integrates literature and molecular database records through identity resolution and provenance control. Retrieved compound names are mapped to standardized structures, stereochemical forms, salts, tautomers, identifiers, botanical sources, and supporting publications. Bioactivity evidence is organized by target identity, assay type, measurement endpoint, units, experimental system, species, conditions, and publication source. FAIR binding resources illustrate why protein–small-molecule measurements should remain connected to target identifiers, measurement types, and source-level provenance [26]. Duplicate database records are traced to their originating evidence where possible, and contradictory structures, activities, or target claims are preserved rather than silently harmonized. Records that cannot be reconciled are labelled unresolved and excluded from high-confidence mechanistic interpretation.

The third stage connects verified entities with pathway, disease, phenotype, ADMET, toxicity, and safety knowledge. Chemical–gene–disease resources can support the retrieval of curated relationships and inferential networks relevant to mechanistic or toxicological hypotheses, but such associations do not establish causality [27]. Pathway mappings are therefore generated only after the compound–target evidence category has been identified, and every path retains its source and relation type. Evidence triangulation evaluates whether apparently convergent findings arise from genuinely independent methods or from repeated imports of the same underlying record. Traditional-use context may be retrieved where supported, but it remains distinct from experimental pharmacology. Safety signals, predicted toxicity, experimental toxicity, interactions, and absence of data are likewise represented as separate categories.

The fourth stage introduces multidisciplinary expert review and validation planning. Natural-product chemistry review addresses identity, analytical characterization, source material, stereochemistry, purity, stability, and extract composition. Pharmacology review examines assay suitability, potency, selectivity, exposure, target context, pathway plausibility, and alternative mechanisms. Toxicology review evaluates predicted hazards, measured toxicity, interactions, safety signals, dose, route, and relevance to the proposed research context. Evaluation of retrieval-supported systems should consider alternative models or configurations, different source corpora, consistency, safety, generalizability, and comparison with expert-derived answers rather than relying on a single performance measure [28]. In the proposed framework, experts can accept, revise, reject, or qualify retrieved relationships and specify experiments that would resolve the most consequential uncertainty.

The fifth stage produces a bounded drug discovery decision output consisting of an evidence-grounded hypothesis, supporting and opposing evidence, unresolved conflicts, uncertainty categories, candidate-prioritisation rationale, validation plan, and audit trail. Retrieval does not eliminate unsupported generation and may introduce errors when irrelevant, incomplete, or source-misaligned information is incorporated into a response [29]. Hallucination control must therefore include claim-to-source checking, passage verification, evidence-type labelling, contradiction detection, unsupported-claim rejection, and explicit acknowledgement of missing information. Expert corrections, newly available sources, experimental findings, and database updates may feed back into the corpus and query process, but generated summaries must never be recycled as primary evidence. **Table 2** presents the proposed retrieval-augmented phytopharmacology framework for evidence-grounded drug discovery reasoning. **Figure 1** illustrates the proposed retrieval-augmented phytopharmacology framework connecting literature evidence, molecular databases, pathway knowledge, expert review, decision logic, validation planning, and uncertainty communication.

Table 2. Proposed Retrieval-Augmented Phytopharmacology Framework for Drug Discovery: Query Formulation, Literature Retrieval, Molecular Database Integration, Pathway Knowledge Mapping, Evidence Triangulation, Expert Review, Decision Logic, Validation Planning, and Uncertainty Boundaries

Framework stage	Core purpose	Required input	Retrieval or interpretive function	Potential output	Reliability concern	Expert review requirement	Validation boundary	Drug discovery decision value

Research-question definition	Convert a broad discovery problem into a bounded evidence question	Botanical source, compound or extract, disease or phenotype context, target or pathway hypothesis, safety question, intended research decision	Define entities, evidence categories, exclusions, and acceptable claim scope	Structured research question and retrieval specification	Ambiguous questions may retrieve unrelated compounds, species, contexts, or evidence types	Domain expert confirms terminology, scope, and intended use	Question definition does not establish that the proposed relationship exists	Clarifies which evidence is needed before prioritisation can begin
Source-corpus definition	Identify admissible and traceable evidence sources	Peer-reviewed literature, review literature where appropriate, curated molecular resources, pathway and safety sources	Select sources according to relevance, curation, accessibility, provenance, and update status	Versioned source inventory and inclusion rationale	Broad or poorly curated corpora may increase irrelevant retrieval and unsupported synthesis	Information specialist and domain expert review source eligibility	Inclusion in the corpus does not imply evidentiary validity	Establishes the evidence universe for a reproducible search
Query formulation and expansion	Retrieve relevant evidence despite terminology variation	Standardized entities, synonyms, taxonomic names, molecular identifiers, ontology terms, disease and assay terminology	Generate source-specific queries and controlled synonym expansions	Search queries, entity mappings, and retrieval logs	Uncontrolled expansion may mix different compounds, taxa, extracts, targets, or indications	Natural-product chemist and pharmacologist inspect high-impact query expansions	Query relevance does not validate retrieved records	Improves recall while preserving domain boundaries
Literature evidence retrieval	Locate source passages and study-level evidence	Structured queries, eligible literature corpus, entity and relation terms	Retrieve articles and passages by concepts, methods, assays, outcomes, and evidence types	Traceable passages, study descriptors, reported findings, and limitations	Abstract-only evidence, passage mismatch, citation laundering, and omission of negative findings	Reviewer verifies full-text context and whether the cited passage supports the claim	Retrieved text does not independently confirm identity, mechanism, efficacy, or safety	Provides claim-level evidence for later synthesis
Molecular database retrieval	Collect structures, identifiers, botanical	Compound names, structures, synonyms, taxa, assay	Search natural-product, phytochemical, structure,	Candidate identity, structure, source, activity,	Duplicate, incomplete, outdated, predicted,	Chemist and data curator inspect structure and	Database presence does not prove occurrence,	Expands the evidence map beyond

	sources, activities, targets, and related records	and target terms	bioactivity, and target resources	and target records	or provenance-poor records	record provenance	activity, target engagement, or relevance	narrative literature
Phytochemical and botanical identity resolution	Determine whether records refer to the same material or entity	Taxonomy, plant part, extraction context, analytical evidence, structures, stereochemistry, identifiers, synonyms	Normalize names and compare botanical and chemical records	Resolved identity, alternative identities, or unresolved conflict set	Synonym collisions, taxonomic changes, extract–compound substitution, and stereochemical ambiguity	Natural-product chemist and taxonomic specialist where required	Unresolved identity prevents high-confidence downstream interpretation	Protects the workflow from propagating incorrect entity mappings
Chemical structure record matching	Align names and identifiers with the correct molecular representation	Canonical and isomeric structures, formula, mass, charge, salt, tautomer and stereochemical data	Compare structure representations and identify duplicates or incompatibilities	Standardized structure record with provenance or multiple unresolved alternatives	Name–structure mismatch and incomplete stereochemistry can contaminate all later retrieval	Cheminformatics and chemistry review of material discrepancies	Structure matching does not confirm that the tested biological material contained the compound	Enables consistent cross-database retrieval and predictive analysis
Bioactivity and target evidence organization	Distinguish measured, functional, phenotypic, predicted, and indirect associations	Assay records, activity endpoints, units, target accessions, species, experimental conditions, source publications	Normalize evidence categories without collapsing incompatible measurements	Assay-aware bioactivity and target evidence profile	Unit incompatibility, assay interference, duplicate measurements, and predicted–measured conflation	Pharmacologist reviews assay quality, potency, selectivity, and biological context	Bioactivity does not establish therapeutic relevance or in vivo activity	Identifies whether a candidate has evidence suitable for mechanistic follow-up
Pathway and disease-context mapping	Place supported targets and phenotypes within biological systems	Verified target evidence, pathway resources, disease or phenotype context, tissue and organism metadata	Retrieve pathway relationships and construct traceable contextual paths	Pathway hypotheses, alternative mechanisms, and context gaps	Database membership, graph proximity, or co-occurrence may be misinterpreted as causality	Pharmacologist or systems biologist reviews directionality, tissue relevance, and alternatives	Pathway mapping supports hypotheses but does not validate mechanism	Helps identify mechanistically informative experiments and competing explanations
ADMET, toxicity, and safety retrieval	Identify developability concerns,	Chemical structure, measured data,	Retrieve and classify measured, predicted,	Evidence-qualified ADMET and safety	Out-of-domain predictions,	Toxicologist and pharmacologist assess	No retrieved profile can declare a	Identifies hazards and validation

	safety evidence, and missing information	predictive resources, toxicogenomics, interaction and pharmacovigilance evidence	curated, and reported safety information	profile with data gaps	uncertain exposure identity, confounding, and false reassurance from absent data	relevance, dose, route, and uncertainty	candidate safe or clinically acceptable	priorities before further investment
Evidence provenance and duplicate control	Determine the origin and independence of retrieved claims	DOI, passage, database version, record identifier, import history, evidence code, publication links	Trace records to originating sources and detect derivative duplicates	Provenance map and independent-evidence count without numerical confidence scoring	Multiple database copies of one study may appear as independent corroboration	Data curator verifies source chains and unresolved provenance	Frequency of repetition does not increase biological validity	Prevents false triangulation and supports transparent evidence weighting
Conflict detection and resolution	Preserve and interpret disagreement across sources	Competing structures, activities, targets, pathways, safety findings, methods, species, contexts, and dates	Create conflict sets and compare evidence according to method and context	Resolved explanation, context-specific interpretation, or unresolved disagreement	Automated summarisation may suppress contradictions or invent reconciliation	Relevant disciplinary experts document adjudication and residual uncertainty	Unresolved conflict reduces confidence and cannot be silently removed	Shows whether additional evidence is required before prioritisation
Evidence triangulation	Assess convergence across independent evidence types	Literature, molecular records, assays, pathways, safety data, and provenance maps	Compare findings across methods while identifying shared sources and incompatible contexts	Evidence synthesis with supporting, opposing, and missing elements	Correlated databases and repeated records may create false convergence	Multidisciplinary panel examines independence and methodological compatibility	Convergence supports plausibility but does not replace experimental validation	Strengthens or weakens the rationale for further investigation
Expert mechanistic plausibility review	Test whether the proposed evidence chain is biologically and chemically coherent	Identity evidence, bioactivity, target, pathway, exposure, phenotype, safety, and alternatives	Interrogate assumptions, causal gaps, dose relevance, and competing explanations	Qualified mechanistic hypothesis and explicit evidence gaps	Fluent generated explanations may conceal missing links or impossible exposure assumptions	Chemistry, pharmacology, systems biology, and toxicology expertise as appropriate	Plausibility is not proof of mechanism	Converts retrieved associations into falsifiable research hypotheses
Candidate-prioritisat	Determine whether a	Identity confidence, evidence	Produce a transparent rationale	Research-priority recommend	Hidden weighting, missing	Multidisciplinary review of	Prioritisation authorizes further	Supports allocation of

ion boundary	candidate merits additional research relative to alternatives	relevance, conflicts, safety gaps, novelty, feasibility, and validation needs	without claiming definitive ranking or success	ation, deferral, or rejection rationale	evidence, correlated inputs, and overconfidence	assumptions, criteria, and uncertainties	evaluation only; it does not validate a drug candidate	experimental attention and resources
Experimental validation planning	Define tests that can confirm, refine, or falsify the leading hypothesis	Major uncertainties, proposed target, pathway context, material identity, exposure and safety gaps	Translate the evidence map into staged experimental questions	Assay plan, controls, materials requirements, endpoints, and decision criteria	Generic or inappropriate experiments may fail to address the actual uncertainty	Experimental pharmacologist, chemist, toxicologist, and assay specialist review the plan	A validation plan is not evidence that validation has occurred	Directs subsequent laboratory work toward the most decision-relevant gaps
Uncertainty communication	Make evidence limits visible and prevent categorical overstatement	Evidence categories, missing metadata, conflicts, predictions, applicability information, and reviewer comments	Label what is observed, predicted, inferred, disputed, missing, or outside scope	Calibrated narrative and structured uncertainty statement	Invented confidence values or vague caveats may obscure the true limitation	Experts confirm that wording matches evidence strength	Uncertainty cannot be converted into certainty by generation or retrieval frequency	Allows cautious comparison without concealing evidence gaps
No clinical-claim boundary	Prevent research evidence from becoming unsupported patient-level advice	Intended use, evidence level, generated claims, clinical relevance, and user context	Detect and block diagnostic, therapeutic, dosing, safety, or patient-management claims	Research-only output with prohibited clinical claims removed or qualified	Preclinical associations may be transformed into treatment recommendations	Clinical and governance review if any clinical interpretation is contemplated	The conceptual framework does not make clinical decisions or treatment recommendations	Maintains a clear translational and ethical limit
Audit trail and feedback updating	Preserve reproducibility and improve future retrieval through verified corrections	Query logs, retrieved passages, database versions, transformations, reviewer decisions, new evidence, and corrections	Store traceable reasoning steps and update sources or mappings under controlled review	Reproducible evidence trail, correction history, and refined retrieval logic	Generated text may be recycled as evidence or undocumented updates may alter conclusions	Curator approves updates and prevents unsupported content entering the evidence corpus	Only verified sources and documented expert annotations may update the framework	Supports correction, version comparison, and iterative improvement

ISSN (Online) 2249-6084 (Print) 2250-1029

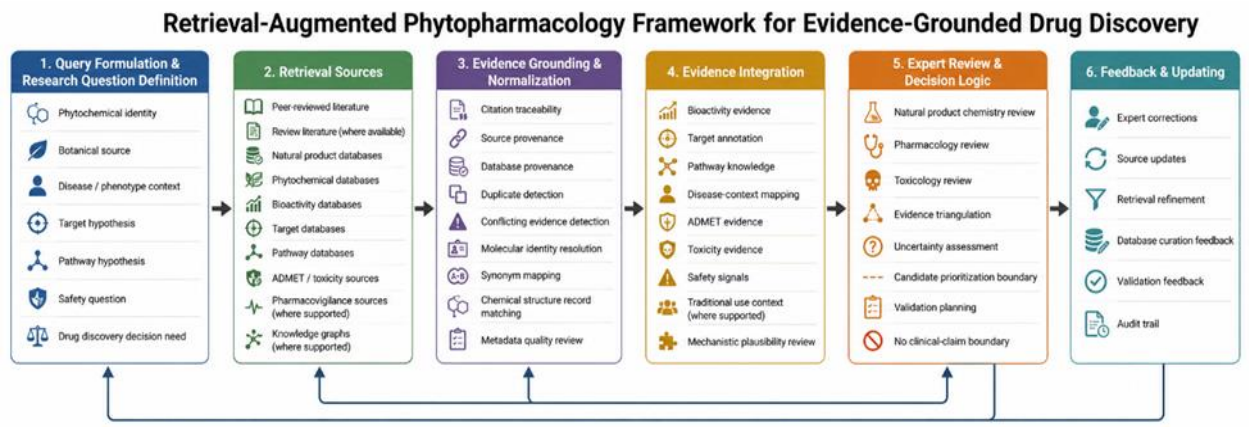

Figure 1. Retrieval-Augmented Phytopharmacology Framework for Evidence-Grounded Drug Discovery

Alt-text description

A conceptual framework showing retrieval queries flowing into literature evidence, molecular databases, pathway knowledge, source grounding, evidence triangulation, expert review, decision logic, validation planning, uncertainty communication, and feedback loops.

CONCLUSION

This article proposes a retrieval-augmented phytopharmacology framework that connects literature evidence, molecular databases, phytochemical and botanical identity, chemical structures, bioactivity records, target annotations, pathway knowledge, ADMET and toxicity information, expert review, drug discovery decision logic, validation planning, and uncertainty communication. Its central contribution is the organization of these components into a traceable research workflow in which retrieval improves source access and evidence structuring without being treated as biological verification. The framework requires source and database provenance, claim-level citation traceability, duplicate and conflict detection, measured-versus-predicted evidence labels, multidisciplinary review, hallucination controls, and an auditable feedback process. Pathway knowledge and graph relationships are used to formulate mechanistic hypotheses rather than confirm mechanisms, while candidate prioritisation remains a bounded recommendation for additional research rather than recognition of a validated drug candidate. Retrieval augmentation may therefore strengthen hypothesis generation and evidence-grounded decision support only when uncertainty remains visible, expert interpretation remains central, experimental validation is planned explicitly, and clinical claims remain outside the permitted scope.

Acknowledgments: None

Conflict of interest: None

Financial support: None

Ethics statement: None

REFERENCES

[1] Liu S, McCoy AB, Wright A. Improving large language model applications in biomedicine with retrieval-augmented generation: A systematic review, meta-analysis, and clinical development guidelines. J Am Med Inform Assoc. 2025;32(4):605-15. doi:10.1093/jamia/ocaf008

[2] Yang R, Ning Y, Keppo E, Liu M, Hong C, Bitterman DS, et al. Retrieval-augmented generation for generative artificial intelligence in health care. NPJ Health Syst. 2025;2:2. doi:10.1038/s44401-024-00004-1

[3] Bender A, Cortés-Ciriano I. Artificial intelligence in drug discovery: What is realistic, what are illusions? Part 1: Ways to make an impact, and why we are not there yet. Drug Discov Today. 2021;26(2):511-24. doi:10.1016/j.drudis.2020.12.009

[4] Thomford NE, Senthebane DA, Rowe A, Munro D, Seele P, Maroyi A, et al. Natural products for drug discovery in the 21st century: Innovations for novel drug discovery. Int J Mol Sci. 2018;19(6):1578. doi:10.3390/ijms19061578

[5] Jeong M, Sohn J, Sung M, Kang J. Improving medical reasoning through retrieval and self-reflection with retrieval-augmented large language models. Bioinformatics. 2024;40(Suppl 1). doi:10.1093/bioinformatics/btae238

[6] Kresevic S, Giuffrè M, Ajcevic M, Accardo A, Crocè LS, Shung DL. Optimization of hepatological clinical guidelines interpretation by large language models: A retrieval augmented generation-based framework.

NPJ Digit Med. 2024;7(1):102. doi:10.1038/s41746-024-01091-y

[7] Lopez I, Swaminathan A, Vedula K, Narayanan S, Haredasht FN, Ma SP, et al. Clinical entity augmented retrieval for clinical information extraction. NPJ Digit Med. 2025;8(1):45. doi:10.1038/s41746-024-01377-1

[8] Ong CS, Obey NT, Zheng Y, Cohan A, Schneider EB. SurgeryLLM: A retrieval-augmented generation large language model framework for surgical decision support and workflow enhancement. NPJ Digit Med. 2024;7(1):364. doi:10.1038/s41746-024-01391-3

[9] Wei CH, Allot A, Leaman R, Lu Z. PubTator central: Automated concept annotation for biomedical full text articles. Nucleic Acids Res. 2019;47(W1). doi:10.1093/nar/gkz389

[10] Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, et al. BioBERT: A pre-trained biomedical language representation model for biomedical text mining. Bioinformatics. 2020;36(4):1234-40. doi:10.1093/bioinformatics/btz682

[11] Sorokina M, Steinbeck C. Review on natural products databases: Where to find data in 2020. J Cheminform. 2020;12(1):20. doi:10.1186/s13321-020-00424-9

[12] Sorokina M, Merseburger P, Rajan K, Yirik MA, Steinbeck C. COCONUT online: Collection of Open Natural Products database. J Cheminform. 2021;13(1):2. doi:10.1186/s13321-020-00478-9

[13] Zeng X, Zhang P, He W, Qin C, Chen S, Tao L, et al. NPASS: Natural product activity and species source database for natural product research, discovery and tool development. Nucleic Acids Res. 2018;46(D1). doi:10.1093/nar/gkx1026

[14] Mohanraj K, Karthikeyan BS, Vivek-Ananth RP, Chand RPB, Aparna SR, Mangalapandi P, et al. IMPPAT: A curated database of Indian medicinal plants, phytochemistry and therapeutics. Sci Rep. 2018;8(1):4329. doi:10.1038/s41598-018-22631-z

[15] Kanehisa M, Furumichi M, Sato Y, Ishiguro-Watanabe M, Tanabe M. KEGG: Integrating viruses and cellular organisms. Nucleic Acids Res. 2021;49(D1). doi:10.1093/nar/gkaa970

[16] Himmelstein DS, Lizee A, Hessler C, Brueggeman L, Chen SL, Hadley D, et al. Systematic integration of biomedical knowledge prioritizes drugs for repurposing. Elife. 2017;6. doi:10.7554/eLife.26726

[17] Chandak P, Huang K, Zitnik M. Building a knowledge graph to enable precision medicine. Sci Data. 2023;10(1):67. doi:10.1038/s41597-023-01960-3

[18] Daina A, Michielin O, Zoete V. SwissADME: A free web tool to evaluate pharmacokinetics, drug-likeness and medicinal chemistry friendliness of small molecules. Sci Rep. 2017;7(1):42717. doi:10.1038/srep42717

[19] Yang H, Lou C, Sun L, Li J, Cai Y, Wang Z, et al. admetSAR 2.0: Web-service for prediction and optimization of chemical ADMET properties. Bioinformatics. 2019;35(6):1067-9. doi:10.1093/bioinformatics/bty707

[20] Holzinger A, Langs G, Denk H, Zatloukal K, Müller H. Causability and explainability of artificial intelligence in medicine. Wiley Interdiscip Rev Data Min Knowl Discov. 2019;9(4). doi:10.1002/widm.1312

[21] Mak KK, Pichika MR. Artificial intelligence in drug development: Present status and future prospects. Drug Discov Today. 2019;24(3):773-80. doi:10.1016/j.drudis.2018.11.014

[22] Paul D, Sanap G, Shenoy S, Kalyane D, Kalia K, Tekade RK. Artificial intelligence in drug discovery and development. Drug Discov Today. 2021;26(1):80-93. doi:10.1016/j.drudis.2020.10.010

[23] Bender A, Cortés-Ciriano I. Artificial intelligence in drug discovery: What is realistic, what are illusions? Part 2: A discussion of chemical and biological data. Drug Discov Today. 2021;26(4):1040-52. doi:10.1016/j.drudis.2020.11.037

[24] Vamathevan J, Clark D, Czodrowski P, Dunham I, Ferran E, Lee G, et al. Applications of machine learning in drug discovery and development. Nat Rev Drug Discov. 2019;18(6):463-77. doi:10.1038/s41573-019-0024-5

[25] Xiong G, Wu Z, Yi J, Fu L, Yang Z, Hsieh C, et al. ADMETlab 2.0: An integrated online platform for accurate and comprehensive predictions of ADMET properties. Nucleic Acids Res. 2021;49(W1). doi:10.1093/nar/gkab255

[26] Liu T, Hwang L, Burley SK, Nitsche CI, Southan C, Walters WP, et al. BindingDB in 2024: A FAIR knowledgebase of protein-small molecule binding data. Nucleic Acids Res. 2025;53(D1). doi:10.1093/nar/gkae1075

[27] Davis AP, Wiegers TC, Johnson RJ, Sciaky D, Wiegers J, Mattingly CJ. Comparative Toxicogenomics Database (CTD): Update 2023. Nucleic Acids Res. 2023;51(D1). doi:10.1093/nar/gkac833

[28] Ke YH, Jin L, Elangovan K, Abdullah HR, Liu N, Sia ATH, et al. Retrieval augmented generation for 10 large language models and its generalizability in assessing medical fitness. NPJ Digit Med. 2025;8(1):187. doi:10.1038/s41746-025-01519-z

[29] Masanneck L, Meuth SG, Pawlitzki M. Evaluating base and retrieval augmented LLMs with document or online support for evidence based neurology. NPJ Digit Med. 2025;8(1):137. doi:10.1038/s41746-025-01536-y