Large Language Models in Computational Pharmacology: A Scoping Review of Knowledge Extraction, Mechanistic Reasoning, Drug Repurposing, and Decision Support

Pieter Botha[1]*, Anel Van Wyk[1], Johan Marais[2]

[1]Department of AI for South African Fynbos and Succulent Bioactives, Faculty of Pharmacy, Stellenbosch University, Stellenbosch, South Africa.

[2]Department of Computational Docking for Anti-Tubercular Natural Products, Faculty of Pharmacy, University of Pretoria, Pretoria, South Africa.

ABSTRACT

Large language models (LLMs) are emerging as generative AI systems with potential relevance to computational pharmacology, including biomedical text mining, drug–target evidence organization, adverse-event extraction, mechanistic hypothesis generation, drug repurposing support, and pharmacology-related decision support. This scoping review maps how LLMs, biomedical language models, domain-specific language models, and retrieval-augmented or tool-augmented systems are being used or proposed across computational pharmacology, while distinguishing evidence organization and hypothesis generation from validated pharmacological or clinical decision-making. PubMed, Scopus, Web of Science, IEEE Xplore, ScienceDirect, SpringerLink, Wiley Online Library, Oxford Academic journals, ACM Digital Library, Nature Portfolio journals, Frontiers journals, MDPI journals, and publisher records were searched for peer-reviewed journal articles published from 1 January 2017 to 9 July 2026; records were screened using predefined eligibility criteria, duplicates were removed, titles and abstracts were screened, full texts were assessed, and 34 studies were included. The mapped literature clustered around biomedical language-model foundations, knowledge extraction, drug–drug interaction extraction, drug–target relation extraction, adverse-event extraction, pharmacovigilance support, mechanistic reasoning assistance, drug repurposing hypothesis generation, medication-information support, and bounded decision-support interfaces. Major gaps involved source grounding, hallucination control, prompt sensitivity, external validation, expert evaluation, reproducibility, privacy, and decision-boundary governance. LLMs may support computational pharmacology by organizing evidence, extracting relations, assisting literature review, and generating hypotheses, but reliable use requires source grounding, task-specific validation, expert oversight, uncertainty communication, and carefully defined boundaries between computational assistance and pharmacological, clinical, or regulatory claims.

Key Words: *Large language models, Computational pharmacology, Scoping review, Knowledge extraction, Mechanistic reasoning, Drug repurposing*

eIJPPR 2026; 16(1):47-59

HOW TO CITE THIS ARTICLE: Botha P, Van Wyk A, Marais J. Large Language Models in Computational Pharmacology: A Scoping Review of Knowledge Extraction, Mechanistic Reasoning, Drug Repurposing, and Decision Support. Int J Pharm Phytopharmacol Res. 2026;16(1):47-59. https://doi.org/10.51847/mvLZkjgRLu

INTRODUCTION

Large language models are increasingly discussed in biomedical contexts because they can generate, summarize, transform, and reason over natural-language inputs in ways that make them attractive for evidence-intensive scientific and clinical tasks. In medicine, these systems are best interpreted as assistance technologies

Corresponding author: Pieter Botha
Address: Department of AI for South African Fynbos and Succulent Bioactives, Faculty of Pharmacy, Stellenbosch University, Stellenbosch, South Africa.
E-mail: ✉ pieter.botha@sun.ac.za
Received: 02 November 2025; **Revised:** 29 January 2026; **Accepted:** 05 February 2026

rather than autonomous knowledge authorities, because generated outputs require verification, contextual interpretation, and human oversight before they can support consequential decisions [1].

Computational pharmacology is especially knowledge-intensive because it integrates literature mining, chemical and biological entities, drug–target evidence, pathway-level interpretation, pharmacokinetic and pharmacodynamic reasoning, drug repurposing hypotheses, adverse-event information, and decision-support boundaries. Recent discussions of LLMs in intelligent medicine and pharmaceutical research position these systems as potentially useful for organizing complex biomedical and pharmacological information, while also emphasizing that pharmaceutical relevance does not by itself establish reliability or translational readiness [2].

The clinical pharmacology context adds a further layer of caution because pharmacological decisions involve drug exposure, safety, interactions, population variability, therapeutic uncertainty, and accountability. LLM applications in clinical pharmacology therefore require explicit validation boundaries, workflow constraints, and expert review rather than assumptions that fluent answers correspond to clinically usable pharmacological reasoning [3].

A scoping review is needed because the literature on LLMs in computational pharmacology is broad, heterogeneous, and still developing across biomedical language models, text-mining systems, generative chatbots, retrieval-augmented systems, chemical tool use, drug repurposing frameworks, pharmacovigilance workflows, and medication-safety interfaces. The objective of this review is to map the breadth and distribution of this literature, identify application domains and evidence types, and synthesize validation gaps, safety concerns, and research priorities without presenting LLM-assisted outputs as validated mechanisms, confirmed repurposing candidates, clinical recommendations, or regulatory-ready systems [4].

Large language models in pharmacology

In pharmacology-related contexts, LLMs can be understood as language-based models that process biomedical, chemical, pharmacological, clinical, or safety-related text to support tasks such as extraction, summarization, question answering, evidence organization, and hypothesis generation. Biomedical language models such as BioBERT established an important foundation for this area by showing that domain-adapted representations can improve biomedical text-mining tasks, including entity and relation-oriented tasks relevant to pharmacological literature mining [5].

Domain-specific pretraining is particularly relevant to computational pharmacology because pharmacological language contains specialized terms for compounds, targets, pathways, diseases, adverse events, formulations, populations, and experimental systems. PubMedBERT illustrates the rationale for biomedical-domain pretraining, in which language-model training on biomedical corpora can improve performance on biomedical natural-language-processing tasks compared with models trained primarily on general-domain language [6].

Generative biomedical language models extend this landscape from representation learning and classification toward text generation, biomedical summarization, and generative mining. BioGPT is relevant to pharmacology-related review questions because it was designed for biomedical text generation and mining, but its role in a scoping synthesis should be framed as support for biomedical language tasks rather than proof that generated biomedical statements are inherently valid or pharmacologically causal [7].

General-purpose and medical LLMs differ from classical text-mining systems because they can produce open-ended answers, explanations, rationales, and conversational outputs rather than only predefined entity labels or relation classes. However, medical LLM evaluation also shows why pharmacological use requires task-specific testing, source verification, uncertainty handling, and human-in-the-loop oversight, especially when outputs may be used near medication safety, clinical pharmacy, or translational decision-support contexts [8].

Scoping review method

This review followed a scoping-review logic because the purpose was to map application domains, model categories, evidence types, validation patterns, limitations, and research gaps rather than to estimate pooled effects or produce clinical recommendations. Reporting was structured using PRISMA-ScR-style elements, including identification, deduplication, screening, eligibility assessment, exclusion reasons, and final inclusion numbers [9].

The review questions asked what types of LLMs are being applied in computational pharmacology; how LLMs are used for knowledge extraction, literature mining, and pharmacological relation extraction; how they are used or proposed for mechanistic reasoning and drug repurposing support; how they are being applied in pharmacology-related decision-support contexts; what evaluation, validation, grounding, and oversight strategies are reported; and what gaps and safety concerns remain visible across the literature. This approach is consistent with scoping-review methodology because the field is heterogeneous, methodologically diverse, and not yet suited to quantitative pooling across comparable outcomes [10].

The PCC framework defined the population as LLM-based, biomedical language model, domain-adapted language model, retrieval-augmented, or generative chatbot systems; the concept as knowledge extraction, literature mining, mechanistic reasoning, drug repurposing support, adverse-event extraction, pharmacovigilance, medication safety, and decision-support boundaries; and the context as computational pharmacology, drug discovery and development, drug–target evidence, pharmacological text mining, clinical pharmacy, medication information, pharmacovigilance, and translational decision-support settings. Searches were conducted in PubMed, Scopus, Web of Science Core Collection, IEEE Xplore, ScienceDirect, SpringerLink, Wiley Online Library, Oxford Academic journals, ACM Digital Library, Nature Portfolio journals, Frontiers journals, MDPI journals, and publisher records for the period from 1 January 2017 to 9 July 2026. Search strings combined terms for "large language model," "LLM," "generative language model," "biomedical language model," "ChatGPT," "retrieval augmented generation," and "domain-specific language model" with pharmacology-related terms including "drug discovery," "drug repurposing," "drug target," "drug interaction," "pharmacovigilance," "adverse event," and "decision support."

Eligibility criteria included peer-reviewed journal articles published from 2017 to 2026 inclusive, DOI availability, focus on LLMs or biomedical/domain-specific language models, relevance to computational pharmacology or adjacent pharmacology-related tasks, and sufficient detail to chart at least one model type, task, data source, evaluation approach, validation type, output category, limitation, or safety concern. Exclusion criteria removed conference-only records, preprint-only records, books, theses, websites, software manuals, vendor or database pages, GitHub repositories, regulatory pages, policy documents, white papers, articles outside the search period, articles without DOI information, general AI studies without LLM relevance, clinical diagnosis-only studies without drug or medication relevance, molecular foundation model-only studies without language or literature relevance, and articles without extractable pharmacology application or evaluation information.

The PRISMA-ScR-style screening log identified 642 records from databases and 28 records from citation chasing or other sources, giving 670 total records before deduplication. After removing 188 duplicate records, 482 records remained for title and abstract screening; 430 were excluded at that stage, leaving 52 full-text articles for eligibility assessment. Eighteen full-text articles were excluded: 3 were not LLM or biomedical language model studies, 4 were not computational-pharmacology, drug-discovery, medication-safety, or pharmacovigilance relevant, 3 were conference-only or preprint-only records, 1 had no DOI, 1 was outside the search period, 2 had insufficient methodological detail, 2 had no extractable pharmacology application, 1 was a duplicate or overlapping report, and 1 was not aligned with the review questions. The final scoping review included 34 studies and 34 final manuscript references. Evidence charting captured LLM category, pharmacology context, application domain, input data source, output type, evaluation method, validation or expert review, source-grounding strategy, safety or hallucination concern, main finding, main limitation, and contribution to the scoping synthesis. **Figure 1** presents the PRISMA-ScR-style flow of record identification, duplicate removal, screening, eligibility assessment, exclusion reasons, and final included studies.

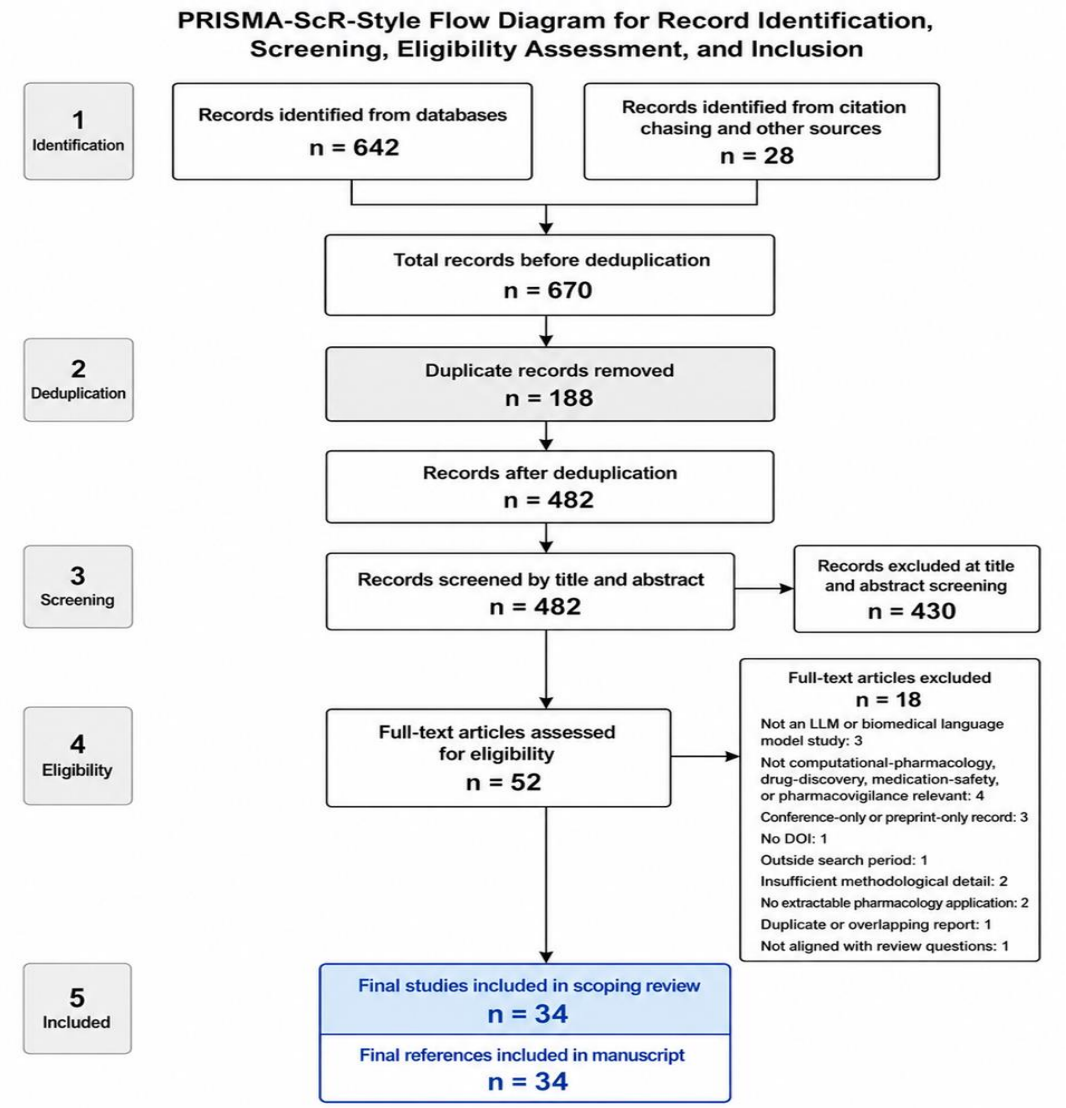

Figure 1. PRISMA-ScR-Style Flow Diagram for Record Identification, Screening, Eligibility Assessment, and Inclusion

Alt-text description

A PRISMA-ScR-style flow diagram showing database records, additional records where applicable, duplicate removal, title and abstract screening, full-text eligibility assessment, reasons for exclusion, and final included studies.

Table 1 summarises the scoping review questions, search strategy, eligibility criteria, evidence-charting domains, and synthesis approach used in this review.

Table 1. Scoping Review Method for Large Language Models in Computational Pharmacology: Review Questions, Search Strategy, Eligibility Criteria, Evidence-Charting Domains, Application Mapping, Validation-Gap Mapping, and Synthesis Approach

Review component	Specification	Rationale	Evidence charted	Mapping output	Limitation controlled
Review rationale	Map LLM applications in computational pharmacology rather than estimate pooled effects.	The literature is broad, heterogeneous, and methodologically diverse.	Model type, task, domain, validation, safety concern.	Scoping evidence map across applications.	Avoids unsupported meta-analysis claims.
Review questions	Six questions covering model types, extraction, reasoning,	Ensures breadth without drifting into a generic AI review.	Application area, output category,	Question-aligned synthesis.	Prevents unstructured

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

	repurposing, decision support, validation, and gaps.		evaluation method, limitation.		narrative selection.
PCC logic	Population: LLM-based systems; Concept: pharmacology-oriented extraction, reasoning, safety, and support; Context: computational pharmacology and adjacent medication domains.	Keeps the review aligned with scoping-review terminology.	Population, concept, context fit.	Eligibility and synthesis boundary.	Controls topic creep into non-pharmacology AI.
Databases searched	PubMed, Scopus, Web of Science, IEEE Xplore, ScienceDirect, SpringerLink, Wiley, Oxford Academic, ACM Digital Library, Nature Portfolio, Frontiers, MDPI, and publisher records.	Captures biomedical, informatics, pharmacology, and computational literature.	Source of record and bibliographic eligibility.	Multidatabase search coverage.	Reduces single-database selection bias.
Search period	1 January 2017 to 9 July 2026.	Captures recent LLM and biomedical language-model literature.	Publication year and DOI status.	Time-bounded included set.	Excludes out-of-period sources.
Search strategy	Boolean combinations of LLM, biomedical language model, ChatGPT, RAG, pharmacology, drug discovery, repurposing, drug target, drug interaction, adverse event, pharmacovigilance, and decision support.	Matches the main review dimensions.	Search-topic alignment.	Candidate record pool.	Reduces irrelevant general AI retrieval.
Inclusion criteria	Peer-reviewed journal article, 2017–2026, DOI, LLM or biomedical language-model relevance, computational pharmacology or medication-safety relevance, extractable charting data.	Ensures all included studies can support scoping synthesis.	Article type, DOI, relevance, extractable data.	Final included evidence base.	Excludes weakly relevant or non-chartable items.
Exclusion criteria	Non-peer-reviewed, conference-only, preprint-only, no DOI, outside period, general AI without LLM relevance, diagnosis-only without drug relevance, insufficient detail.	Maintains scope and source quality.	Exclusion reason.	Counted full-text exclusion log.	Prevents unsupported source inclusion.
Evidence charting	Model category, pharmacology context, application domain, input source, output type, evaluation, validation, grounding, safety concern, finding, limitation.	Enables structured mapping across heterogeneous studies.	Required charting variables.	Scoping review charting matrix.	Avoids invented performance or validation claims.
Application mapping	Literature mining, NER, DDI extraction, DTI extraction, ADE extraction, pharmacovigilance, mechanistic reasoning, repurposing, evidence synthesis, decision support,	Organizes evidence by pharmacological function.	Domain and LLM function.	Application domain synthesis.	Prevents overemphasis on any single use case.

	human oversight, RAG support.				
Validation-gap mapping	Hallucination, unsupported citations, grounding weakness, outdated knowledge, prompt sensitivity, external validation, expert evaluation, clinical boundary, medication safety, bias, reproducibility, privacy, regulatory uncertainty.	Identifies safety-critical gaps across domains.	Limitation and mitigation need.	Validation and safety gap synthesis.	Prevents overclaiming readiness.
Synthesis approach	Narrative scoping synthesis organized by application domain, evidence cluster, validation pattern, and decision boundary.	Appropriate for heterogeneous scoping evidence.	Cross-study patterns and gaps.	Domain-by-domain synthesis.	Avoids causal, clinical, or regulatory conclusions not supported by charted evidence.

Knowledge extraction and literature mining

Knowledge extraction and literature mining formed one of the clearest evidence clusters in the mapped literature because pharmacological knowledge is often embedded in large volumes of biomedical text. BioBERT-based approaches to drug–drug interaction extraction show how biomedical language models can support relation extraction from text, but the scoping interpretation is that these systems identify candidate textual relations rather than establish clinically actionable interaction guidance [11].

Relation-focused architectures further illustrate how LLM-related biomedical NLP systems can be adapted to pharmacology-specific extraction tasks. A relation BioBERT model combined with bidirectional recurrent modeling was evaluated for drug–drug interaction extraction, supporting the claim that transformer-based representations can assist DDI relation classification while still depending on task-specific datasets and evaluation boundaries [12].

Drug–target relation extraction is another central computational pharmacology use case because target evidence is distributed across literature, molecular descriptions, gene descriptions, and biomedical databases. BERT-based mining of drug–target interactions from PubMed-scale literature supports large-scale evidence retrieval, but extracted drug–target links should be interpreted as literature-derived candidates that require curation and biological validation before mechanistic or therapeutic claims are made [13].

More recent ensemble transformer work using chemical and gene descriptions extends this extraction logic by combining biomedical literature signals with entity descriptions to improve drug–target interaction mining. This supports the scoping synthesis that source-enriched transformer systems may improve pharmacological relation extraction, while still leaving unresolved questions about external validation, false-positive relations, corpus bias, and the boundary between extracted associations and experimentally confirmed pharmacological interactions [14].

Adverse-event and pharmacovigilance extraction broaden the knowledge-extraction cluster from literature-based discovery toward medication safety. BERT-based adverse-event extraction from social media demonstrates potential pharmacovigilance support, but social media text introduces noise, reporting bias, ambiguity, privacy sensitivity, and source-verification challenges that prevent such outputs from being treated as validated safety signals without expert review [15].

Fine-tuned transformer frameworks for adverse drug reaction detection similarly show that biomedical language models can support safety-related classification and extraction tasks. In a scoping-review interpretation, these studies are important because they connect LLM-related methods to medication safety, but they also reinforce that benchmark performance and text classification do not independently establish pharmacovigilance causality, clinical severity, or regulatory actionability [16].

Clinical-note adverse-event extraction represents a more clinically embedded setting, yet the reviewed evidence indicates methodological heterogeneity across data sources, annotation schemes, extraction targets, and evaluation approaches. A systematic review of adverse drug event extraction from clinical notes supports the finding that this area remains fragmented and requires stronger reporting, external validation, and expert adjudication before LLM-assisted extraction can be relied upon in safety-critical clinical workflows [17].

Mechanistic reasoning and drug repurposing

Mechanistic reasoning and drug repurposing represented a second major evidence cluster, but the mapped literature showed that LLMs are primarily being used to organize

evidence, structure hypotheses, and connect heterogeneous information rather than to confirm pharmacological causality. A multisource prompt framework for drug repurposing supports the claim that prompted LLM workflows can assist candidate prioritization and rationale generation when multiple evidence sources are incorporated, but such outputs remain hypothesis-generating and require downstream validation [18].

Generative AI approaches to repurposing may become more useful when combined with real-world clinical evidence, disease-specific data, and explicit validation pathways. Work on prioritizing Alzheimer's disease drug repurposing candidates illustrates how generative AI can be linked to real-world validation logic, but the scoping interpretation remains cautious: candidate ranking and evidence-supported prioritization do not establish therapeutic efficacy or clinical utility by themselves [19].

Long chain-of-thought prompting has also been explored for drug repositioning knowledge extraction, suggesting that LLM reasoning strategies can help structure complex evidence trails across drugs, diseases, and mechanistic claims. However, chain-of-thought outputs should be treated as reasoning assistance rather than biological proof, because plausible mechanistic narratives may still contain unsupported links, incomplete evidence, or causal overextension [20].

Tool-using and chemically oriented LLM workflows extend mechanistic reasoning into chemical research and small-molecule discovery, but these systems remain bounded by source quality, tool reliability, human interpretation, and experimental confirmation. Autonomous chemical research workflows show that LLM-linked agents can coordinate literature, planning, and experimental or tool-mediated tasks, but autonomy should not be equated with validated pharmacological discovery [21]. Chemistry-tool augmentation demonstrates that external tools can help ground selected chemical outputs, while still requiring verification of tool choice, input validity, and interpretation [22]. Broader reviews of language models, multimodal LLMs, and small-molecule LLM applications position these systems across drug discovery and development, but they consistently require careful distinction between computational assistance and experimentally validated discovery [23]. Small-molecule LLM applications are therefore best understood as exploratory aids for design, prioritization, and evidence organization rather than as independent engines of validated molecular discovery [24]. A ChatGPT-based drug-discovery case study further supports the view that chatbot systems can assist exploratory workflows, but case-level workflow support should not be generalized into validated drug repurposing or therapeutic claims [25].

Table 2 maps LLM application domains in computational pharmacology, including knowledge extraction, mechanistic reasoning, drug repurposing, pharmacovigilance, and decision-support functions.

Table 2. Application Domains of Large Language Models in Computational Pharmacology: Knowledge Extraction, Literature Mining, Drug–Target Relation Extraction, Adverse-Event Extraction, Mechanistic Reasoning, Drug Repurposing, Pharmacovigilance, Evidence Synthesis, and Decision-Support Boundaries

Application domain	LLM function	Input data source	Output type	Validation approach	Key limitation	Safety or reliability concern	Scoping synthesis implication
Biomedical literature mining	Extract, summarize, and organize biomedical evidence	Abstracts, full-text articles, curated biomedical corpora	Evidence summaries and extracted statements	Benchmark evaluation and manual verification where reported	Coverage depends on corpus and retrieval strategy	Missing, outdated, or unsupported evidence	Literature mining is one of the clearest near-term support functions
Pharmacological named entity recognition	Identify drugs, targets, diseases, phenotypes, and adverse events	Biomedical text and annotated corpora	Entity spans and entity classes	Annotated benchmark testing	Entity ambiguity and synonym variation	Misclassification of drug or target terms	Entity recognition underpins higher-level pharmacology extraction
Drug–drug interaction extraction	Detect and classify	Biomedical literature and	Candidate DDI relation labels	Dataset-based relation	Dataset specificity and limited	False-negative or false-positive	Supports medication-safety evidence organization,

	interaction relations	relation datasets		extraction evaluation	external validation	interaction signals	not clinical advice
Drug–target relation extraction	Mine drug–target associations	PubMed-scale literature, chemical descriptions, gene descriptions	Candidate drug–target relations	Comparative model evaluation and curation needs	Extracted associations may not be experimentally confirmed	Mechanistic overinterpretation	Useful for hypothesis generation and evidence mapping
Adverse-event extraction	Extract ADE or ADR mentions	Social media, clinical notes, biomedical text	Adverse-event entities or classifications	Dataset evaluation and expert review where available	Noisy language, underreporting, and annotation heterogeneity	False safety signals or missed harms	Supports pharmacovigilance triage but requires safety review
Pharmacovigilance support	Structure and triage safety information	Safety narratives, adverse-event terms, drug names	Structured safety outputs and candidate signals	Guardrail testing and domain review	Safety-critical terminology is fragile	Wrong drug, event, seriousness, or causality interpretation	Requires guardrails and human pharmacovigilance oversight
Mechanistic reasoning assistance	Link drugs, targets, pathways, and disease mechanisms	Literature, pathways, knowledge bases, extracted relations	Mechanistic hypotheses and evidence chains	Expert review and experimental follow-up	Reasoning fluency may exceed evidence strength	False causal narratives	Best framed as evidence organization and hypothesis generation
Pathway reasoning assistance	Organize pathway-level evidence	Pathway literature, biological databases, extracted mechanisms	Pathway hypotheses	Biological validation required	Incomplete or outdated pathway context	Unsupported pathway claims	Useful for mapping plausible mechanisms, not proving them
Drug repurposing support	Prioritize existing drugs for new indications	Literature, real-world evidence, multi-source biomedical data	Ranked candidates and rationales	Preclinical, clinical, or real-world validation where available	Candidate ranking may be mistaken for efficacy	Premature therapeutic interpretation	Supports prioritization only within validation pipelines
Evidence synthesis	Combine extracted evidence across sources	Studies, relations, notes, retrieved documents	Evidence maps and summaries	Source checking and expert synthesis	Selective retrieval and summarization bias	Unsupported citations or missing counterevidence	Supports scoping synthesis and research planning
Medication information support	Answer drug-information or pharmacy questions	Medication questions, pharmacy scenarios, drug references	Answers, explanations, or recommendations	Pharmacist or expert review in evaluation studies	Uneven response quality	Incorrect, incomplete, or fabricated medication information	Requires pharmacist oversight and source verification

		where available					
Clinical or translational decision-support interface	Assist with medication-safety or translational review	Clinical vignettes, medication directions, EHR-like cases, retrieved evidence	Alerts, flags, or copilot suggestions	Expert-panel or task-specific evaluation	Often simulated or bounded evaluation	Automation bias and patient-safety risk	Should remain human-in-the-loop and bounded
Human-in-the-loop review	Present outputs for expert adjudication	Model outputs, sources, uncertainty flags	Reviewable recommendations or evidence summaries	Expert audit	Oversight design often underreported	False confidence if outputs appear authoritative	Essential governance layer across safety-critical domains
Retrieval-augmented pharmacology support	Ground answers in retrieved sources	Indexed literature, guidelines, databases, reference corpora	Source-linked answers and summaries	Retrieval relevance and answer-faithfulness testing	Retrieval may surface irrelevant or outdated sources	Source mismatch and citation unreliability	Promising grounding strategy, but not sufficient without validation

Decision-support applications

Decision-support applications formed a third evidence cluster, especially around medication information, medication-safety screening, clinical pharmacy support, and bounded clinical decision-support interfaces. LLM evaluation for preventing medication direction errors in online pharmacies supports the possibility that language models can assist medication-safety review under defined task conditions, but such systems should be interpreted as error-detection support rather than autonomous medication-safety authorities [26].

Clinical pharmacy and medication therapy management studies provide a more practice-facing view of LLM decision support. Evaluation of ChatGPT in clinical pharmacy and medication therapy management suggests that conversational LLMs may assist with structured pharmacy cases, but medication therapy decisions remain sensitive to patient context, dose interpretation, contraindications, comorbidities, and professional accountability [27].

Question-answering evaluations in pharmacy practice and medication information further show why decision-support claims must remain bounded. Assessment of ChatGPT for answering clinical pharmacy questions indicates that responses may be uneven and may include misleading or unreliable elements, supporting the need for source verification and pharmacist review [28]. Evaluation of ChatGPT as a medication-related information resource similarly supports caution because medication answers can be incomplete, context-dependent, or unsuitable for direct use without professional interpretation [29].

Retrieval-augmented and clinical decision-support systems may address some limitations by linking outputs to retrieved evidence and expert-evaluated workflows. A medication-safety study of a large language model as a clinical decision-support system supports the scoping conclusion that LLMs may augment medication safety when framed as a copilot across defined clinical specialties, but this does not remove the need for clinician oversight, accountability, validation, and careful management of safety-critical decision boundaries [30].

Scoping synthesis

Across the mapped literature, LLM applications in computational pharmacology clustered into four broad layers: biomedical language-model foundations, pharmacological knowledge extraction, reasoning and repurposing support, and decision-support interfaces. Retrieval-augmented generation emerged as a cross-cutting strategy for improving source grounding, but healthcare RAG evidence also shows that retrieval does not automatically solve answer correctness, citation faithfulness, or evaluation gaps [31].

The pharmacovigilance cluster highlighted safety-critical terminology, drug-name precision, adverse-event interpretation, and uncertainty handling as essential design requirements. Guardrail-focused work in pharmacovigilance supports the synthesis that LLMs require structured safeguards when used near medical safety settings, especially because wrong drug names, incorrect adverse-event terms, or unsupported safety interpretations can have downstream consequences [32].

The strongest mapped evidence was concentrated in bounded extraction and evaluation settings, including biomedical text mining, DDI extraction, DTI extraction, adverse-event extraction, medication-question evaluation, and medication-safety task support. By contrast, mechanistic reasoning, pathway interpretation, and drug repurposing were more often exploratory or framework-oriented, requiring careful language to avoid presenting evidence organization as causal proof. Broader assessments of medical LLM pitfalls support this distinction because hallucination, bias, factual inconsistency, and overreliance risks remain relevant when fluent outputs are introduced into scientific or clinical reasoning contexts [33].

Implementation readiness varied substantially across domains. Some studies reported benchmark or task-specific evaluation, some used expert review or real-world validation logic, and others were conceptual, case-based, or review-level contributions. Clinician-focused implementation guidance supports the synthesis that safe LLM use requires task-specific model selection, user training, oversight procedures, source verification, monitoring, and transparent communication of model limitations [34].

Taken together, the scoping evidence indicates that LLMs are most defensible as tools for organizing, extracting, retrieving, summarizing, and prioritizing pharmacological evidence, while their use for mechanism explanation, drug repurposing, and decision support must remain explicitly bounded. The main research priorities are pharmacology-specific hallucination benchmarks, citation-faithfulness audits, external validation across corpora and institutions, structured expert evaluation, reproducibility reporting for prompts and model versions, privacy-preserving deployment methods, and governance frameworks that keep human experts responsible for safety-critical interpretation.

Table 3 presents the scoping synthesis of evidence clusters, validation gaps, safety concerns, and research priorities for LLMs in computational pharmacology.

Table 3. Scoping Synthesis of Large Language Models in Computational Pharmacology: Evidence Clusters, Validation Gaps, Source-Grounding Needs, Hallucination Risk, Human Oversight, Decision-Support Boundaries, and Research Priorities

Synthesis theme	Mapped evidence cluster	Main contribution	Main limitation	Validation gap	Safety or reliability concern	Human oversight need	Research priority
Biomedical language-model foundations	BioBERT, PubMedBERT, BioGPT, medical LLMs	Establishes language-model capability for biomedical text tasks	Not always pharmacology-specific	Domain and task transfer validation	Corpus bias and overgeneralization	Biomedical NLP review	Pharmacology-specific benchmark suites
Knowledge extraction	NER, DDI extraction, DTI extraction, ADE extraction	Converts unstructured text into structured candidate evidence	Extracted relations may not be causal or complete	External validation and curator review	False-positive or false-negative evidence	Pharmacologist and curator adjudication	Shared extraction datasets and curation workflows
Literature mining and evidence retrieval	PubMed-scale mining, biomedical summarization, retrieved evidence	Helps organize large evidence spaces	Retrieval quality and coverage vary	Retrieval relevance and citation-faithfulness testing	Missing counterevidence or unsupported citations	Source verification	Evidence-grounded retrieval pipelines
Adverse-event extraction	Social media, clinical-note, and safety-text extraction	Supports pharmacovigilance triage and signal discovery	Noisy data and heterogeneous annotation	Clinical and pharmacovigilance validation	Misclassified adverse events or drugs	Pharmacovigilance expert review	Safety-critical ADE benchmarks
Pharmacovigilance support	Guardrails and safety-oriented	Adds structured safeguards	Early-stage and setting-	Prospective safety	Wrong drug or event coding	Safety scientist oversight	Guardrail validation standards

	LLM workflows	around drug-safety tasks	specific evidence	workflow validation			
Mechanistic reasoning	Pathway, target, and evidence-chain reasoning assistance	Organizes mechanistic hypotheses	Reasoning may sound causal without proof	Experimental and expert validation	Mechanistic overclaiming	Pharmacologist review	Mechanism-verification frameworks
Drug repurposing support	Prompt frameworks, candidate prioritization, case studies	Helps rank and explain candidate hypotheses	Candidate lists may be mistaken for discovery	Preclinical, clinical, or real-world validation	Premature therapeutic claims	Translational expert review	Prospective repurposing validation studies
Tool-augmented chemistry	LLM agents and chemistry-tool augmentation	Links language reasoning to tools and workflows	Tool choice and interpretation remain uncertain	Tool-output verification	False confidence from tool-mediated answers	Chemist and pharmacologist supervision	Auditable tool-use protocols
Medication information support	Pharmacy Q&A and medication-resource evaluations	Tests LLM responses in practical medication contexts	Response quality is uneven	Pharmacist-scored validation	Incorrect dosing, contraindication, or interaction information	Pharmacist review	Medication-information safety benchmarks
Decision-support interfaces	Medication-safety and clinical copilot studies	Explores bounded LLM assistance in safety workflows	Often vignette-based or task-specific	Real-world workflow validation	Automation bias and clinical overreliance	Clinician-in-the-loop governance	Prospective human-factors evaluation
Source grounding	Retrieval-augmented generation and evidence-linked outputs	May improve traceability	Retrieved sources may be irrelevant or outdated	Retrieval and answer-faithfulness audits	Citation mismatch	Expert source audit	Standardized grounding metrics
Reproducibility and transparency	Prompting, fine-tuning, model-version dependence	Identifies need for reporting standards	Prompt and model details often incomplete	Reproducibility testing	Non-replicable outputs	Audit and documentation	Prompt/model reporting guidelines
Safety governance	Hallucination, bias, privacy, uncertainty, regulatory uncertainty	Frames responsible use requirements	Governance evidence remains immature	Safety-case evaluation	Patient, research, and policy harm	Institutional oversight	Domain-specific LLM governance frameworks

CONCLUSION

This scoping review maps an emerging and heterogeneous literature on large language models in computational pharmacology across knowledge extraction, literature mining, drug–drug interaction extraction, drug–target relation extraction, adverse-event extraction, pharmacovigilance support, mechanistic reasoning assistance, drug repurposing hypothesis generation, medication-information support, and bounded decision-support applications. The mapped evidence suggests that LLMs can help organize complex pharmacological evidence and support hypothesis generation, but they should not be treated as independent validators of mechanisms, repurposing candidates, medication safety,

clinical recommendations, or regulatory decisions. Reliable use requires source grounding, transparent reporting, task-specific validation, external evaluation, expert oversight, uncertainty communication, privacy-aware workflows, and clearly defined decision boundaries that preserve human accountability in safety-critical pharmacology contexts.

Acknowledgments: None

Conflict of interest: None

Financial support: None

Ethics statement: None

REFERENCES

[1] Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. Nat Med. 2023;29(8):1930-40. doi:10.1038/s41591-023-02448-8

[2] Liu XH, Lu ZH, Wang T, Liu F. Large language models facilitating modern molecular biology and novel drug development. Front Pharmacol. 2024;15:1458739. doi:10.3389/fphar.2024.1458739

[3] Lu J, Choi K, Eremeev M, Gobburu J, Goswami S, Liu Q, et al. Large language models and their applications in drug discovery and development: A primer. Clin Transl Sci. 2025;18(4). doi:10.1111/cts.70205

[4] Zheng Y, Koh HY, Ju J, Yang M, May LT, Webb GI, et al. Large language models for drug discovery and development. Patterns (N Y). 2025;6(10):101346. doi:10.1016/j.patter.2025.101346

[5] 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

[6] Gu Y, Tinn R, Cheng H, Lucas MR, Usuyama N, Liu X, et al. Domain-specific language model pretraining for biomedical natural language processing. ACM Trans Comput Healthc. 2022;3(1):1-23. doi:10.1145/3458754

[7] Luo R, Sun L, Xia Y, Qin T, Zhang S, Poon H, et al. BioGPT: Generative pre-trained transformer for biomedical text generation and mining. Brief Bioinform. 2022;23(6). doi:10.1093/bib/bbac409

[8] Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, et al. Large language models encode clinical knowledge. Nature. 2023;620(7972):172-80. doi:10.1038/s41586-023-06291-2

[9] Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, et al. PRISMA extension for scoping reviews (PRISMA-ScR): Checklist and explanation. Ann Intern Med. 2018;169(7):467-73. doi:10.7326/M18-0850

[10] Munn Z, Peters MDJ, Stern C, Tufanaru C, McArthur A, Aromataris E. Systematic review or scoping review? Guidance for authors when choosing between a systematic or scoping review approach. BMC Med Res Methodol. 2018;18(1):143. doi:10.1186/s12874-018-0611-x

[11] Zhu Y, Li L, Lu H, Zhou A, Qin X. Extracting drug-drug interactions from texts with BioBERT and multiple entity-aware attentions. J Biomed Inform. 2020;106:103451. doi:10.1016/j.jbi.2020.103451

[12] KafiKang M, Hendawi A. Drug-drug interaction extraction from biomedical text using relation BioBERT with BLSTM. Mach Learn Knowl Extr. 2023;5(2):669-83. doi:10.3390/make5020036

[13] Aldahdooh J, Vähä-Koskela M, Tang J, Tanoli Z. Using BERT to identify drug-target interactions from whole PubMed. BMC Bioinformatics. 2022;23(1):245. doi:10.1186/s12859-022-04768-x

[14] Aldahdooh J, Tanoli Z, Tang J. Mining drug-target interactions from biomedical literature using chemical and gene descriptions-based ensemble transformer model. Bioinform Adv. 2024;4(1). doi:10.1093/bioadv/vbae106

[15] Dong F, Guo W, Liu J, Patterson TA, Hong H. BERT-based language model for accurate drug adverse event extraction from social media: Implementation, evaluation, and contributions to pharmacovigilance practices. Front Public Health. 2024;12:1392180. doi:10.3389/fpubh.2024.1392180

[16] Hussain S, Afzal H, Saeed R, Iltaf N, Umair MY. Pharmacovigilance with transformers: A framework to detect adverse drug reactions using BERT fine-tuned with FARM. Comput Math Methods Med. 2021;2021:5589829. doi:10.1155/2021/5589829

[17] Modi S, Kasmiran KA, Mohd Sharef N, Sharum MY. Extracting adverse drug events from clinical notes: A systematic review of approaches used. J Biomed Inform. 2024;151:104603. doi:10.1016/j.jbi.2024.104603

[18] Wei J, Zhuo L, Fu X, Zeng X, Wang L, Zou Q, et al. DrugReAlign: A multisource prompt framework for drug repurposing based on large language models. BMC Biol. 2024;22(1):226. doi:10.1186/s12915-024-02028-3

[19] Yan C, Grabowska ME, Dickson AL, Li B, Wen Z, Roden DM, et al. Leveraging generative AI to prioritize drug repurposing candidates for Alzheimer's

disease with real-world clinical validation. NPJ Digit Med. 2024;7:46. doi:10.1038/s41746-024-01038-3

[20] Kang H, Li J, Hou L, Xu X, Zheng S, Li Q. Large language model-enhanced drug repositioning knowledge extraction via long chain-of-thought: Development and evaluation study. JMIR Med Inform. 2025;13. doi:10.2196/77837

[21] Boiko DA, MacKnight R, Kline B, Gomes G. Autonomous chemical research with large language models. Nature. 2023;624(7992):570-8. doi:10.1038/s41586-023-06792-0

[22] Bran AM, Cox S, Schilter O, Baldassari C, White AD, Schwaller P. Augmenting large language models with chemistry tools. Nat Mach Intell. 2024;6(5):525-35. doi:10.1038/s42256-024-00832-8

[23] Chakraborty C, Bhattacharya M, Pal S, Chatterjee S, Das A, Lee SS. AI-enabled language models (LMs) to large language models (LLMs) and multimodal large language models (MLLMs) in drug discovery and development. J Adv Res. 2025;78:377-89. doi:10.1016/j.jare.2025.02.011

[24] Ma J, Liu J, Xu D, Song X, Zhang Z. Application and prospects of large language models in small-molecule drug discovery. Anal Chem. 2025;97(50):27453-77. doi:10.1021/acs.analchem.5c04083

[25] Wang R, Feng H, Wei GW. ChatGPT in drug discovery: A case study on anticocaine addiction drug development with chatbots. J Chem Inf Model. 2023;63(22):7189-209. doi:10.1021/acs.jcim.3c01429

[26] Pais C, Liu J, Voigt R, Gupta V, Wade E, Bayati M. Large language models for preventing medication direction errors in online pharmacies. Nat Med. 2024;30(6):1574-82. doi:10.1038/s41591-024-02933-8

[27] Roosan D, Padua P, Khan R, Khan H, Verzosa C, Wu Y. Effectiveness of ChatGPT in clinical pharmacy and the role of artificial intelligence in medication therapy management. J Am Pharm Assoc (2003). 2024;64(2):422-8.e8. doi:10.1016/j.japh.2023.11.023

[28] Munir F, Gehres A, Wai D, Song L. Evaluation of ChatGPT as a tool for answering clinical questions in pharmacy practice. J Pharm Pract. 2024;37(6):1303-10. doi:10.1177/08971900241256731

[29] Grossman S, Zerilli T, Nathan JP. Appropriateness of ChatGPT as a resource for medication-related questions. Br J Clin Pharmacol. 2024;90(10):2691-5. doi:10.1111/bcp.16212

[30] Ong JCL, Jin L, Elangovan K, Lim GYS, Lim DYZ, Sng GGR, et al. Large language model as clinical decision support system augments medication safety in 16 clinical specialties. Cell Rep Med. 2025;6(10):102323. doi:10.1016/j.xcrm.2025.102323

[31] Amugongo LM, Mascheroni P, Brooks S, Doering S, Seidel J. Retrieval augmented generation for large language models in healthcare: A systematic review. PLOS Digit Health. 2025;4(6). doi:10.1371/journal.pdig.0000877

[32] Hakim JB, Painter JL, Ramcharran D, Kara V, Powell G, Sobczak P, et al. The need for guardrails with large language models in pharmacovigilance and other medical safety critical settings. Sci Rep. 2025;15(1):27886. doi:10.1038/s41598-025-09138-0

[33] Omiye JA, Gui H, Rezaei SJ, Zou J, Daneshjou R. Large language models in medicine: The potentials and pitfalls: A narrative review. Ann Intern Med. 2024;177(2):210-20. doi:10.7326/M23-2772

[34] Li HY, Fu JF, Python A. Implementing large language models in health care: Clinician-focused review with interactive guideline. J Med Internet Res. 2025;27. doi:10.2196/71916