Pharmacophore

ISSN-2229-5402

Journal home page: http://www.pharmacophorejournal.com

LARGE LANGUAGE MODELS FOR PHARMACEUTICAL KNOWLEDGE MANAGEMENT: A CRITICAL REVIEW

Alejandro Torres[1]*, Miguel Fernandez[1]

1. ***Department of Intelligent Drug Systems, Faculty of Pharmacy, University of Chile, Santiago, Chile.***

ARTICLE INFO

Received:
03 August 2025
Received in revised form:
06 November 2025
Accepted:
08 November 2025
Available online:
28 December 2025

Keywords: Large language models, Pharmaceutical informatics, Retrieval-augmented generation, Hallucination, Regulatory intelligence, Pharmacovigilance

ABSTRACT

Large language models are increasingly being introduced into pharmaceutical knowledge management to support regulatory intelligence, safety surveillance, scientific literature review, and document search. Retrieval-augmented generation has become especially attractive because it promises to ground responses in approved labels, scientific publications, patents, and internal reports. Despite this enthusiasm, the evidence base remains uneven and fragmented. Current systems often demonstrate impressive linguistic fluency, yet fluency is frequently mistaken for factual reliability, domain competence, and regulatory readiness. This critical review evaluates the use of large language models and retrieval-augmented generation in pharmaceutical knowledge management. It focuses on hallucination control, domain-specific evaluation, trustworthiness, and the conditions required for safe deployment in regulated workflows. Retrieval-augmented generation reduces some factual errors but does not eliminate hallucination, source misuse, or incomplete reasoning. Evaluation methods remain immature, with many studies relying on metrics that do not adequately measure pharmaceutical correctness, completeness, or actionability. Unverified outputs from large language models may create risks for patient safety, pharmacovigilance, regulatory compliance, and internal decision-making. Responsible deployment requires expert oversight, traceable sources, robust evaluation, and explicit governance rather than confidence in model fluency alone. Large language models may become valuable tools for pharmaceutical knowledge work, but they should not yet be treated as autonomous knowledge authorities. The field must build pharmaceutical-specific benchmarks, stronger fact-checking protocols, and auditable governance frameworks before these systems can be trusted in high-stakes contexts.

To Cite This Article: Torres A, Fernandez M. Large Language Models for Pharmaceutical Knowledge Management: A Critical Review. Pharmacophore. 2025;16(6):12-21. https://doi.org/10.51847/XPhOzUYFSY

Introduction

The pharmaceutical sector produces and depends on vast quantities of unstructured and semi-structured text, including regulatory labels, safety reports, clinical trial documents, scientific articles, patents, and internal technical memoranda. Early transformer architectures created the technical foundation for large language models by enabling scalable attention over long textual contexts [1], while biomedical adaptations such as BioBERT and SciBERT demonstrated that domain pretraining could improve extraction and classification over biomedical and scientific corpora [2, 3]. These developments have encouraged pharmaceutical organizations to view language models as tools for converting fragmented document repositories into searchable and conversational knowledge systems. The promise is substantial, but it rests on the assumption that linguistic pattern recognition can be translated into dependable knowledge management.

The pace of adoption has outstripped the maturity of evidence in several pharmaceutical settings. Domain-specific models such as GatorTron, BioGPT, PharmBERT, and RxBERT show that biomedical and drug-label language can be modeled more effectively than with general systems alone [4-8], yet adaptation to terminology does not automatically guarantee regulatory correctness or safety relevance. Applications such as adverse event extraction, drug label analysis, and automated comparison of safety wording illustrate real utility [9-16], but they also expose the fragility of models when labels are ambiguous, incomplete, or modified over time. The central problem is that many deployments treat language models as accelerators of expert work before adequately defining the boundaries of acceptable error.

This disconnect is especially important because pharmaceutical knowledge management is not merely an information retrieval problem. A small error in dose interpretation, contraindication wording, drug-drug interaction reasoning, or adverse event classification may have consequences for patient safety, signal detection, or regulatory submissions [17, 18]. Retrieval-

Corresponding Author: Alejandro Torres; Department of Intelligent Drug Systems, Faculty of Pharmacy, University of Chile, Santiago, Chile. E-mail: alejandro.torres@gmail.com.

augmented systems promise to reduce reliance on parametric memory by grounding outputs in documents [19-22], but grounding is not equivalent to truth when retrieved passages are irrelevant, misread, or selectively summarized. The pharmaceutical context therefore demands a higher standard than general usefulness: it requires traceable, complete, and verifiable claims under conditions of uncertainty.

This review critically examines large language models and retrieval-augmented generation as tools for pharmaceutical knowledge management rather than presenting them as a finished technological solution. It evaluates the literature on model adaptation, document retrieval, hallucination mitigation, benchmark design, and regulatory trust, drawing on biomedical language modeling, pharmacovigilance, clinical summarization, and safety-critical AI studies [23]. The focus is thematic and evaluative rather than systematic; it does not claim exhaustive coverage of every model or dataset. Instead, it asks whether current evidence is strong enough to justify confidence in LLM-based pharmaceutical knowledge tools, and where the field must improve before operational deployment becomes defensible.

The Landscape of LLMs in Pharmaceutical Knowledge Management

Core Application Areas

Current applications cluster around regulatory intelligence, pharmacovigilance document processing, drug label interpretation, scientific literature synthesis, and extraction from trial or safety documents. Drug labeling has received particular attention because structured product labels contain authoritative but linguistically complex information about indications, warnings, dosing, adverse reactions, and interactions [7, 8, 11-16]. Pharmacovigilance studies show that language models can help identify adverse event content and compare label changes [9, 10, 17], but these tasks remain vulnerable to temporal drift when labels are revised and to semantic ambiguity when adverse reactions are described using overlapping clinical terms. The literature therefore supports LLMs as accelerators for information triage, but not yet as independent arbiters of pharmaceutical meaning.

LLM Architectures and Adaptation Strategies

The architectural trajectory begins with transformer-based pretraining [1] and extends through biomedical domain adaptation, including BioBERT, SciBERT, PubMedBERT, BioGPT, GatorTron, PharmBERT, and RxBERT [2-8]. These models demonstrate that pretraining on biomedical literature, clinical notes, or drug labeling can improve lexical and contextual fit, particularly where terminology differs from general language. However, model adaptation often improves task performance without solving the deeper problem of factual accountability: a fine-tuned system may classify labels more effectively while still being unable to explain uncertainty, identify missing evidence, or recognize when a question exceeds the scope of the corpus [7, 8, 12]. The field has therefore been too quick to equate domain adaptation with domain trustworthiness.

RAG as a Preferred Paradigm

Retrieval-augmented generation has become the favored paradigm because pharmaceutical knowledge is distributed across controlled documents, updated repositories, and proprietary archives that cannot be fully encoded in model parameters. Biomedical RAG studies show that grounding generation in retrieved sources can improve access to clinical and scientific information [19-22, 24, 25], while tool-augmented systems suggest a path toward connecting models with domain databases rather than relying only on pretrained memory [23]. This architecture is particularly appealing for internal R&D documents and regulatory archives because it can, in principle, preserve source traceability. Yet RAG also shifts the failure point from generation alone to the entire retrieval, ranking, chunking, and synthesis pipeline, making system validation more complex rather than simply safer.

Retrieval-Augmented Generation: Promise and Pitfalls

How RAG is Implemented in Pharma

In pharmaceutical settings, RAG typically involves parsing source documents into chunks, embedding those chunks, retrieving semantically similar passages, and passing the retrieved text to a generative model for synthesis. Biomedical implementations have applied this pattern to scientific literature, electronic health records, clinical guidelines, and knowledge graphs [19-25], and the same logic is increasingly applied to labels, regulatory correspondence, patents, and internal reports. The critical design choices are not trivial: chunk size may separate a warning from its qualifying condition, embeddings may miss rare drug names, and ranking algorithms may privilege semantic similarity over regulatory relevance. A pharmaceutical RAG system is therefore only as reliable as its document preprocessing, metadata governance, and retrieval validation.

Table 1 summarizes the main retrieval-design risks that can affect pharmaceutical RAG reliability and links each risk to a practical validation control.

Table 1. Key Design Risks and Validation Controls in Pharmaceutical RAG Systems

RAG design element	Pharmaceutical reliability risk	Practical validation control
Document chunking	Safety warnings, contraindications, or regulatory qualifications may be split across separate chunks.	Test whether retrieved chunks preserve complete clinical or regulatory meaning.

Embedding model	Rare drug names, formulation terms, abbreviations, or sponsor-specific terminology may be poorly represented.	Evaluate retrieval using pharmaceutical synonym sets, brand/generic names, and rare entity queries.
Retrieval ranking	Semantically similar passages may be ranked above legally or clinically more relevant source text.	Compare ranking outputs against expert-curated relevance judgments.
Metadata governance	Source date, jurisdiction, version, document type, or approval status may be lost during indexing.	Require metadata filters for label version, regulatory region, document class, and update date.
Retrieved-to-generated synthesis	The generative model may overgeneralize, omit uncertainty, or blend evidence across incompatible sources.	Audit outputs against retrieved passages and require citation-linked answer verification.

Benefits over Standalone LLMs

Compared with standalone LLMs, RAG offers clear conceptual advantages for pharmaceutical knowledge management because it reduces dependence on static parametric knowledge and can expose the sources used to generate an answer. Studies of retrieval-augmented biomedical systems show that external context can improve response grounding and support more transparent information access [19-22, 24, 25]. In regulated knowledge work, the ability to cite an approved label, protocol, or scientific article is not merely a convenience but a prerequisite for auditability. Nevertheless, citation display can create a false sense of assurance if the cited passage does not actually support the generated claim or if the system omits conflicting evidence.

Evidence of Residual Hallucination

The strongest critique of RAG is that retrieval reduces hallucination without eliminating it. Biomedical evaluations show that LLMs can still produce unsupported or misleading claims even when external evidence is supplied [19-22, 26]. In pharmaceutical tasks, this residual risk is amplified because generated answers may compress complex label language into overconfident statements, conflate related adverse events, or overlook qualifiers such as population restrictions and monitoring conditions [9-18]. The practical implication is uncomfortable but unavoidable: source-grounded generation cannot be accepted as factual unless the relationship between answer and evidence is itself evaluated.

RAG-Specific Failure Modes

RAG introduces distinctive failure modes that are sometimes obscured by the broader discussion of hallucination. Retrieval can fail because the relevant document is missing, poorly indexed, embedded with insufficient domain sensitivity, or ranked below less relevant passages [21, 22]. Even when retrieval succeeds, the generator may misuse sources, blend incompatible passages, over-summarize uncertainty, or ignore context outside the window [19, 24, 25]. In pharmaceutical organizations, the additional “garbage in, garbage out” problem is severe because internal documents may include drafts, superseded regulatory positions, non-final interpretations, or inconsistent terminology that a fluent model may synthesize without recognizing their status.

Hallucination Control: Techniques and Their Limits

Current Mitigation Strategies

Common mitigation strategies include prompt engineering, explicit source citation, self-consistency, retrieval grounding, chain-of-verification, fact-checking modules, and post-generation human review. Medical hallucination benchmarks and factuality frameworks have helped distinguish fluent but unsupported outputs from claims that can be verified against evidence [26]. However, most mitigation methods operate probabilistically rather than deterministically, and they often depend on the same model family to critique or verify its own outputs. This creates a circularity problem: a model that lacks domain understanding may not reliably detect the very pharmaceutical errors it generates.

Domain-Specific Challenges

Pharmaceutical language is unusually unforgiving because clinically meaningful distinctions may depend on small textual differences. A model must distinguish indication from off-label use, warning from contraindication, adverse event association from causality, and dosage recommendation from administration constraint [11-18]. Domain-specific models trained on labels and biomedical corpora improve terminology handling [7, 8, 11, 12], yet subtle semantic distinctions remain difficult when the relevant evidence spans tables, footnotes, population qualifiers, and evolving label versions. Hallucination control in this domain must therefore address not only invented facts but also oversimplification, missing qualifiers, and inappropriate generalization.

Success Rates and Residual Risk

Published studies often show performance improvements from biomedical pretraining, label-specific adaptation, or retrieval augmentation [2-8, 19-25], but these improvements should not be interpreted as proof that systems are safe enough for autonomous pharmaceutical decision-making. Studies on medical hallucination, biomedical factuality, and safety-critical summarization emphasize that residual errors persist even in advanced models [26]. The difficulty is that a low aggregate error rate may still be unacceptable when the remaining errors concern contraindications, safety signals, or regulatory commitments. Pharmaceutical evaluation must therefore focus less on average performance and more on the severity, detectability, and downstream consequences of rare failures.

The Need for Human-in-the-Loop Validation

Human-in-the-loop validation remains essential because current LLMs cannot reliably determine when their outputs are incomplete, outdated, or insufficiently supported. Pharmacovigilance and medication safety studies illustrate that expert interpretation is needed to assess whether extracted or generated content is clinically and regulatory meaningful [17, 18]. Expert review is not merely a final quality check; it should shape corpus curation, retrieval design, prompt constraints, error taxonomies, and escalation workflows. The field's challenge is to design human oversight that is scalable and auditable rather than symbolic.

Table 2 defines the major trustworthiness failure modes of pharmaceutical LLM systems and links each failure mode to domain-specific evaluation criteria and deployment controls.

Table 2. Pharmaceutical LLM Trustworthiness Failure Modes, Evaluation Criteria, and Deployment Controls

Trustworthiness domain	Pharmaceutical failure mode	Why it matters in regulated knowledge work	Evaluation criterion	Required control before deployment
Source retrieval	Relevant approved label, safety document, or trial evidence is not retrieved	The answer may appear grounded while excluding the most authoritative or current source	Retrieval recall for current, authoritative, task-relevant documents	Version-controlled source repositories, metadata validation, and domain-specific retrieval testing
Source ranking	Less relevant but semantically similar passages are ranked above regulatory-relevant evidence	Semantic similarity may obscure jurisdiction, document status, or label authority	Ranking quality by authority, recency, jurisdiction, and task relevance	Regulatory-aware ranking rules and expert-reviewed retrieval validation sets
Chunking and preprocessing	Warning, qualifier, dosage condition, or population restriction is separated from its context	A technically correct passage may become unsafe when interpreted without its limiting condition	Qualifier preservation across chunk boundaries	Chunking protocols designed around label sections, tables, footnotes, and warning structures
Generation	The LLM produces fluent but unsupported or partially supported claims	Fluency may be mistaken for factual reliability, especially in safety-sensitive domains	Claim-level source support and contradiction checking	Evidence-to-claim mapping, citation verification, and refusal when support is insufficient
Citation behavior	The model cites a source that does not actually support the generated statement	Displayed citations may create false assurance and weaken auditability	Citation faithfulness and passage-to-claim alignment	Citation-level review, automated evidence matching, and human verification for high-risk outputs
Pharmaceutical semantics	The system confuses indication, off-label use, warning, contraindication, adverse event association, or causality	Small semantic differences can affect patient safety, regulatory interpretation, or pharmacovigilance decisions	Domain-specific correctness judged against pharmaceutical expert rubrics	Expert-designed semantic error taxonomy and calibrated reviewer panels
Completeness	The answer omits relevant contraindications, monitoring requirements, conflicting evidence, or uncertainty	Incomplete answers may be more dangerous than visibly uncertain ones	Completeness against predefined evidence requirements	Checklist-based completeness review and targeted testing on high-risk cases
Temporal validity	The answer relies on outdated labels, superseded safety language, or draft internal documents	Pharmaceutical knowledge changes over time and must remain version-specific	Current-source verification and document-status awareness	Source-date controls, document-status labels, and retrieval exclusion of superseded sources
Uncertainty handling	The model gives an overconfident answer when evidence is incomplete or ambiguous	Users may over-trust outputs that should trigger expert review or escalation	Calibration of uncertainty, refusal quality, and escalation appropriateness	Mandatory uncertainty statements and escalation pathways for ambiguous evidence
Workflow actionability	Output is linguistically correct but not usable for regulatory intelligence, safety review, or medical information workflows	Knowledge tools must support professional decisions, not merely generate plausible summaries	Task-specific actionability and expert acceptability	Workflow-specific validation, user-role controls, and documented human approval

Domain Evaluation: The Achilles' Heel of Pharma LLMs

Existing Benchmarks and Their Inadequacy

General benchmarks cannot adequately capture pharmaceutical knowledge work because they rarely test the exact forms of reasoning required for labels, safety narratives, regulatory correspondence, or drug-drug interaction interpretation. Biomedical benchmarks such as those used for clinical knowledge and medical hallucination provide useful starting points, but they do not fully evaluate source fidelity, version control, jurisdictional differences, or the consequences of misinterpreting regulatory text. Label-focused datasets and extraction tasks offer stronger domain alignment [7, 8, 11-16], yet they remain narrower than

the workflows in which organizations want to deploy LLMs. The gap between benchmark competence and operational reliability is therefore one of the field's central unresolved problems.

The Over-Reliance on Automated Metrics

Automated metrics such as overlap scores, classification accuracy, and extraction F-measures are useful for development but weak proxies for pharmaceutical correctness. A summary can achieve lexical similarity while omitting a contraindication, changing the strength of a warning, or failing to preserve a population qualifier [26]. Studies of biomedical text generation and clinical summarization increasingly recognize that factuality, safety, and expert acceptability require evaluation beyond generic NLP metrics [24, 26]. Pharmaceutical LLM research should therefore reduce its dependence on surface-level metrics and evaluate whether outputs are accurate, complete, source-faithful, and actionable for the intended professional task.

Human Expert Evaluation

Human expert evaluation remains the most credible method for judging pharmaceutical outputs, but it is expensive, slow, and inconsistently reported. Expert review is especially important in pharmacovigilance and medication safety because the validity of an output depends on clinical context, regulatory wording, and the seriousness of potential downstream harm [17, 18]. The weakness of the current literature is not that expert evaluation is absent everywhere, but that it is often small-scale, poorly standardized, and difficult to compare across studies. Without shared rubrics and inter-reviewer calibration, expert evaluation can become anecdotal rather than a reproducible standard.

Toward Better Evaluation

A stronger evaluation framework should assess factual accuracy, source fidelity, completeness, uncertainty communication, safety impact, and workflow actionability rather than treating answer generation as an isolated NLP task. Biomedical factuality frameworks and RAG benchmarks provide useful foundations [21, 22, 26], while pharmaceutical label datasets and safety applications indicate where domain-specific criteria must be added [7-18]. Evaluation should also test adversarial and failure-prone cases, including outdated labels, conflicting evidence, rare adverse reactions, ambiguous drug names, and incomplete internal documents. Until such frameworks become routine, claims that LLMs are "ready" for pharmaceutical knowledge management will remain premature.

Trust, Accountability, and Regulatory Readiness

The Trust Deficit

The trust deficit around pharmaceutical LLMs is justified because current systems can sound authoritative while remaining uncertain, incomplete, or wrong. Clinical knowledge models have shown impressive competence on medical question-answering tasks, yet safety-critical studies warn that benchmark performance does not automatically translate into dependable use in real professional settings. In pharmaceutical knowledge management, this distinction is especially important because users may over-trust fluent explanations about drug safety, label interpretation, or regulatory evidence. Trust should therefore be earned through transparent validation, calibrated uncertainty, and demonstrated failure handling rather than inferred from general model capability.

Accountability and Traceability

Accountability requires more than displaying citations beside generated text. RAG systems must preserve the relationship between retrieved passages, generated claims, document versions, user prompts, and subsequent human edits [19-22]. Pharmacovigilance and label-comparison tools illustrate why traceability matters: safety conclusions may depend on whether the system consulted the current label, an older label, a draft update, or an incomplete source [9, 10, 14-16]. A defensible pharmaceutical LLM workflow must therefore log retrieval decisions, source provenance, model versions, reviewer actions, and final approved outputs in a manner suitable for audit.

Regulatory Signals

Regulatory readiness should not be confused with technical plausibility. Although LLMs may assist with drafting, summarization, and evidence retrieval, their use in regulated submissions or labeling workflows would require documented validation, source control, reproducibility, and human accountability [10]. The literature on drug labeling, adverse event extraction, and medication direction safety shows that models can support expert workflows [8-12], but it does not establish that generated content can be accepted without independent verification. For regulatory use, the key question is not whether an LLM can produce a plausible answer, but whether the answer can be traced, reproduced, challenged, and corrected under formal quality systems.

Beyond RAG – Emerging Architectures and Alternative Paradigms

Knowledge-Graph-Enhanced LLMs

Knowledge-graph-enhanced LLMs are attractive because pharmaceutical knowledge is highly relational: drugs connect to targets, indications, adverse events, contraindications, formulations, trial populations, and regulatory histories. KRAGEN demonstrates how biomedical problem-solving can be supported by combining retrieval with structured knowledge graphs

[22], while tool-augmented biomedical systems such as GeneGPT show the value of linking language models to domain resources rather than relying solely on internal parameters [23]. For pharmaceutical applications, knowledge graphs may reduce certain hallucinations by constraining outputs to explicit relationships. However, they also introduce new risks when graph coverage is incomplete, relations are outdated, or uncertain evidence is represented as if it were definitive.

Multi-Agent and Tool-Using LLMs

Tool-using LLMs offer a promising alternative to monolithic generation because they can delegate tasks to calculators, databases, search systems, chemical resources, or pharmacovigilance modules. GeneGPT is an early example of using domain tools to improve biomedical information access [23], and RAG systems in healthcare show how retrieval can be combined with generation to support more grounded responses [20, 21, 24, 25]. In pharmaceutical workflows, tool use could help verify doses, identify drug-drug interactions, retrieve label versions, or compare regulatory requirements. Yet multi-agent architectures can also obscure responsibility, because an error may arise from the planner, the tool, the retrieved data, the synthesis step, or the handoff among agents.

Symbolic-Neuro Hybrid Systems

Symbolic-neuro hybrid systems may be better suited than unconstrained generation for safety-critical pharmaceutical constraints. Rule engines can enforce deterministic checks for contraindications, dose limits, mandatory warnings, or document-status restrictions, while LLMs can support search, summarization, and explanation [11, 12, 18]. This division of labor is appealing because it reserves hard safety boundaries for systems that are auditable and predictable. The limitation is that hybrid systems require careful knowledge engineering, maintenance, and validation; without sustained governance, symbolic rules can become stale while neural components continue to generate persuasive but unsupported interpretations.

Cross-Domain Lessons From Clinical and Legal AI

Clinical NLP and Medical Evidence Synthesis

Clinical NLP offers useful lessons because it has confronted similar tensions between automation, evidence synthesis, and patient safety. Clinical and biomedical LLM studies show that performance gains are real [4, 6, 24], but hallucination and safety evaluations emphasize that expert review remains essential when outputs could influence care [26]. Pharmaceutical knowledge systems should adopt the clinical lesson that factual correctness is task-specific, context-dependent, and not adequately captured by general fluency. The most transferable principle is that LLMs should be treated as decision-support tools embedded in verification workflows, not as autonomous authorities.

Legal AI and Document Review

Legal AI is relevant because pharmaceutical knowledge management also involves sensitive documents, professional accountability, version control, and high consequences for misinterpretation. RAG-style document review resembles legal retrieval workflows in that an answer must be grounded in specific sources, but pharmaceutical systems must additionally respect scientific uncertainty, safety evidence, and regulatory jurisdiction [19-22]. The lesson from document-intensive professional domains is that audit trails, privilege-like access controls, and clear responsibility boundaries matter as much as model accuracy. Pharmaceutical organizations should therefore evaluate LLMs not only as NLP systems, but as components of controlled professional processes.

A Critical Assessment Framework for Pharmaceutical Knowledge LLMs

Proposed Dimensions

A credible assessment framework for pharmaceutical LLMs should include factual accuracy, source fidelity, completeness, safety relevance, domain specificity, uncertainty communication, and workflow actionability. Existing factuality and hallucination frameworks provide a foundation for evaluating whether outputs are supported by evidence [26], while RAG benchmarks highlight robustness, self-awareness, and retrieval dependence [20-22]. Pharmaceutical label and safety datasets add domain-specific pressure points, including adverse event classification, label changes, and drug-drug interaction interpretation [7-18]. The framework should also distinguish harmless wording imperfections from errors that could affect clinical judgment, regulatory interpretation, or patient safety.

Figure 1 illustrates the trust-to-regulatory-readiness architecture required before LLM and RAG systems can be responsibly used in pharmaceutical knowledge management.

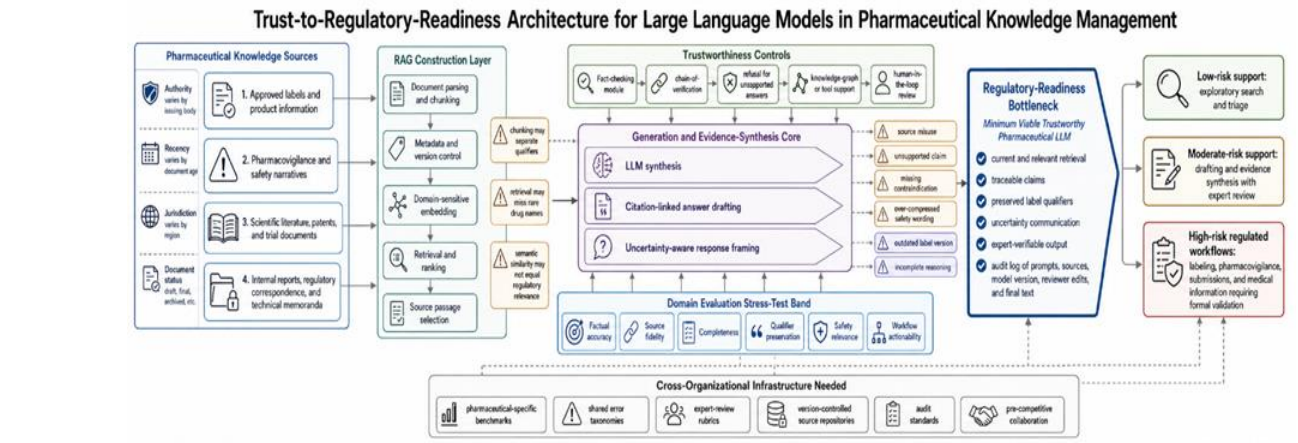

Figure 1. Trust-to-Regulatory-Readiness Architecture for Large Language Models in Pharmaceutical Knowledge Management

Minimum Viable Trustworthy LLM for Pharma

A minimum viable trustworthy pharmaceutical LLM should demonstrate more than high benchmark scores. It should retrieve current and relevant sources, preserve qualifiers from labels and scientific evidence, communicate uncertainty, refuse unsupported answers, and produce outputs that experts can efficiently verify [19-22, 26]. It should also maintain logs of inputs, retrieved evidence, generated claims, reviewer decisions, and final approved text so that accountability does not disappear after deployment [9, 10]. Without these capabilities, the system may still be useful for exploratory search or drafting, but it should not be considered reliable for operational knowledge decisions.

A Call for Pre-Competitive Collaboration

The field needs pre-competitive collaboration because no single organization can adequately solve benchmarking, safety evaluation, and governance for pharmaceutical LLMs in isolation. Public resources such as annotated drug labels and adverse event datasets show the value of shared infrastructure [14-16], while domain-specific models demonstrate how open scientific and biomedical corpora can accelerate progress [2-8]. Independent evaluation would be particularly valuable for testing RAG systems across label versions, safety narratives, regulatory documents, and conflicting evidence. Without shared benchmarks and audit standards, the literature will remain difficult to compare and vendors will be able to claim readiness using narrow or opaque evaluations.

Future Directions and Grand Challenges

Toward Zero-Hallucination Pharmaceutical LLMs

The aspiration of zero-hallucination pharmaceutical LLMs is understandable but technically and conceptually difficult. Hallucination is not only fabrication; in pharmaceutical contexts it also includes missing qualifiers, overgeneralizing indications, compressing uncertainty, or citing evidence that does not support the claim [26]. RAG, knowledge graphs, and tool use can reduce these risks [20-23], but none guarantees complete factual reliability when source documents are incomplete, inconsistent, or outdated. Future research must therefore shift from claiming hallucination reduction to formally characterizing residual error types and their safety consequences.

Scalable Human-Expert Evaluation

Scalable expert evaluation is one of the most urgent unresolved problems. Human review is necessary for judging clinical relevance, regulatory wording, and safety implications [17, 18], but it is expensive and difficult to standardize across organizations and therapeutic areas. A practical path forward may combine expert-designed rubrics, calibrated reviewer panels, targeted sampling of high-risk outputs, and automated pre-screening for unsupported claims [26]. The key challenge is to make expert oversight reproducible and auditable rather than relying on informal professional intuition after the system has already generated a plausible answer.

Integration with Pharmaceutical Workflows

Integration with pharmaceutical workflows will determine whether LLMs create durable value or merely add another layer of technical complexity. Labeling, pharmacovigilance, medical information, regulatory intelligence, and R&D knowledge search each have different tolerance for error, different evidence standards, and different documentation requirements [8-18]. RAG systems must therefore be configured around specific workflows rather than deployed as general chat interfaces over document repositories [19-22, 24, 25]. The most realistic future is not full automation, but carefully bounded human-AI collaboration in which LLMs accelerate search and synthesis while experts retain responsibility for interpretation and approval.

Table 3 maps pharmaceutical knowledge-management workflows to appropriate LLM roles, validation burdens, human oversight requirements, and deployment boundaries.

Table 3. Workflow-Specific Governance and Validation Architecture for Pharmaceutical LLM Deployment

Pharmaceutical workflow	Appropriate LLM/RAG role	Evidence sensitivity	Primary validation burden	Human oversight requirement	Deployment boundary
Exploratory literature search	Retrieve, cluster, and summarize scientific literature for early orientation	Moderate; errors may misdirect research interpretation but are usually reviewable	Source coverage, retrieval relevance, summary faithfulness, and omission detection	Researcher review before any scientific conclusion is accepted	Suitable for low-risk discovery support, not for final evidence claims
Regulatory intelligence	Track, compare, and summarize regulatory documents, guidance, correspondence, and jurisdiction-specific changes	High; outdated or jurisdictionally incorrect information may affect compliance decisions	Version control, jurisdiction tagging, citation fidelity, and audit-log reproducibility	Regulatory expert approval required before action	Use as decision-support only within controlled review workflows
Pharmacovigilance triage	Assist with adverse event extraction, case narrative review, duplicate detection, and signal-relevant text organization	Very high; missed or misclassified safety information may affect patient safety and signal detection	Recall for safety-relevant content, seriousness classification, causality wording, and false-negative analysis	Pharmacovigilance specialist review mandatory	May support triage but should not autonomously determine safety conclusions
Drug-label interpretation	Answer questions about indications, contraindications, dosing, warnings, adverse reactions, and interactions	Very high; small wording changes may alter clinical or regulatory meaning	Qualifier preservation, label-section awareness, current-label retrieval, and claim-level verification	Medical, regulatory, or labeling expert review required	Only bounded use with current approved sources and traceable evidence
Medical information response drafting	Draft responses to professional or patient-facing inquiries using approved information	Very high; inaccurate wording may influence clinical understanding or medication use	Approved-source restriction, response completeness, citation support, and audience-appropriate wording	Final approval by qualified medical information staff	Drafting support only; no autonomous outbound communication
R&D knowledge management	Search internal reports, patents, trial documents, and technical memoranda for synthesis and hypothesis support	Moderate to high; source status, confidentiality, and incomplete evidence are major risks	Access control, document-status labeling, retrieval precision, and uncertainty communication	Subject-matter expert review before scientific or strategic decisions	Useful for synthesis and triage when document governance is strong
Regulatory submission support	Assist with drafting, cross-checking, or locating evidence for submission documents	Extremely high; reproducibility, traceability, and accountability are mandatory	Full audit trail, source-to-claim mapping, model-version control, and formal validation	Formal quality-system review and accountable human sign-off	Not suitable without validated, auditable, governed deployment
Internal policy and SOP search	Retrieve controlled procedural knowledge for operational staff	High; outdated or incorrect SOP interpretation may cause process noncompliance	Current-document retrieval, superseded-document exclusion, and exact procedural wording preservation	Process owner review for high-impact procedures	Suitable only when connected to controlled document management systems

Strengths and Limitations of This Review

Strengths

This review's principal strength is its focused critique of the three issues most consequential for pharmaceutical adoption: retrieval-augmented generation, hallucination control, and domain evaluation. By linking foundational language-model research [1-6] with drug-label modeling, pharmacovigilance, safety evaluation, and biomedical RAG studies [7], it examines LLMs as socio-technical systems rather than isolated algorithms. The review also emphasizes failure modes that are sometimes underplayed in promotional discussions, including source misuse, missing qualifiers, weak benchmark validity, and premature trust. This focus is appropriate because pharmaceutical knowledge management is a regulated, evidence-sensitive activity in which plausible language is not sufficient.

Limitations

This review is qualitative and thematic, so it does not provide the exhaustive coverage or formal evidence grading expected from a systematic or scoping review. It reflects the state of the peer-reviewed literature available through 2025, and the rapid movement of LLM research means that some relevant methods may appear first as preprints, vendor reports, or internal validation studies not captured here [20, 21]. Another limitation is that the pharmaceutical LLM literature remains uneven, with stronger evidence for label extraction and biomedical question-answering than for proprietary R&D knowledge management or regulatory submission workflows [7-18]. Consequently, several conclusions are necessarily cautionary: they

identify what must be demonstrated before deployment rather than claiming that the evidence already proves safety or readiness.

Conclusion

LLMs hold transformative potential for pharmaceutical knowledge management, but the field is at an inflection point where technical capability has outpaced evidence of safety and reliability. Their ability to search, summarize, and synthesize complex text could improve regulatory intelligence, safety surveillance, and scientific decision-support. Yet this promise will only be realized if the field resists the temptation to treat fluent generation as verified knowledge.

RAG architectures address some hallucination risks by grounding responses in external documents, but they do not eliminate residual errors, weak retrieval, source misuse, or incomplete reasoning. The most serious barrier is not the absence of impressive demonstrations, but the absence of mature evaluation frameworks that reflect the demands of pharmaceutical correctness. A system that performs well on a benchmark may still be unsafe if it mishandles a label qualifier, omits a safety warning, or cites evidence inaccurately.

A coordinated effort is urgently needed to develop pharmaceutical-specific benchmarks, rigorous fact-checking protocols, transparent audit mechanisms, and reproducible expert evaluation. These foundations should be built collaboratively across academia, industry, regulators, and standards bodies. Without such shared infrastructure, the field will remain fragmented, and claims of trustworthiness will be difficult to verify.

Until these foundations are established, LLMs should be deployed cautiously and within clearly bounded workflows. They are best understood as assistants for retrieval, triage, drafting, and synthesis rather than autonomous decision-makers. Extensive human oversight, source traceability, and explicit acknowledgment of current limitations should remain mandatory in pharmaceutical applications. The responsible path forward is not rejection of LLMs, but disciplined adoption grounded in evidence, governance, and humility.

Acknowledgments: None

Conflict of interest: None

Financial support: None

Ethics statement: None

References

1. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. Adv Neural Inf Process Syst. 2017;30:5998-6008.
2. 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.
3. Beltagy I, Lo K, Cohan A. SciBERT: A pretrained language model for scientific text. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). 2019;3615-20.
4. Gu Y, Tinn R, Cheng H, Lucas M, Usuyama N, Liu X, et al. Domain-specific language model pretraining for biomedical natural language processing. ACM Trans Comput Healthc. 2021;3(1):1-23.
5. Yang X, Chen A, PourNejatian N, Shin HC, Smith KE, Parisien C, et al. GatorTron: A large clinical language model to unlock patient information from unstructured electronic health records. arXiv preprint arXiv:2203.03540. 2022.
6. 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):bbac409.
7. ValizadehAslani T, Shi Y, Ren P, Wang J, Zhang Y, Hu M, et al. PharmBERT: a domain-specific BERT model for drug labels. Brief Bioinform. 2023;24(4):bbad226.
8. Wu L, Gray M, Dang O, Xu J, Fang H, Tong W. RxBERT: enhancing drug labeling text mining and analysis with AI language modeling. Exp Biol Med. 2023;248(21):1937-43.
9. Wu L, Fang H, Qu Y, Xu J, Tong W. Leveraging FDA labeling documents and large language model to enhance annotation, profiling, and classification of drug adverse events with AskFDALabel. Drug Saf. 2025;48(6):655-65.
10. Neyarapally GA, Wu L, Xu J, Zhou EH, Dang O, Lee J, et al. Description and validation of a novel AI tool, LabelComp, for the identification of adverse event changes in FDA labeling. Drug Saf. 2024;47(12):1265-74.
11. Wu Y, Liu Z, Wu L, Chen M, Tong W. BERT-based natural language processing of drug labeling documents: a case study for classifying drug-induced liver injury risk. Front Artif Intell. 2021;4:7834.
12. Shi Y, Wang J, Ren P, ValizadehAslani T, Zhang Y, Hu M, et al. Fine-tuning BERT for automatic ADME semantic labeling in FDA drug labeling to enhance product-specific guidance assessment. J Biomed Inform. 2023;138:1045.

13. Shi Y, Ren P, Zhang Y, Gong X, Hu M, Liang H. Information extraction from FDA drug labeling to enhance product-specific guidance assessment using natural language processing. Front Res Metr Anal. 2021;6:670006.
14. Demner-Fushman D, Shooshan SE, Rodriguez L, Aronson AR, Lang F, Rogers W, et al. A dataset of 200 structured product labels annotated for adverse drug reactions. Sci Data. 2018;5(1):180001.
15. Pandey A, Kreimeyer K, Foster M, Dang O, Ly T, Wang W, et al. Adverse event extraction from structured product labels using the event-based text-mining of health electronic records (ETHER) system. Health Informatics J. 2019;25(4):1232-43.
16. Tanaka Y, Chen HY, Belloni P, Gisladottir U, Kefeli J, Patterson J, et al. OnSIDES database: Extracting adverse drug events from drug labels using natural language processing models. Med. 2025;6(7):1011-25.
17. Zitu MM, Owen D, Manne A, Wei P, Li L. Large language models for adverse drug events: A clinical perspective. J Clin Med. 2025;14(15):5490.
18. Sicard J, Montastruc F, Achalme C, Jonville-Bera AP, Songue P, Babin M, et al. Can large language models detect drug–drug interactions leading to adverse drug reactions? Ther Adv Drug Saf. 2025;16:20420986251339358.
19. Soong D, Sridhar S, Si H, Wagner JS, Sá AC, Yu CY, et al. Improving accuracy of GPT-3/4 results on biomedical data using a retrieval-augmented language model. PLOS Digit Health. 2024;3(8):e0000568.
20. 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):e0000877.
21. Li M, Zhan Z, Yang H, Xiao Y, Zhou H, Huang J, et al. Benchmarking retrieval-augmented large language models in biomedical NLP: Application, robustness, and self-awareness. Sci Adv. 2025;11(47):eadr1443.
22. Matsumoto N, Moran J, Choi H, Hernandez ME, Venkatesan M, Moore JH. KRAGEN: a knowledge graph-enhanced RAG framework for biomedical problem solving using large language models. Bioinformatics. 2024;40(6):btae353.
23. Jin Q, Yang Y, Chen Q, Lu Z. GeneGPT: augmenting large language models with domain tools for improved access to biomedical information. Bioinformatics. 2024;40(2):btae075.
24. Alkhalaf M, Yu P, Yin M, Deng C. Applying generative AI with retrieval augmented generation to summarize and extract key clinical information from electronic health records. J Biomed Inform. 2024;156:104662.
25. 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.
26. Wysocka M, Wysocki O, Delmas M, Mutel V, Freitas A. Large language models, scientific knowledge and factuality: A framework to streamline human expert evaluation. J Biomed Inform. 2024;158:104724.