Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 01/08/2026 - **Accepted:** 13/09/2026
Vol: 01.04
DOI: 10.66104/tbe64a08
Pages: 1-25

LARGE LANGUAGE MODELS IN MEDICATION MANAGEMENT: CLINICAL UTILITY, SAFETY ASSURANCE, AND PHARMACIST-LED GOVERNANCE

Júlia Costa Oliveira Ornelas
Faculdade Presidente Antônio Carlos, Brazil
Corresponding author: jujuornelas@yahoo.com
Orcid ID: https://orcid.org/0009-0007-2672-4217

Abstract
Large language models (LLMs) can produce fluent medication explanations, draft patient-facing materials, summarize records, and assist with information retrieval. Their linguistic competence has encouraged proposals for pharmaceutical-care use, yet fluency is not equivalent to clinical reliability. This integrative narrative review evaluates LLMs as governed components of medication-management systems rather than autonomous digital pharmacists. Evidence was synthesized across pharmacy practice, clinical informatics, patient safety, human factors, and AI governance. The analysis distinguishes low-risk language transformation from high-risk clinical judgment and examines drug information, medication reconciliation, interaction screening, adherence support, pharmacovigilance, documentation, education, and multilingual communication. Published evaluations suggest potential workflow utility in constrained tasks but substantial variation across models, prompts, versions, questions, evaluators, and settings. Hazards include confabulation, fabricated citations, critical omission, automation bias, privacy loss, unequal language performance, model drift, and weak reproducibility. Retrieval-augmented generation can improve traceability but cannot eliminate retrieval or synthesis failure. The Pharmacist-Led LLM Assurance Framework for Medication Management (PLLAMM) organizes implementation around five stages: Govern, Map, Measure, Manage, and Monitor. Current evidence supports only conditional utility under pharmacist supervision for bounded, auditable tasks; it does not justify unsupervised medication counseling, autonomous dosing, interaction management, or therapeutic decision-making.

Keywords: large language models; pharmaceutical care; medication safety; clinical pharmacy; generative artificial intelligence; drug information; patient counseling; AI governance

1. Introduction

Medication management is an information-intensive process that connects prescribing, dispensing, administration, monitoring, patient behavior, and communication. Pharmacists routinely translate technical evidence into individual recommendations while considering indication, dose, formulation, interactions, contraindications, organ function, adherence, affordability, and patient preference. Errors can emerge when relevant information is missing, outdated, poorly communicated, or incorrectly applied. The World Health Organization has

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 01/08/2026 - **Accepted:** 13/09/2026
Vol: 01.04
DOI: 10.66104/tbe64a08
Pages: 1-25

therefore treated medication-related harm as a major patient-safety priority and emphasized coordinated improvement across high-risk situations, polypharmacy, and transitions of care (World Health Organization [WHO], 2017).

LLMs introduce a new interface to this problem. Trained on large text collections, they generate responses by predicting linguistic sequences rather than consulting a stable, curated knowledge base in the manner of a conventional drug-information system. Their outputs can appear coherent, tailored, and authoritative even when unsupported. This combination of accessibility and plausible language is valuable for communication but risky when users interpret generated text as verified clinical knowledge (Thirunavukarasu et al., 2023; World Health Organization, 2024).

Pharmacy is a particularly revealing domain for evaluating LLMs because many tasks have explicit safety constraints. A response about administration timing may be useful only if it accounts for formulation; an interaction statement may require severity, mechanism, patient context, and a management plan; a dose recommendation may depend on renal or hepatic function. Partial correctness can therefore be clinically misleading. Evaluations must measure completeness, harmful omission, citation integrity, reproducibility, and appropriateness, not merely whether an answer contains some correct facts (Khatri et al., 2025; Munir et al., 2024).

The public debate frequently alternates between two unsupported extremes: that conversational AI will replace pharmacists, or that any hallucination makes the technology useless. Both positions ignore task structure and system design. LLMs may provide value in low-risk drafting, summarization, translation, and navigation while remaining unsuitable for independent therapy selection. Safety depends on where the model is placed, what information it can access, how its output is verified, and who remains accountable (National Institute of Standards and Technology [NIST], 2024; WHO, 2024).

This article asks under which conditions LLMs can support medication management without weakening professional responsibility or patient safety. Its objectives are to explain the technology in clinically relevant terms; organize uses by task and risk; synthesize evidence on performance; analyze technical, ethical, and operational failure modes; and propose an implementation framework led by pharmacists. The article focuses on LLM-mediated pharmaceutical care rather than autonomous diagnosis and treats deployment as a sociotechnical intervention involving people, data, workflows, institutions, and vendors (Sittig & Singh, 2010; Thirunavukarasu et al., 2023).

The relevant question is not whether LLMs are generally accurate, but whether a specified model configuration can safely support a specified medication-management task under pharmacist-led governance. This task-configuration-workflow thesis guides the review and the proposed assurance framework.

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 01/08/2026 - **Accepted:** 13/09/2026
Vol: 01.04
DOI: 10.66104/tbe64a08
Pages: 1-25

2. Methodology

2.1 Review design and purpose

This article uses an integrative narrative review with thematic synthesis because scholarship on large language models (LLMs) in medication management spans clinical evaluations, benchmark studies, systematic and scoping reviews, patient-safety and human-factors research, technical literature, and regulatory or governance documents. The review does not claim the exhaustive retrieval, protocol registration, or pooled estimation of a systematic review or meta-analysis. Its purpose is to distinguish technical capability from clinically dependable performance and to translate heterogeneous evidence into risk-proportionate requirements for pharmacist-led evaluation, deployment, and oversight.

2.2 Information sources and final search date

The final search update was conducted on September 6, 2026. Sources were identified through PubMed/MEDLINE, Google Scholar, the ACM Digital Library, IEEE Xplore, Crossref-linked publisher platforms, and institutional repositories. Official materials from WHO, NIST, FDA, EMA, ANVISA, the Brazilian federal government, and the Conselho Federal de Farmácia were sought for governance, safety, privacy, professional accountability, and implementation. Backward citation tracing located foundational publications. Because the original search was narrative and iterative, complete contemporaneous retrieval and deduplication counts were not preserved; Appendix A reports the reproducible database-specific strategies and the retained-corpus audit trail without presenting reconstructed counts as prospective records.

2.3 Search strategy

Search combinations included ("large language model" OR LLM OR "generative artificial intelligence" OR ChatGPT) AND (pharmacy OR pharmacist OR "medication management" OR "pharmaceutical care"); and the same model terms combined with "drug information," counseling, reconciliation, interactions, dosing, adherence, pharmacovigilance, or "medication error." Focused searches addressed hallucination, retrieval-augmented generation, source attribution, automation bias, external validation, model drift, human oversight, clinical governance, and regulation. Search syntax was adapted to the functions of each database or platform.

2.4 Eligibility and exclusion criteria

Eligible sources made a direct contribution to at least one analytical domain through empirical evidence, systematic or scoping synthesis, transparent technical or methodological analysis, patient-safety or human-factors scholarship, legal authority, or institutional guidance. Priority was given to peer-reviewed research and authoritative primary documents. Sources were excluded when they consisted mainly of promotional claims, anonymous or non-reproducible comparisons, duplicate propositions, unverifiable methods, or demonstrations that did not distinguish feasibility from clinical benefit. Regulatory and institutional documents were used

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 01/08/2026 - **Accepted:** 13/09/2026
Vol: 01.04
DOI: 10.66104/tbe64a08
Pages: 1-25

within their normative or contextual scope rather than treated as independent evidence of effectiveness.

2.5 Selection process and corpus composition

The author conducted the complete relevance screen, full-text eligibility assessment, data extraction, thematic coding, appraisal, and synthesis. The retained corpus contains 28 sources: 22 analytical sources and six sources used principally for regulatory, legal, ethical, or contextual interpretation. To address classification reliability, an independent verifier, Eduardo Alves Barbosa (not an author), assessed a stratified sample of eight sources (28.6%), covering the evidence roles and analytical domains represented in the corpus. The verifier applied the prespecified eligibility, source-role, domain, and appraisal criteria independently. Agreement was 100% for eligibility and 100% for evidence-role/domain classification, with no disagreements requiring reconciliation. Cohen's kappa for the multicategory evidence-role/domain classification was 1.00, exceeding the journal threshold of 0.75. A separate kappa was not calculated for eligibility because both raters classified all eight sampled sources as included, producing no category variance; percentage agreement is therefore reported for that decision.

2.6 Data extraction and thematic coding

For each retained source, the author recorded publication type, model or system evaluated, model version when reported, prompt or task, medication-management function, setting, data source, access to external knowledge, evaluator, comparator, outcome, clinically important failure, and stated limitation. Evidence was coded across six domains: information retrieval and synthesis; medication review and reconciliation; patient communication; pharmacovigilance and signal detection; safety assurance and evaluation; and pharmacist-led governance. Accuracy was separated from completeness, source fidelity, harmful omission, reproducibility, workflow effect, patient comprehension, and clinical outcomes.

2.7 Critical appraisal and synthesis rules

Evidence was appraised according to task representativeness, clarity of the reference standard, model and prompt documentation, comparator quality, sample and case diversity, external or prospective validation, error severity, citation verification, subgroup analysis, reproducibility, and relevance to real clinical workflows. High examination scores, fluent language, or performance on curated questions were not treated as evidence of safe autonomous use. Greater weight was assigned when independent evidence converged and when evaluation captured consequential omissions or incorrect actions rather than aggregate accuracy alone. Claims remained conditional when evidence was retrospective, single-model, rapidly outdated, or weakly connected to patient outcomes.

2.8 Single-author procedure and methodological limitations

The manuscript remains a single-author work: the author designed the review, screened all records, extracted data, coded themes, appraised evidence, interpreted findings, and wrote the manuscript. Eduardo Alves Barbosa had a limited non-author role as independent verifier of a prespecified 28.6%

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 01/08/2026 - **Accepted:** 13/09/2026
Vol: 01.04
DOI: 10.66104/tbe64a08
Pages: 1-25

source sample. The verifier agreed with all eight eligibility decisions and all eight evidence-role/domain classifications. Observed agreement was 100%, and Cohen's kappa for the multicategory evidence-role/domain classification was 1.00. No reconciliation changes were required. Authorship is not assigned because the verification role did not include study conception, substantive interpretation, drafting, or final responsibility for the manuscript.

3. Technical Foundations for Medication-Related Use

3.1 Probabilistic generation and clinical meaning

An LLM represents statistical relationships among tokens and produces text conditioned on the prompt and preceding context. This mechanism enables flexible language generation but does not guarantee that statements correspond to current pharmaceutical evidence. A model can combine familiar patterns into an incorrect dose, nonexistent source, or overconfident recommendation. The clinically important distinction is between linguistic plausibility and epistemic grounding (Ji et al., 2023; Thirunavukarasu et al., 2023).

Instruction tuning and reinforcement from human feedback can improve helpfulness and reduce undesirable outputs, but they may also encourage a system to provide an answer when refusal or clarification would be safer. Health questions frequently omit age, pregnancy, indication, comorbidities, laboratory values, allergies, formulation, or concurrent medicines. A safe system must identify missing decision variables and request clarification rather than complete the pattern with assumptions (WHO, 2024).

3.2 Context windows, temporal limits, and version instability

The context window limits how much patient information, guideline text, and interaction history can be considered at once. Large windows do not ensure that every detail receives equal attention, and long records can contain contradictions or obsolete entries. Medication review requires temporal reasoning: a discontinued medicine, a historical allergy, and a current prescription have different meanings. Systems must preserve provenance and chronology rather than reduce a record to undifferentiated text (Nerella et al., 2024; Sittig & Singh, 2010).

Knowledge can also be temporally incomplete. New safety communications, labeling changes, shortages, approvals, and guidelines may postdate training. Internet access does not automatically solve the problem because retrieved pages may be outdated, promotional, or irrelevant to the jurisdiction. Clinical outputs should therefore cite dated, authoritative sources and disclose the retrieval time, model version, and applicable population (NIST, 2024; WHO, 2024).

Commercial models change without full public documentation. A prompt that performs well today may behave differently after a model update, safety-policy change, or retrieval modification. This creates a validation problem: evidence belongs to a particular system configuration, not to a permanent product name. Pharmacy implementations need version control, regression testing, and criteria for suspending use after material change (CHART Collaborative, 2025; NIST, 2024).

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 01/08/2026 - **Accepted:** 13/09/2026
Vol: 01.04
DOI: 10.66104/tbe64a08
Pages: 1-25

3.3 Retrieval-augmented generation and domain adaptation

Retrieval-augmented generation (RAG) connects a language model to an external collection and supplies retrieved passages during response generation. In pharmacy, the collection may include approved labeling, formularies, protocols, monographs, and institutional policies. RAG can improve currency and citation traceability because answers are conditioned on selected sources rather than memory alone (Lewis et al., 2020).

RAG is not a guarantee of truth. A retriever may miss the relevant document, select a similarly named drug, rank a secondary source above an official one, or omit a decisive contraindication. The generator may then misinterpret the retrieved passage or combine incompatible recommendations. Evaluation must separately test document ingestion, retrieval recall, source authority, synthesis fidelity, and abstention when evidence is insufficient (Lewis et al., 2020; NIST, 2024).

Fine-tuning or domain adaptation can improve terminology and task behavior, but it creates maintenance obligations and may not make factual knowledge transparent. A locally adapted model can still be outdated or confidently wrong. For high-stakes medication information, curated retrieval, deterministic calculations, rule-based checks, and human review may be preferable to embedding all functions in one generative model (Singhal et al., 2023; WHO, 2024).

4. A Risk-Based Taxonomy of Pharmacy Tasks

Risk assignment uses six criteria: potential severity, reversibility, detectability, dependence on patient-specific data, need for deterministic calculation, and required professional judgment. Table 1 applies these criteria.

Table 1. Risk-Based Taxonomy of LLM-Supported Pharmacy Tasks

Task category	Potential LLM role	Primary failure mode	Minimum safeguard	Risk level and basis
Language transformation	Rewrite verified content for readability or format.	Omission or distortion during simplification.	Compare output with approved source.	Low: limited harm; reversible; readily detectable.
Documentation	Draft a note from structured encounter data.	Unsupported facts become part of the record.	Pharmacist review, provenance, and signature.	Low-moderate: record propagation possible; review detects most errors.
Patient education	Explain reviewed medicine information.	Incomplete precautions or false personalization.	Approved sources and escalation rules.	Moderate: omission may harm; depends on patient context.
Drug information	Support search, extraction, and synthesis.	Fabricated citations or selective evidence.	Open and verify every cited source.	Moderate-high: source fidelity and professional judgment required.
Reconciliation	Normalize lists and flag discrepancies.	Incorrect resolution of conflicting records.	Patient interview and pharmacist confirmation.	High: discrepancies may affect transitions; patient-specific verification required.
Interaction/dosing	Explain validated alerts or calculations.	Missed modifiers, unit error, unsafe action.	Deterministic tools plus clinical review.	Very high: severe harm, calculations/thresholds, low error tolerance.
Therapy selection	Organize options for professional deliberation.	Overconfident recommendation from incomplete data.	Qualified decision-maker retains authority.	Very high: direct therapeutic consequence and irreducible professional judgment.

4.1 Language transformation and administrative support

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 01/08/2026 - **Accepted:** 13/09/2026
Vol: 01.04
DOI: 10.66104/tbe64a08
Pages: 1-25

The lowest-risk opportunities involve transforming text whose clinical content has already been verified. Examples include changing reading level, drafting appointment reminders, structuring a pharmacist's note, translating a reviewed instruction, or producing alternative wording for education. Even these tasks require checking because simplification can remove conditions, translation can alter dose units, and summaries can omit exceptions. The model should transform an approved source rather than invent new clinical content (NIST, 2024; WHO, 2024).

Documentation assistance may reduce clerical burden by converting a structured encounter into a draft note. The safe unit is a draft that the pharmacist verifies and signs, not an autonomous record. Systems should distinguish patient statements from professional assessment, avoid importing unsupported facts, and show the source segments used. Audit logs and correction mechanisms are essential because generated errors can become durable when copied into electronic records (Sittig & Singh, 2010).

4.2 Patient education and medication counseling

LLMs can generate accessible explanations of purpose, administration, expected effects, and common precautions. They can adapt tone and vocabulary and may help patients formulate questions before speaking with a professional. Studies of patient-facing medication information have found responses that are often understandable and partly correct, but completeness varies by drug and evaluation criterion. Missing monitoring, contraindication, or escalation advice can remain clinically important even when the prose appears helpful (Abanmy et al., 2025).

Counseling is not a generic monograph. Effective communication should be prioritized according to the patient's regimen, risks, goals, language, and capacity. An LLM without a verified medication list or relevant clinical data cannot safely personalize recommendations. Consumer systems should explicitly separate general education from individualized advice and direct urgent, complex, pediatric, pregnancy, or high-risk questions to qualified care (WHO, 2024).

4.3 Drug information and literature support

Drug-information work requires question clarification, systematic source selection, critical appraisal, synthesis, citation, and a response tailored to the decision. LLMs can help decompose a question, generate search concepts, summarize retrieved documents, and draft a response. They are weaker substitutes for source verification because generated citations can be nonexistent or unrelated and confident language can conceal uncertainty (Alkaissi & McFarlane, 2023; Khatri et al., 2025).

Real-world evaluations show why task definitions matter. Khatri and colleagues reported limited accuracy and reproducibility across broad drug-information categories, while other studies found stronger performance in constrained calculations or factual questions and weaker results in patient cases. These differences do not contradict one another; they indicate that

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 01/08/2026 - **Accepted:** 13/09/2026
Vol: 01.04
DOI: 10.66104/tbe64a08
Pages: 1-25

performance is sensitive to scope, prompt, model, and scoring. A local service must validate the exact questions and sources it expects to encounter (Khatri et al., 2025; Munir et al., 2024).

4.4 Medication reconciliation and regimen review

Medication reconciliation involves comparing sources, resolving discrepancies, and determining what a patient is actually taking. An LLM can normalize names, identify duplicate entries, summarize changes, or propose questions. It should not independently decide that a discrepancy is intentional or discontinue a medicine. Reliable reconciliation requires patient interview, clinical context, and access to authoritative records that may disagree (WHO, 2017). Table 2 translates these boundaries into permitted, supervised, and prohibited functions for each reconciliation subtask.

Table 2. Permitted Boundaries for LLM Use in Medication Reconciliation

Reconciliation subtask	Status	Required control
Normalize medicine names and formatting	Permitted	Use a controlled dictionary; preserve source entries.
Match route, strength, and formulation	Supervised	Flag uncertainty; pharmacist confirms equivalence.
Detect possible duplicates	Supervised	Treat as a flag, not a discontinuation decision.
Identify discrepancies across lists	Supervised	Display provenance, dates, and conflicting entries.
Draft patient-interview questions	Permitted	Use approved prompts; avoid leading questions.
Draft a reconciliation note	Supervised	Use reconciled data only; pharmacist edits and signs.
Resolve discrepancies or discontinue therapy	Prohibited	Patient interview and accountable clinician/pharmacist decision required.
Create the final reconciled medication list	Prohibited autonomously	Pharmacist confirmation and documented accountability required.

Comprehensive medication management adds indication, effectiveness, safety, adherence, and patient goals. Case-based studies suggest that advanced LLMs can identify some drug-related problems, but benchmark success does not show reliable management across uncommon conditions, multiple organ dysfunction, complex interactions, or incomplete histories. Pharmacist review remains essential because the task combines evidence with individualized judgment (Roosan et al., 2024; Yang et al., 2025).

4.5 Interaction screening, dosing, and therapeutic decisions

Interaction screening appears structured but is clinically nuanced. A meaningful answer must identify the interacting agents, mechanism, severity, evidence quality, timing, patient modifiers, and management options. Generative systems can omit a key interaction or overstate a theoretical one. They should therefore supplement, not replace, validated interaction databases and pharmacist interpretation (Sicard et al., 2025).

Dosing is high risk because equations, units, formulations, organ function, indication, age, weight, and monitoring requirements interact. LLMs may perform familiar calculations correctly yet fail when data are missing or cases deviate from training patterns. Deterministic calculators and rules should handle arithmetic and threshold logic, while the LLM may explain a verified result. Autonomous dose changes are not justified by current evidence (Munir et al., 2024; WHO, 2024).

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 01/08/2026 - **Accepted:** 13/09/2026
Vol: 01.04
DOI: 10.66104/tbe64a08
Pages: 1-25

Therapy selection is the highest-risk category because it requires diagnosis, goals, comparative evidence, contraindications, patient preference, cost, and follow-up. Examination performance demonstrates encoded medical knowledge but does not establish accountability or safe care. LLM recommendations should remain advisory, traceable to sources, and subject to licensed clinical judgment (Singhal et al., 2023; Thirunavukarasu et al., 2023).

5. Evidence on Accuracy, Completeness, and Clinical Usefulness

5.1 Why reported performance varies

Published accuracy figures vary because studies ask different questions and score different outcomes. Multiple-choice examinations favor recognition among fixed options, while drug-information practice requires clarification and documentation. A model may receive credit for naming an appropriate drug even if it omits dose, duration, monitoring, or an important contraindication. Studies should report item-level criteria and harmfulness, not only total accuracy (Lee et al., 2024; Thirunavukarasu et al., 2023).

Prompting also changes results. Providing patient context, asking for stepwise reasoning, requiring citations, or allowing retrieval can improve some responses, but it can also produce longer answers with more opportunities for error. A test that selects the best prompt after observing results risks overstating expected performance. Protocols should predefine prompts and report all repetitions, failures, and model settings (CHART Collaborative, 2025).

Human scoring introduces further variation. Pharmacists may disagree about completeness or acceptable management when evidence is limited. Blinded independent reviewers, explicit rubrics, adjudication, and inter-rater agreement improve credibility. Evaluations should also compare the model with realistic alternatives: a pharmacist using ordinary references, a conventional database, or the same workflow without AI (Bossuyt et al., 2015; CHART Collaborative, 2025).

5.2 Factual questions versus patient cases

Constrained factual questions usually provide all needed information and have stable answers, making them easier for LLMs. Patient cases require recognizing what is missing, balancing competing goals, and avoiding premature closure. Munir and colleagues found marked variation across pharmacy question categories, including substantially weaker performance on patient cases than on some structured categories. Such findings support task-specific authorization rather than a general label of competence (Munir et al., 2024).

Advanced models can perform impressively in selected case sets. Roosan and colleagues reported strong results for GPT-4 in clinical-pharmacy cases, illustrating the effect of model generation and study design. The appropriate interpretation is that an LLM can assist with certain well-described cases, not that it has been validated for autonomous practice. External validation with prospective patients, workflow outcomes, and safety endpoints remains necessary (Roosan et al., 2024).

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 01/08/2026 - **Accepted:** 13/09/2026
Vol: 01.04
DOI: 10.66104/tbe64a08
Pages: 1-25

5.3 Reproducibility and update effects

LLM output can vary across repeated runs because generation is probabilistic and system parameters may be hidden. A response that is correct once may be incomplete later. Reproducibility is especially important for medication advice because patients expect equivalent questions to receive consistent safety information. Studies should repeat prompts, quantify variation, and report worst-case as well as average performance (Khatri et al., 2025).

Model updates can invalidate prior findings. Naming a product without a release date or version is insufficient. Evaluations should record the exact model, access route, date, system instructions, temperature where available, tools, retrieval collection, and prompt sequence. Health institutions need regression suites that test common, high-risk, and adversarial cases before accepting an update (CHART Collaborative, 2025; NIST, 2024).

5.4 From answer quality to patient outcomes

Most current evidence evaluates generated answers rather than medication errors, adherence, quality of life, hospital use, or patient understanding over time. Answer quality is an intermediate outcome. A system may produce accurate text without changing behavior, or it may improve access while creating new risk through false reassurance. Clinical adoption requires prospective evaluation of outcomes and unintended consequences (Ong et al., 2025; WHO, 2024).

Efficiency claims also require full accounting. Time saved drafting an answer may be offset by verification, correction, documentation, integration, and incident management. Benefit is more plausible when pharmacists verify short, source-linked drafts than when they must audit long, unstructured output. Studies should measure total task time, cognitive workload, error detection, and downstream rework, not generation speed alone (Sittig & Singh, 2010).

6. Core Safety Risks

6.1 Confabulation and fabricated evidence

Confabulation occurs when a model generates unsupported content that fits linguistic patterns. In pharmacy, the error may concern a dose, interaction, approval status, administration instruction, or citation. Fabricated references are especially hazardous because they create an appearance of evidence and waste verification time. Users should never infer reliability from confident tone or citation formatting (Alkaissi & McFarlane, 2023; Ji et al., 2023).

Requiring citations can improve auditability only when links resolve to relevant sources and the claims match those sources. Systems should preferably retrieve from a controlled collection and expose the quoted supporting passage. Pharmacists must still open the source because a real citation can be mischaracterized. Citation validity, entailment, currency, and jurisdiction are separate checks (NIST, 2024).

6.2 Omission, overgeneralization, and false reassurance

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 01/08/2026 - **Accepted:** 13/09/2026
Vol: 01.04
DOI: 10.66104/tbe64a08
Pages: 1-25

An answer can be factually correct yet unsafe because it omits a condition. Saying that a medicine is generally compatible may be misleading for pregnancy, renal impairment, a modified-release formulation, or a specific interacting regimen. Evaluations should therefore use mandatory-information checklists and identify critical omissions separately from ordinary incompleteness (Abanmy et al., 2025; Khatri et al., 2025).

LLMs also tend to produce a complete-sounding response even when the question lacks clinical details. False reassurance may delay professional assessment, while excessive warnings may generate anxiety or nonadherence. Safe systems should express uncertainty, state assumptions, identify missing variables, and use calibrated escalation rather than generic disclaimers appended to an otherwise decisive answer (WHO, 2024).

6.3 Automation bias and deskilling

Automation bias occurs when users accept system output too readily or reduce independent checking. Fluent explanations can intensify this tendency because they resemble expert communication. Less experienced users may receive useful structure while being less able to detect subtle errors. Training must include known failure examples, verification procedures, and permission to reject or suspend the tool (Goddard et al., 2012; NIST, 2024).

Long-term dependence can also weaken information-search and critical-appraisal skills if the system replaces rather than supports professional reasoning. Pharmacy education should teach students to compare generated answers with primary or authoritative sources and document disagreements. Assessment should measure whether learners retain the ability to work safely without the tool (Thirunavukarasu et al., 2023).

6.4 Bias, language, and health equity

Training data reflect unequal representation, historical practice, and dominant languages. Performance demonstrated in English cannot be assumed for Brazilian Portuguese, regional vocabulary, low-literacy communication, or indigenous languages. Translation can alter units and culturally specific instructions. Local validation must include the populations, languages, literacy levels, and access conditions of intended use (WHO, 2021, 2024).

Digital systems can expand access but also deepen exclusion when they require reliable connectivity, modern devices, or advanced literacy. Patients with the greatest medication-management burden may have the least ability to identify erroneous output. Accessibility, disability, age, socioeconomic status, and digital confidence should be included in usability testing rather than treated as post-deployment concerns (Milne-Ives et al., 2020; WHO, 2021).

6.5 Privacy, confidentiality, and security

Medication questions can reveal diagnoses, pregnancy, mental health, substance use, finances, and identity. Entering such data into a public chatbot may disclose sensitive information outside the protections of a clinical record. Institutions must determine data roles, retention, training use, access, location, deletion, incident response, and subcontractors before use. In Brazil,

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 01/08/2026 - **Accepted:** 13/09/2026
Vol: 01.04
DOI: 10.66104/tbe64a08
Pages: 1-25

health information is sensitive personal data under the General Data Protection Law (Brazil, 2018).

Prompt injection and contaminated documents create additional security risks when an LLM retrieves from external sources or acts within systems. Malicious text can attempt to redirect behavior, expose data, or produce unsafe instructions. Least privilege, input isolation, source allowlists, output validation, and human authorization are necessary when systems connect to records, prescribing, or communication channels (NIST, 2024).

7. Human Factors and Professional Responsibility

7.1 The pharmacist as accountable decision-maker

Professional accountability cannot be transferred to a probabilistic model. A pharmacist who uses an LLM remains responsible for the clinical appropriateness of communicated information within the applicable legal and professional framework. Institutions and vendors also have duties, but distributed responsibility should not become an accountability gap. Governance must specify who approves use, who verifies output, and who responds to incidents (WHO, 2021, 2024).

Human oversight must be meaningful. Requiring a signature after presenting a polished recommendation can create ceremonial review under time pressure. Interfaces should display source passages, uncertainties, missing data, and alternative interpretations so that the pharmacist can interrogate the result. Workload and alert design determine whether oversight operates in practice (Goddard et al., 2012; Sittig & Singh, 2010).

7.2 Preserving the therapeutic relationship

Pharmaceutical care involves listening, negotiation, empathy, and understanding why a person uses or avoids a medicine. A chatbot may make information available outside office hours, but it cannot assume that text alone captures patient priorities or safeguarding needs. AI-mediated communication should create routes to human contact and should not be used to ration professional attention away from patients who need it most (WHO, 2021).

Disclosure supports trust. Patients should know when content is generated, whether a pharmacist reviewed it, what data were used, and how to report a concern. Disclosure should be clear without shifting the burden of technical risk to patients. Consent may be required for some data uses, but consent does not replace institutional responsibility for safety and necessity (Brazil, 2018; WHO, 2021).

7.3 Workflow fit and new categories of error

Technology can reduce one error while creating another. Automated summaries may improve visibility but copy an incorrect medication across records; a chatbot may improve access while delaying escalation; a generated note may save time but blur the source of an assessment. Sociotechnical evaluation maps interactions among people, tasks, interfaces, organizations, and external rules (Sittig & Singh, 2010).

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 01/08/2026 - **Accepted:** 13/09/2026
Vol: 01.04
DOI: 10.66104/tbe64a08
Pages: 1-25

Implementation should begin with observation of current work. If a service's main problem is missing patient data, adding a generative model will not solve it. If delays come from repetitive formatting of verified information, drafting support may help. A clear problem statement prevents adoption from becoming a search for uses after a technology has already been purchased (NIST, 2024).

8. Pharmacist-Led Assurance Framework

The Pharmacist-Led LLM Assurance Framework for Medication Management (PLLAMM) has five connected stages: Govern, Map, Measure, Manage, and Monitor. It is an evidence-informed implementation guide, not a validated clinical intervention. Each stage produces auditable evidence and a release decision proportionate to task risk. Table 3 summarizes the evidence required and the release question for each PLLAMM stage.

Table 3. Pharmacist-Led LLM Assurance Framework for Medication Management

Assurance stage	Required evidence	Release question
Govern	Named owners, policies, prohibited uses, incident authority.	Is accountability explicit and enforceable?
Map	Workflow, users, data, affected decisions, harm pathways.	Is the intended use narrow and necessary?
Measure	Representative tests, critical failures, equity, reproducibility.	Does evidence meet predeclared risk thresholds?
Manage	Controls, verification, escalation, access, security, training.	Can failures be prevented, detected, and contained?
Monitor	Logs, audits, complaints, overrides, drift, update testing.	Does real-world performance remain acceptable?

8.1 Define the intended use and prohibited uses

Every implementation should state the user, setting, task, input, output, decision affected, and expected benefit. It should also list prohibited uses, such as autonomous dose adjustment, unsupervised counseling for high-risk medicines, or insertion of generated text into records without review. Narrow authorization makes evaluation possible and limits the consequences of failure (NIST, 2024; WHO, 2024).

Risk classification should consider severity, reversibility, detectability, exposure, and dependence on patient-specific data. Rewriting an approved leaflet is not equivalent to choosing anticoagulant therapy. Higher-risk tasks require stronger evidence, more structured inputs, deterministic controls, and closer human oversight. Some uses may remain unacceptable even if average accuracy appears high (WHO, 2024).

8.2 Build an authoritative evidence layer

A pharmacy system should use an approved source hierarchy: current regulatory labeling, institutional formulary and policies, validated drug databases, guidelines, and primary literature when necessary. Documents need ownership, versioning, effective dates, jurisdiction, and retirement procedures. Retrieval should preserve document identity and passage location so that a reviewer can reconstruct the answer (Lewis et al., 2020; NIST, 2024).

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 01/08/2026 - **Accepted:** 13/09/2026
Vol: 01.04
DOI: 10.66104/tbe64a08
Pages: 1-25

Conflicts between sources require explicit rules. A national label, specialty guideline, and local protocol may differ because they answer different questions. The model should not silently average them. It should present the conflict, identify applicability, and escalate to a pharmacist. Governance of the knowledge base is therefore as important as model selection (WHO, 2024).

8.3 Validate with representative and adversarial cases

Validation sets should represent routine questions, rare but severe risks, incomplete prompts, ambiguous drug names, look-alike products, dose units, special populations, polypharmacy, and local language. Cases should be separated into development and locked test sets. Testing only textbook questions creates inflated confidence and misses the situations in which safe refusal or clarification is the correct response (Bossuyt et al., 2015; CHART Collaborative, 2025).

Outcome measures should include factual accuracy, completeness, critical omission, harmfulness, citation validity, retrieval success, reproducibility, calibration, refusal quality, equity, time, and user workload. Thresholds should be defined before testing and tied to task risk. A favorable average cannot compensate for catastrophic failure on a small set of high-risk cases (NIST, 2024).

Comparison should be made against current practice. If a pharmacist using standard databases is faster and safer, the model adds little. If a source-linked draft reduces time without reducing quality, the benefit is credible. Prospective shadow testing allows the system to run without influencing care while discrepancies are measured (Ong et al., 2025).

8.4 Design verification and escalation

Verification should be proportional and structured. A pharmacist should receive the source, relevant patient data, generated claim, uncertainty, and required checklist. For dose or interaction questions, deterministic calculations and database checks should be displayed separately from generated explanation. This prevents the LLM from becoming an opaque single point of failure (NIST, 2024).

Non-negotiable safeguard rule: LLMs may explain validated outputs from approved deterministic calculators, drug-interaction databases, and institutional protocols, but must not independently generate dose changes, interaction-severity classifications, or therapeutic substitutions without pharmacist verification.

Escalation rules should activate when data are missing, sources conflict, retrieval fails, the medicine is high risk, the patient belongs to a special population, or the question concerns an urgent symptom. A safe response may be to stop and route the case rather than generate more text. Escalation performance should be tested as a primary outcome (WHO, 2024).

8.5 Monitor, audit, and manage change

Deployment is the beginning of validation, not its end. Logs should support review of prompts, retrieved sources, outputs, corrections, overrides, latency, and incidents while respecting

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 01/08/2026 - **Accepted:** 13/09/2026
Vol: 01.04
DOI: 10.66104/tbe64a08
Pages: 1-25

privacy. Sampling should include ordinary and high-risk cases. Patients and staff need simple reporting routes, and governance teams need authority to suspend the system (NIST, 2024).
Material changes include model updates, prompt revisions, new tools, knowledge-base updates, interface changes, and altered vendor data practices. Each change should be assessed for revalidation. Regression tests should be run before release and after incidents. Vendor assurances cannot replace local evidence because risk depends on the specific workflow and population (CHART Collaborative, 2025; NIST, 2024).

9. Application Areas and Boundary Conditions

9.1 Adherence support

LLMs can generate reminders, explore barriers conversationally, and adapt education to a patient's questions. Adherence is not merely forgetting; it can reflect adverse effects, cost, beliefs, complexity, or disagreement with therapy. A system that pushes reminders without identifying the barrier may be ineffective or coercive. Pharmacist-supervised tools should support shared decisions and escalation rather than assume that greater adherence is always the correct outcome (Milne-Ives et al., 2020; WHO, 2021).
Evidence for conversational agents shows promise but heterogeneous interventions and outcomes. Engagement often declines over time, and short studies may overestimate sustained use. Trials should measure clinically meaningful adherence, persistence, safety, patient experience, and equity, not only message counts or satisfaction (Milne-Ives et al., 2020).

9.2 Pharmacovigilance

Conversational systems may help patients describe suspected adverse reactions and structure reports. An LLM can ask follow-up questions about timing, dose, dechallenge, concomitant medicines, and outcomes. It should not determine causality autonomously. Reports need human review, deduplication, coding, and submission through established pharmacovigilance systems (WHO, 2019).
The same system can introduce reporting bias by leading patients toward expected symptoms or summarizing uncertain narratives too decisively. Interfaces should preserve the original account and distinguish generated structure from patient language. Evaluation should compare completeness and signal quality with ordinary reporting, while monitoring whether particular groups are excluded (WHO, 2019, 2021).

9.3 Community pharmacy

Community pharmacies receive frequent questions about nonprescription medicines, missed doses, administration, minor ailments, and interactions. An LLM may support intake and prepare a structured summary before the pharmacist intervenes. The main safety requirement is triage: the system must recognize red flags, uncertainty, vulnerable populations, and questions that require clinical assessment (Shin et al., 2024).

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 01/08/2026 - **Accepted:** 13/09/2026
Vol: 01.04
DOI: 10.66104/tbe64a08
Pages: 1-25

Operational benefit depends on integration with verified references and dispensing systems without exposing unnecessary data. A public chatbot used by a patient differs from a controlled professional copilot. Claims from one setting cannot be transferred to the other because available data, user expertise, and safeguards differ (Thirunavukarasu et al., 2023).

9.4 Hospitals and transitions of care

Hospitals offer richer data and more complex risk. LLMs could summarize medication histories, discharge changes, and monitoring plans, but source errors can propagate quickly. Transition documents should be generated only from reconciled data, label uncertainty, and remain subject to pharmacist review. High-risk medicines require standardized content and teach-back rather than open-ended generation alone (WHO, 2017).

Integration with electronic health records raises access-control and provenance requirements. The system should retrieve only necessary information, distinguish note authorship, and avoid using generated conclusions as source facts in subsequent generations. Monitoring should detect copy-forward and contradiction across versions of the medication list (Sittig & Singh, 2010).

9.5 Education and professional development

LLMs can create cases, explain mechanisms, simulate counseling, and provide formative feedback. Their educational value is greatest when students critique output and verify sources. Unsupervised answer generation can undermine assessment validity and expose learners to plausible errors. Curricula should teach prompt design, evidence retrieval, bias, privacy, and the boundaries of automation (Thirunavukarasu et al., 2023).

Faculty need transparent policies distinguishing permitted assistance from authorship or misconduct. Because systems and access differ, assessment should not reward only familiarity with a commercial model. Students must demonstrate independent competence in calculations, literature evaluation, counseling, and patient safety (WHO, 2024).

10. Brazilian and Resource-Constrained Contexts

In Brazil, LLM deployment intersects with universal-health-system objectives, geographic inequality, Portuguese-language performance, professional regulation, and the LGPD. Imported evaluations may not reflect national medicine names, formularies, labeling, clinical protocols, or access pathways. Systems intended for Brazilian care require validation with local sources such as ANVISA-approved information and applicable therapeutic guidelines (Brazil, 2018). Institutional deployment should align professional accountability with Conselho Federal de Farmácia rules and organizational pharmacy governance; ground medicine information in ANVISA-authorized labeling, electronic leaflets, safety communications, and applicable SUS clinical protocols; and document why each data use is necessary. Health information is sensitive personal data under the LGPD, so an institution must identify a lawful basis, minimize collection, control access, define retention, manage processors, and establish incident response rather than treating consent as a universal solution. Validation must use Brazilian Portuguese,

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 01/08/2026 - **Accepted:** 13/09/2026
Vol: 01.04
DOI: 10.66104/tbe64a08
Pages: 1-25

locally used medicine names, formulations, literacy levels, regional vocabulary, and SUS referral pathways.

Resource constraints strengthen both the appeal and the danger of automation. A chatbot may extend access where pharmacists are scarce, but the same scarcity reduces the availability of human verification and escalation. Technology should not be presented as an equivalent substitute for missing professional care. Safer designs connect remote users to pharmacists and reserve automation for bounded support (WHO, 2021).

Local development should include pharmacists, patients, informaticians, privacy specialists, accessibility experts, and representatives of underserved communities. Participatory design can identify language, trust, workflow, and access problems that benchmark teams overlook. Procurement should require data portability, audit access, security obligations, service continuity, and notification of model changes (NIST, 2024; WHO, 2021).

Cost evaluation must include integration, knowledge-base maintenance, validation, training, verification time, privacy controls, monitoring, vendor dependence, and incident response. A nominally inexpensive public model can become costly when staff must repair unreliable output. Conversely, a controlled system may justify its cost if it reduces documentation burden without compromising safety. Total cost should be compared with simpler alternatives (Sittig & Singh, 2010).

11. Discussion

The literature supports a conditional conclusion: LLMs are useful language technologies but not self-validating clinical authorities. Their strongest current role is to transform, organize, and communicate information within a system that supplies verified knowledge and professional review. Their weakest role is autonomous judgment under incomplete information, where fluent output can conceal missing evidence or patient context (Khatri et al., 2025; WHO, 2024).

Performance heterogeneity is not a minor methodological inconvenience; it is a central finding. Results change with task, prompt, model, version, retrieval, repetition, and evaluator. The appropriate question is therefore not whether "ChatGPT is accurate," but whether a specified configuration meets a predeclared threshold for a specified task, population, and workflow. General claims should be rejected when the evidence is local (CHART Collaborative, 2025).

RAG and specialized models shift rather than eliminate risk. They can improve grounding and relevance, but retrieval quality, document governance, and synthesis become new failure points. A transparent system that sometimes abstains may be safer than a more capable system that produces untraceable answers. Architecture should favor inspectability and containment over the appearance of conversational completeness (Lewis et al., 2020; NIST, 2024).

Human oversight must be designed, funded, and measured. If verification takes too long, staff will bypass it; if interfaces hide sources, reviewers cannot evaluate claims; if productivity targets reward acceptance, automation bias becomes organizational. Implementation research should examine incentives and workload alongside accuracy (Goddard et al., 2012; Sittig & Singh, 2010).

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 01/08/2026 - **Accepted:** 13/09/2026
Vol: 01.04
DOI: 10.66104/tbe64a08
Pages: 1-25

Pharmacists should have a leading role because they understand drug-information methods, medication-use processes, and the consequences of incomplete counseling. Their role extends beyond final approval to defining the task, selecting sources, designing test cases, setting escalation rules, monitoring errors, and educating users. Governance led only by technology teams risks optimizing fluency or speed rather than medication safety (WHO, 2017, 2024).

The evidence does not support replacing pharmaceutical care with consumer chatbots. It does support carefully bounded copilots that reduce clerical work, improve access to verified content, and help professionals communicate. These systems should be judged by net patient and workflow outcomes, not demonstration quality or examination scores (Ong et al., 2025; Thirunavukarasu et al., 2023).

12. Research Priorities

Future studies should move from static questions to prospective workflow evaluations. Multicenter research should measure medication discrepancies, preventable harm, patient comprehension, adherence, total task time, workload, and equity. Study protocols should register the model configuration and lock prompts before analysis to reduce selective reporting (CHART Collaborative, 2025).

To move evaluation beyond answer quality, Table 4 specifies patient-, safety-, workflow-, governance-, and equity-centered outcomes recommended for prospective studies.

Table 4. Recommended Patient-, Safety-, Workflow-, and Equity-Centered Endpoints

Domain	Recommended endpoint	Interpretation
Medication safety	Medication-error and preventable adverse-event rates	Direct clinical safety, not answer quality alone.
Reconciliation	Discrepancy detection and verified resolution	Separate flags from correctly resolved discrepancies.
Output safety	Critical omission and harmful-action rates	Weight consequential failures above minor inaccuracies.
Communication	Patient comprehension and teach-back success	Assess understanding after exposure.
Adherence	Validated adherence/persistence measures	Examine benefit and unintended pressure.
Pharmacovigilance	Completeness and usefulness of adverse-event reports	Preserve original patient narrative.
Workflow	Pharmacist workload and verification time	Include correction and escalation burden.
Governance	Appropriate abstention and escalation	Measure both missed and excessive escalation.
Equity	Performance by language, literacy, disability, and subgroup	Report gaps, not only pooled averages.
Trust	Calibrated patient/professional trust	Distinguish justified reliance from automation bias.

Evaluation datasets should include incomplete and contradictory records, uncommon medicines, special populations, multilingual questions, and adversarial inputs. Public test sets need careful governance to prevent contamination and should be complemented by confidential local cases. Reporting should include the full distribution of outcomes and examples of critical failure without exposing patient information (NIST, 2024).

Comparative research should test generative systems against pharmacists with ordinary references, established clinical decision support, and hybrid workflows. The economically relevant comparison is not AI versus unaided memory but the new system versus current best

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 01/08/2026 - **Accepted:** 13/09/2026
Vol: 01.04
DOI: 10.66104/tbe64a08
Pages: 1-25

feasible practice. Factorial designs can isolate the contributions of retrieval, model generation, and pharmacist review (Bossuyt et al., 2015).

Longitudinal studies are required because performance, behavior, and dependence change over time. Researchers should examine whether users become better at verification or more complacent, whether vendor updates improve or regress safety, and whether initial efficiency survives integration costs. Post-deployment surveillance should connect incidents to system configuration and corrective action (NIST, 2024).

Research in Portuguese and in low-resource settings is essential. It should examine health literacy, regional vocabulary, access, disability, and culturally appropriate escalation. Equitable design cannot be inferred from average performance in English-speaking academic centers (WHO, 2021).

13. Limitations

Methodological limitations include the integrative narrative design, compact 28-source corpus, author-led full-corpus screening and coding, absence of a prospectively registered protocol, and incomplete contemporaneous retrieval and deduplication counts. The appendices improve transparency but cannot retroactively convert the work into a systematic review. Independent verification covered a stratified 28.6% sample rather than the entire corpus; although agreement was complete (100%; Cohen's kappa = 1.00 for evidence-role/domain classification), sampling leaves residual risk of undetected disagreement elsewhere in the corpus.

Evidentiary limitations include heterogeneous designs, reference standards, prompts, models, outcomes, and frequent reliance on simulated tasks or answer-quality endpoints rather than patient outcomes. Technological limitations include rapid model replacement, undocumented vendor changes, prompt sensitivity, retrieval failure, model drift, and restricted reproducibility. Clinical limitations arise because risk and acceptable performance differ across education, documentation, reconciliation, dosing, interactions, and therapy selection; results cannot be transferred across tasks without local validation. Jurisdictional limitations include variation in law, professional scope, formularies, clinical pathways, language, infrastructure, and access. PLLAMM is therefore a governance framework requiring prospective validation, not evidence that any product is clinically effective.

14. Conclusion

LLMs have potential workflow utility for organizing information, drafting documentation, adapting verified education, and navigating evidence collections. Conditional utility may arise when the task is bounded, the source is authoritative, the output is readily verifiable, and a pharmacist remains accountable. Current evidence is dominated by intermediate performance measures and does not establish broad patient-outcome benefit.

The principal risks are confabulation, omission, fabricated evidence, inconsistency, bias, privacy loss, automation bias, and silent change. A response may look professional while being clinically incomplete. Accordingly, average accuracy is not enough: assurance must address critical failures, reproducibility, sources, population, workflow, and human behavior (Khatri et al., 2025; NIST, 2024).

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 01/08/2026 - **Accepted:** 13/09/2026
Vol: 01.04
DOI: 10.66104/tbe64a08
Pages: 1-25

Pharmacist-led governance provides the practical bridge between innovation and safety. A defensible implementation defines intended and prohibited uses, uses authoritative evidence, validates representative cases, separates deterministic functions from generation, supports meaningful verification, escalates uncertainty, and monitors real-world performance. Tools that cannot meet these requirements should not influence medication decisions (WHO, 2024).

Current evidence supports LLMs as supervised components of pharmaceutical care, not autonomous digital pharmacists. The profession should neither surrender judgment to persuasive text nor reject tools that can be governed effectively. The appropriate goal is augmented pharmaceutical care in which technology increases access and efficiency while responsibility, evidence appraisal, and patient-centered decisions remain human (WHO, 2021, 2024).

Ethical Approval

Ethical approval was not required because this study analyzed publicly available literature and institutional documents and involved no human participants, identifiable personal data, animals, or intervention.

Funding

The author received no financial support for the research, authorship, or publication of this article.

Conflicts of Interest

The author declares no conflict of interest.

Author and Independent Verification Contributions

Júlia Costa Oliveira Ornelas is the sole author and is responsible for conceptualization, methodology, full-corpus screening, extraction, thematic analysis, interpretation, writing, and final accountability. Eduardo Alves Barbosa is identified only as an independent methodological verifier for a prespecified eight-source sample; he is not an author and is not responsible for the manuscript's conclusions. His completed verification record and any disagreements must be retained with the review documentation.

References

Abanmy, N. O., Al-Ghreimil, N., Alsabhan, J. F., Al-Baity, H., & Aljadeed, R. (2025). Evaluating the accuracy of ChatGPT in delivering patient instructions for medications: An exploratory case study. Frontiers in Artificial Intelligence, 8, 1550591. https://doi.org/10.3389/frai.2025.1550591

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 01/08/2026 - **Accepted:** 13/09/2026
Vol: 01.04
DOI: 10.66104/tbe64a08
Pages: 1-25

Alkaissi, H., & McFarlane, S. I. (2023). Artificial hallucinations in ChatGPT: Implications in scientific writing. Cureus, 15(2), e35179. https://doi.org/10.7759/cureus.35179

Bossuyt, P. M., Reitsma, J. B., Bruns, D. E., et al. (2015). STARD 2015: An updated list of essential items for reporting diagnostic accuracy studies. BMJ, 351, h5527. https://doi.org/10.1136/bmj.h5527

Brazil. (2018). Law No. 13,709 of August 14, 2018: General Personal Data Protection Law.

Goddard, K., Roudsari, A., & Wyatt, J. C. (2012). Automation bias: A systematic review of frequency, effect mediators, and mitigators. Journal of the American Medical Informatics Association, 19(1), 121-127. https://doi.org/10.1136/amiajnl-2011-000089

Grant, M. J., & Booth, A. (2009). A typology of reviews: An analysis of 14 review types and associated methodologies. Health Information and Libraries Journal, 26(2), 91-108. https://doi.org/10.1111/j.1471-1842.2009.00848.x

CHART Collaborative. (2025). Reporting guideline for chatbot health advice studies: The CHART statement. JAMA Network Open, 8(8), e2530220. https://doi.org/10.1001/jamanetworkopen.2025.30220

Ji, Z., Lee, N., Frieske, R., et al. (2023). Survey of hallucination in natural language generation. ACM Computing Surveys, 55(12), Article 248. https://doi.org/10.1145/3571730

Khatri, S., et al. (2025). Accuracy and reproducibility of ChatGPT responses to real-world drug information questions. Journal of the American College of Clinical Pharmacy. https://doi.org/10.1002/jac5.70038

Lee, P., et al. (2024). Assessing the research landscape and clinical utility of large language models: A scoping review. BMC Medical Informatics and Decision Making, 24, 72. https://doi.org/10.1186/s12911-024-02459-6

Lewis, P., Perez, E., Piktus, A., et al. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. Advances in Neural Information Processing Systems, 33, 9459-9474.

Milne-Ives, M., de Cock, C., Lim, E., et al. (2020). The effectiveness of artificial intelligence conversational agents in health care: Systematic review. Journal of Medical Internet Research, 22(10), e20346. https://doi.org/10.2196/20346

Munir, F., et al. (2024). Evaluation of ChatGPT as a tool for answering clinical pharmacy questions. Journal of Pharmacy Technology. https://doi.org/10.1177/08971900241256731

National Institute of Standards and Technology. (2024). Artificial intelligence risk management framework: Generative artificial intelligence profile (NIST AI 600-1). https://doi.org/10.6028/NIST.AI.600-1

Nerella, S., Bandyopadhyay, S., Zhang, J., et al. (2024). Transformers and large language models in healthcare: A review. Artificial Intelligence in Medicine, 154, 102900. https://doi.org/10.1016/j.artmed.2024.102900

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 01/08/2026 - **Accepted:** 13/09/2026
Vol: 01.04
DOI: 10.66104/tbe64a08
Pages: 1-25

Ong, J. C. L., et al. (2025). A scoping review on generative AI and large language models for reducing medication-related harm. npj Digital Medicine. https://doi.org/10.1038/s41746-025-01565-7

Roosan, D., Padua, P., Khan, R., Khan, H., Verzosa, C., & Wu, Y. (2024). Effectiveness of ChatGPT in clinical pharmacy and the role of artificial intelligence in medication therapy management. Journal of the American Pharmacists Association, 64(2), 422-428.e8. https://doi.org/10.1016/j.japh.2023.11.023

Shin, E., Hartman, M., & Ramanathan, M. (2024). Performance of the ChatGPT large language model for decision support in community pharmacy. British Journal of Clinical Pharmacology, 90(12), 3320-3333. https://doi.org/10.1111/bcp.16215

Sicard, J., et al. (2025). Can large language models detect drug-drug interactions in pharmacovigilance cases? Therapeutic Advances in Drug Safety, 16. https://doi.org/10.1177/20420986251339358

Singhal, K., Azizi, S., Tu, T., et al. (2023). Large language models encode clinical knowledge. Nature, 620, 172-180. https://doi.org/10.1038/s41586-023-06291-2

Sittig, D. F., & Singh, H. (2010). A new sociotechnical model for studying health information technology in complex adaptive healthcare systems. Quality and Safety in Health Care, 19(Suppl. 3), i68-i74. https://doi.org/10.1136/qshc.2010.042085

Sukhera, J. (2022). Narrative reviews: Flexible, rigorous, and practical. Journal of Graduate Medical Education, 14(4), 414-417. https://doi.org/10.4300/JGME-D-22-00480.1

Thirunavukarasu, A. J., Ting, D. S. J., Elangovan, K., et al. (2023). Large language models in medicine. Nature Medicine, 29, 1930-1940. https://doi.org/10.1038/s41591-023-02448-8

World Health Organization. (2017). Medication without harm: Global patient safety challenge on medication safety.

World Health Organization. (2019). WHO guideline: Recommendations on digital interventions for health system strengthening.

World Health Organization. (2021). Ethics and governance of artificial intelligence for health: WHO guidance.

World Health Organization. (2024). Ethics and governance of artificial intelligence for health: Guidance on large multi-modal models.

Yang, H., Hu, M., Most, A., Hawkins, W. A., Murray, B., Smith, S. E., & Li, S. (2025). Evaluating accuracy and reproducibility of large language model performance on critical care assessments in pharmacy education. Frontiers in Artificial Intelligence, 8, 1514896. https://doi.org/10.3389/frai.2024.1514896

Appendix A. Reproducible Search Strategies and Audit Trail

The searches were iterative and narrative. The strings below reproduce the final strategy; database-level retrieval and duplicate counts were not preserved contemporaneously and are therefore reported as 'not

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 01/08/2026 - **Accepted:** 13/09/2026
Vol: 01.04
DOI: 10.66104/tbe64a08
Pages: 1-25

recorded' rather than reconstructed. This is a transparency limitation, not a claim of systematic-review completeness. Table A1 presents the database-specific fields, filters, reproducible queries, and status of the retrieval counts.

Table A1. Database-Specific Search Strategy

Source	Fields/filters	Final reproducible query	Retrieval count
PubMed/MEDLINE	Title/Abstract; English/Portuguese; through 6 Sep 2026	((large language model*[Title/Abstract] OR LLM[Title/Abstract] OR ChatGPT[Title/Abstract] OR generative AI[Title/Abstract]) AND (pharmacy[Title/Abstract] OR pharmacist*[Title/Abstract] OR medication management[Title/Abstract] OR pharmaceutical care[Title/Abstract] OR drug information[Title/Abstract] OR medication reconciliation[Title/Abstract] OR drug interaction*[Title/Abstract] OR dosing[Title/Abstract] OR pharmacovigilance[Title/Abstract] OR medication error*[Title/Abstract]))	Not recorded
ACM Digital Library	Title/abstract/keywords; through 6 Sep 2026	("large language model" OR ChatGPT OR "generative AI") AND (medication OR pharmacy OR clinical) AND (safety OR evaluation OR governance OR retrieval)	Not recorded
IEEE Xplore	All metadata; through 6 Sep 2026	("large language model" OR ChatGPT OR "generative AI") AND (pharmacy OR medication OR drug information) AND (safety OR validation OR decision support)	Not recorded
Google Scholar	First 200 relevance-ranked results per focused query; through 6 Sep 2026	Exact model terms combined separately with pharmacy, medication management, counseling, reconciliation, interactions, dosing, adherence, pharmacovigilance, medication error, automation bias, governance, and model drift	Not recorded
Publisher/Crossref platforms	Title/keyword and DOI verification	Focused title, author, DOI, and citation tracing for candidate and foundational sources	Not recorded
Institutional/legal sources	Official domains; current documents through 6 Sep 2026	WHO/NIST/FDA/EMA/ANVISA/CFF/LGPD terms combined with AI, medication safety, privacy, governance, and professional accountability	Not recorded

Appendix B. Source-Selection Flow

Table B1 reports the available source-selection audit trail and distinguishes recorded corpus counts from retrieval information that was not preserved contemporaneously.

Table B1. Retained-Corpus Audit Flow

Stage	Count	Audit statement
Database/platform records identified	Not recorded	The original narrative search did not retain export-level counts.
Duplicates removed	Not recorded	No reproducible deduplication log was preserved.
Titles/abstracts screened	Not recorded	Eligibility concepts are reported in Section 2.4.
Full texts assessed/excluded	Not recorded	Full-text exclusion reasons were not logged prospectively.
Sources retained	28	22 analytical; 6 regulatory/legal/ethical/contextual.
Independent verification sample	8 (28.6%)	Stratified across roles and domains; see Appendix D.

Appendix C. Source Characteristics and Critical Appraisal

Table C1 characterizes every retained source and makes the source-level critical appraisal visible by reporting evidence role, task or domain, source type, and principal limitation.

Table C1. Source-Level Characteristics and Appraisal Matrix

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 01/08/2026 - **Accepted:** 13/09/2026
Vol: 01.04
DOI: 10.66104/tbe64a08
Pages: 1-25

Source	Year	Evidence role	Task/domain	Type	Principal limitation
1. Abanmy, N. O., Al-Ghreimil, N., Alsabhan, J. F., Al-Baity, H., & Aljadeed, R.	2025	Clinical	Patient instructions	Empirical case evaluation	Model/prompt specificity; no patient outcomes
2. Alkaissi, H., & McFarlane, S. I.	2023	Technical	Hallucination/citations	Commentary/case examples	Not medication-workflow validation
3. Bossuyt, P. M., Reitsma, J. B., Bruns, D. E., et al.	2015	Governance	Evaluation reporting	Reporting guideline	Diagnostic focus; indirect application
4. Brazil.	2018	Governance	Privacy	Law	Jurisdiction-specific legal authority
5. Goddard, K., Roudsari, A., & Wyatt, J. C.	2012	Technical	Automation bias	Systematic review	Predates current LLM interfaces
6. Grant, M. J., & Booth, A.	2009	Governance	Review method	Methodological typology	General, not pharmacy-specific
7. CHART Collaborative.	2025	Governance	Chatbot reporting	Reporting guideline	Reporting standard, not effectiveness evidence
8. Ji, Z., Lee, N., Frieske, R., et al.	2023	Technical	Hallucination	Technical review	Limited direct clinical workflow evidence
9. Khatri, S., et al.	2025	Clinical	Drug information	Empirical evaluation	Model/version and local question dependence
10. Lee, P., et al.	2024	Clinical	Clinical utility landscape	Scoping review	Heterogeneous tasks and intermediate endpoints
11. Lewis, P., Perez, E., Piktus, A., et al.	2020	Technical	RAG	Technical research	Not designed for clinical safety
12. Milne-Ives, M., de Cock, C., Lim, E., et al.	2020	Clinical	Conversational agents/adherence	Systematic review	Pre-LLM heterogeneity and short follow-up
13. Munir, F., et al.	2024	Clinical	Clinical pharmacy questions	Empirical evaluation	Curated questions; limited patient outcomes
14. National Institute of Standards and Technology.	2024	Governance	AI risk management	Government framework	Non-prescriptive and not pharmacy-specific
15. Nerella, S., Bandyopadhyay, S., Zhang, J., et al.	2024	Technical	Healthcare LLM mechanisms	Narrative review	Broad healthcare scope
16. Ong, J. C. L., et al.	2025	Clinical	Medication-related harm	Scoping review	Emerging, heterogeneous evidence
17. Roosan, D., Padua, P., Khan, R., Khan, H., Verzosa, C., & Wu, Y.	2024	Clinical	Medication therapy management	Empirical evaluation	Single configuration; simulated tasks
18. Shin, E., Hartman, M., & Ramanathan, M.	2024	Clinical	Community pharmacy decisions	Empirical evaluation	Context/model-specific
19. Sicard, J., et al.	2025	Clinical	Drug interactions	Empirical evaluation	Case selection and model-version limits
20. Singhal, K., Azizi, S., Tu, T., et al.	2023	Technical	Clinical knowledge	Benchmark evaluation	Benchmark performance is not workflow safety
21. Sittig, D. F., & Singh, H.	2010	Governance	Sociotechnical safety	Conceptual model	Predates generative AI
22. Sukhera, J.	2022	Governance	Narrative review rigor	Methodological commentary	No task-specific appraisal instrument
23. Thirunavukarasu, A. J., Ting, D. S. J., Elangovan, K., et al.	2023	Technical	LLMs in medicine	Narrative review	Broad and rapidly aging evidence
24. World Health Organization.	2017	Governance	Medication safety	Global guidance	Normative; not LLM evaluation
25. World Health Organization.	2019	Governance	Digital interventions	Guideline	Predates current LLM evidence
26. World Health Organization.	2021	Governance	AI ethics	Global guidance	High-level normative framework
27. World Health Organization.	2024	Governance	Generative AI governance	Global guidance	Normative; implementation requires localization
28. Yang, H., Hu, M., Most, A., Hawkins, W. A., Murray, B., Smith, S. E., & Li, S.	2025	Clinical	Pharmacy education	Empirical evaluation	Education endpoint, not patient outcome

Appendix D. Independent Verification Record

The sample was prespecified to represent the evidence roles and analytical domains in the retained corpus. Eduardo Alves Barbosa completed the independent eligibility, evidence-role/domain, and

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 01/08/2026 - **Accepted:** 13/09/2026
Vol: 01.04
DOI: 10.66104/tbe64a08
Pages: 1-25

overall-appraisal assessment without changing the author's classifications. All eight sources were independently confirmed as eligible, and every evidence-role/domain classification agreed. Observed agreement was 8/8 (100%). Cohen's kappa for the multicategory evidence-role/domain classification was 1.00. Eligibility kappa was not estimable because both raters assigned the same inclusion category to every sampled source; the absence of category variance makes chance-corrected agreement undefined for that decision. No disagreements required reconciliation. Table D1 identifies the eight verified sources, their coverage rationale, the independent decisions, and the absence of reconciliation changes.

Table D1. Prespecified 28.6% Verification Sample

Source no.	Coverage rationale	Verifier eligibility	Role/domain agreement	Appraisal agreement	Resolution
1	patient communication	Agreed: include	Agreed	Agreed	No change
5	human factors	Agreed: include	Agreed	Agreed	No change
9	drug information	Agreed: include	Agreed	Agreed	No change
11	retrieval/system risk	Agreed: include	Agreed	Agreed	No change
13	clinical pharmacy	Agreed: include	Agreed	Agreed	No change
16	medication harm synthesis	Agreed: include	Agreed	Agreed	No change
18	community pharmacy	Agreed: include	Agreed	Agreed	No change
27	AI governance	Agreed: include	Agreed	Agreed	No change

Verification statistics: sources independently double-coded = 8/28 (28.6%); eligibility agreement = 8/8 (100%; kappa not estimable because all ratings were 'include'); evidence-role/domain agreement = 8/8 (100%); Cohen's kappa = 1.00; disagreements = 0; reconciliation changes = 0. Independent verifier: Eduardo Alves Barbosa.

CC BY