Review Article

Where Does the Answer Come From? A Scoping Review of Evidence Tracing, Hallucination, and Safeguards in Generative Medicines Information

Nguyen Thanh Huy[1]*, Pham Quang Minh[1], Le Thi Bich[2], Tran Van Nam[3]

[1]Department of Generative AI and Medicines Information, Faculty of Pharmacy, Hanoi University of Pharmacy, Hanoi, Vietnam. [2]Department of Evidence Tracing and Hallucination Detection, Faculty of Pharmacy, Can Tho University of Medicine and Pharmacy, Can Tho, Vietnam. [3]Department of AI Safeguards in Pharmacy, Faculty of Pharmacy, Hue University, Hue, Vietnam.

Abstract

Generative artificial intelligence can produce medicines information rapidly, yet fluent answers can conceal uncertainty about where claims originated, whether cited sources exist, and whether retrieved evidence supports the wording presented. This scoping review mapped how generative medicines information systems attach evidence to claims, define and measure hallucination and citation error, implement retrieval and provenance, evaluate claim–source alignment, and apply human verification, conflict handling, and escalation safeguards. A protocol-led scoping review used the Population–Concept–Context framework and followed JBI and PRISMA-ScR principles. Peer-reviewed studies of generative systems producing medication or medicines information were eligible when they reported extractable evidence-tracing, citation, hallucination, retrieval, or safeguard methods. Records were screened, charted, appraised, and synthesized through evidence-tracing and safeguard taxonomies. The evidence base was methodologically heterogeneous and concentrated in retrospective, cross-sectional, or simulated evaluations. Systems attached evidence through model-generated references, search-linked citations, constrained document retrieval, or retrieval-augmented generation, but citation presence did not reliably establish bibliographic validity, relevance, or claim-level support. Hallucination definitions varied across fabricated references, incorrect citation elements, unsupported factual statements, and source–claim mismatch. Retrieval reduced some unsupported generation yet introduced failures involving document selection, temporal validity, partial entailment, and evidence conflict. Human review was frequently recommended, but escalation thresholds, reviewer workload, audit trails, and prospective workflow performance were rarely evaluated. Evidence tracing in generative medicines information remains fragmented. Trustworthy evaluation requires separate assessment of source existence, source quality, retrieval completeness, claim-level support, contradiction, and human adjudication. Current safeguards alone do not establish clinical safety or deployment readiness.

Keywords: Scoping review, Digital pharmacy, Artificial intelligence, Pharmacy practice, Medication safety, Evidence synthesis

Introduction

Generative language models have moved medicines information from static retrieval toward interactive synthesis. A user can request an explanation of an interaction, a comparison of therapies, a counselling script, or a summary of an adverse-effect signal and receive a fluent response within seconds. Contemporary medical large language models can integrate broad textual knowledge and adapt outputs to conversational context, creating opportunities for education, documentation, and preliminary information support [1]. Their apparent coherence, however, is not equivalent to evidential transparency. Foundation models learn statistical relationships across extensive corpora, while the precise documents contributing to an individual output are generally not recoverable from model parameters alone [2].

This distinction is especially important for medicines information. Medication claims may depend on indication, dose, formulation, timing, comorbidity, interacting treatments, population characteristics, and the date of the underlying evidence. A response may be linguistically plausible while relying on outdated, incomplete, weak, or nonexistent sources. GPT-4 evaluations in medicine have therefore emphasized that impressive task performance can coexist with factual mistakes, unstable reasoning, and

Address for correspondence: Nguyen Thanh Huy, Department of Generative AI and Medicines Information, Faculty of Pharmacy, Hanoi University of Pharmacy, Hanoi, Vietnam.
huy.nguyen@hup.edu.vn
Received: 01 May 2026; **Accepted:** 05 September 2026

How to cite this article: Huy NT, Minh PQ, Bich LT, Nam TV. Where Does the Answer Come From? A Scoping Review of Evidence Tracing, Hallucination, and Safeguards in Generative Medicines Information. Arch Pharm Pract. 2026;17(3):84-93. https://doi.org/10.51847/MDBmOwWqrB

uncertainty about appropriate use [3]. When a generated answer includes citations, users may interpret them as an audit trail even when references are fabricated, bibliographic details are incorrect, retrieved sources are irrelevant, or the cited text does not support the associated claim.

Evidence tracing denotes the processes through which a generated claim can be connected to identifiable source material and examined for support. It includes, but is not limited to, reference generation, document retrieval, provenance recording, citation placement, claim segmentation, source–claim alignment, contradiction detection, and preservation of an audit trail. Hallucination is related but broader: it may concern invented facts, fabricated references, incorrect citation elements, unsupported synthesis, or statements that exceed retrieved evidence. Safeguards may operate before generation, during retrieval and drafting, after generation through automated verification, or through professional review and escalation. Patient-safety analyses caution that generative systems must be evaluated according to the consequences of errors, not fluency or average accuracy alone [4].

This scoping review asked how generative medicines-information systems attach evidence to claims; how hallucination and citation error are defined and measured; which retrieval, provenance, and alignment designs are used; what verification, human-review, conflict-handling, and escalation safeguards have been tested; and where evaluations remain confined to laboratory or simulated settings. The purpose was to map methods and boundaries rather than estimate a pooled effect or establish clinical effectiveness. It also identified reporting gaps that obstruct comparison across systems.

Review Design, Questions, and Protocol

Protocol

The review was designed as a scoping review because the objective was to map concepts, technical approaches, evaluation practices, and evidence gaps across a heterogeneous field rather than estimate an intervention effect. Reporting followed the PRISMA extension for scoping reviews, including explicit eligibility criteria, information sources, selection processes, charting procedures, and presentation of evidence distributions [5]. The protocol was finalized before the definitive search. It specified the review questions, date boundaries, eligible source types, screening rules, extraction fields, exclusion hierarchy, taxonomy domains, and procedures for resolving uncertainty. Methodological decisions were aligned with updated JBI guidance emphasizing transparent question formulation, structured charting, and descriptive synthesis appropriate to the review purpose [6].

Protocol deviations were logged with timing, rationale, and expected effects on eligibility or interpretation.

PCC Formulation

The population comprised users, patients, health professionals, pharmacists, or simulated cases receiving or evaluating medicines information generated by a large language model or related generative system. The concept comprised evidence tracing, citation generation, retrieval-augmented generation, provenance, claim–source alignment, hallucination, verification, human review, evidence-conflict management, escalation, and auditability. The context included community, hospital, ambulatory, academic, regulatory, pharmacovigilance, research, and public-facing settings. A scoping approach was retained because the questions concerned the extent, characteristics, definitions, and evaluation of evidence rather than comparative effectiveness, consistent with guidance distinguishing scoping from systematic effectiveness reviews [7].

Review Questions

Six questions organized the review: how evidence was attached to generated medicines-information claims; how hallucination and citation error were defined and measured; which retrieval and provenance designs were used; how claim–source alignment was assessed; which human-verification and escalation safeguards were tested; and which safeguards had only laboratory or simulated evaluation. An existing scoping review of generative systems and medication-related harm informed terminology for medication tasks and safety consequences but did not determine eligibility because its question was broader than evidence tracing [8].

Definitions

Generation meant producing new natural-language medicines information rather than retrieving an unchanged document. A citation was any explicit source identifier presented as supporting an answer. Evidence tracing required a recoverable connection between a claim and source material. Hallucination covered fabricated or unsupported content, including nonexistent references and source–claim mismatch. A safeguard was a technical, procedural, or human control intended to detect, prevent, expose, or manage unsupported generation. These definitions separated source existence, bibliographic correctness, relevance, entailment, and clinical appropriateness.

Eligibility, Definitions, and Information Sources

Eligibility Criteria

Reports published from 2017 through August 3, 2026 were eligible when they evaluated a generative system producing medication, drug, pharmacotherapy, pharmacovigilance, or medicines-information content and reported extractable methods or findings concerning citations, retrieval, provenance, hallucination, source support, verification, conflict handling, or human oversight. Methodological articles were retained when they governed review conduct or defined an evaluation construct. Protocol guidance supported

prospective specification of objectives, concepts, eligibility boundaries, and planned presentation [9].

Reports were excluded when they evaluated only examination performance or general clinical accuracy without evidence-tracing or safeguard information; addressed non-medicines applications; used non-generative information extraction alone; lacked sufficient methodological detail; were editorials, abstracts, preprints, or non-peer-reviewed materials; or duplicated a more complete report. Scoping reviews were understood as evidence syntheses that systematically identify and map the breadth and characteristics of evidence without requiring pooled effect estimation [10]. One primary reason was assigned to every full-text exclusion.

Information Sources and Search Boundary

MEDLINE/PubMed, Embase, Scopus, and Web of Science Core Collection were searched from January 1, 2017 through August 3, 2026. Supplementary searching included backward reference checking, forward citation searching, and targeted searches for named systems, citation metrics, provenance methods, and retrieval-augmented medication applications. Gray literature was examined only to identify terminology or linked peer-reviewed reports and was not included as evidence. Reporting of databases, platforms, dates, search strings, limits, and supplementary methods followed systematic-search expectations [11].

Search concepts combined controlled vocabulary and free-text terms for generative artificial intelligence or large language models; medication, drug information, pharmacy, pharmacotherapy, or pharmacovigilance; and citation, reference, retrieval, provenance, grounding, hallucination, verification, safeguard, alignment, contradiction, human review, or escalation. Database syntax was adapted without changing the conceptual blocks. Search documentation followed PRISMA-S principles so that source selection, query translation, deduplication inputs, and supplementary searching remained reproducible [12]. No language restriction was applied during retrieval; reports required sufficient English information for reliable screening and extraction.

Search, Selection, Extraction, Appraisal, and Synthesis

Screening and Data Charting

Search results were merged in a master ledger and deduplicated by DOI or persistent identifier, then normalized title, first author, year, and report characteristics. Two reviewers independently screened titles and abstracts, assessed full texts, and resolved disagreements through discussion or third-reviewer adjudication. Data charting followed recommendations to extract information directly answering scoping questions while separating source characteristics, methods, findings, and reviewer interpretation [13]. Fields included setting, task, model, knowledge source, retrieval design, citation behavior, evaluation unit, reference standard, hallucination definition, alignment method, safeguards, human role, conflict handling, outcomes, and validation level.

Table 1 organizes the evidence, constructs, relationships, uncertainties, and boundary conditions required for eligibility, definitions, and evidence-extraction framework for the review Where Does the Answer Come From? A Scoping Review of Evidence Tracing, Hallucination, and Safeguards in Generative Medicines Information.

Table 1. Eligibility, definitions, and evidence-extraction framework for the review Where Does the Answer Come From? A Scoping Review of Evidence Tracing, Hallucination, and Safeguards in Generative Medicines Information.

Review Element	Operational Definition	Eligibility or Decision Rule	Data Field	Rationale	Bias or Ambiguity Risk	Planned Synthesis Use
Review Design	Systematic mapping of concepts, evidence types, methods, and gaps	Retain scoping questions; do not pool heterogeneous outcomes	Review purpose; question type	Appropriate for heterogeneous and emerging evidence [5]	Drift toward effectiveness review	Organize descriptive and thematic mapping
Population	People, professionals, simulated patients, cases, or prompts receiving or evaluating generated medicines information	Include real or simulated users and medication cases	User group; case source; intended audience	Captures patient-facing and professional contexts	Simulated cases may not represent practice	Stratify by audience and setting
Generative System	Model producing new natural-language medication content	Exclude unchanged document retrieval and non-generative classifiers	Model; version; access date; prompting	Separates generation from conventional search	Incomplete model-version reporting	Compare system and prompting characteristics
Medicines Information	Content concerning drugs, pharmacotherapy, medication use, safety, counselling, or surveillance	Require a medication-specific task or output	Task; medicine domain; clinical purpose	Maintains direct pharmacy relevance	Broad clinical studies may contain minor medication content	Classify application domains
Evidence Tracing	Recoverable connection between a generated claim and identifiable source material	Include explicit citation, retrieval, provenance, attribution, or alignment evaluation	Source identifier; attribution level; audit trail	Distinguishes traceable evidence from unsupported fluency	Citation presence may be mistaken for support	Construct evidence-tracing taxonomy

Hallucination	Fabricated, incorrect, unsupported, or source-incongruent generated content	Extract the study's definition and denominator without harmonizing incompatible measures	Error unit; definition; numerator; denominator	Preserves methodological differences	Studies use the same label for different errors	Compare definitions and measurement approaches
Citation Validity	Existence and bibliographic correctness of a cited source	Assess separately from relevance and claim support	DOI or identifier; metadata accuracy; source existence	Prevents conflation of authenticity and evidential fit	Partial metadata errors may be classified inconsistently	Map bibliographic evaluation methods
Claim–Source Alignment	Degree to which a source supports, partially supports, contradicts, or does not address a claim	Require an explicit human or automated comparison	Claim unit; source passage; rating scale; adjudicator	Evaluates support beyond citation existence	Claim segmentation and entailment judgments vary	Classify alignment approaches
Safeguard	Technical, procedural, interface, or human control addressing unsupported generation	Include safeguards that are described sufficiently for extraction	Safeguard stage; trigger; action; fallback	Maps prevention, detection, and response controls	Recommendations may be reported as tested controls	Separate proposed from evaluated safeguards
Human Review	Professional verification, correction, approval, or escalation before information use	Extract reviewer expertise, instructions, workload, and authority	Reviewer type; number; agreement; escalation rule	Human oversight is not a uniform intervention	Expert review may conceal system deficiencies	Compare oversight and escalation designs
Validation Level	Laboratory, retrospective, simulated workflow, prospective, or operational evaluation	Record the highest level directly demonstrated	Setting; external data; prospective use; workflow integration	Separates benchmark performance from implementation evidence	Authors may overstate real-world relevance	Identify evidence maturity and gaps
Evidence Conflict	Disagreement among retrieved sources, guidelines, dates, or professional judgments	Include explicit detection, display, prioritization, or escalation procedures	Conflict type; resolution rule; uncertainty display	Medication evidence may be conditional or time-sensitive	Silent source selection can obscure disagreement	Map conflict-handling mechanisms

Appraisal and Taxonomies

Each report was appraised for sampling, comparator quality, reproducibility, blinding, denominator reporting, model-version transparency, external validation, and applicability. Bibliographic evaluation separated source existence, metadata correctness, and topical relevance, reflecting multidimensional approaches such as the Reference Hallucination Score [14]. Citation fabrication and retrieval performance were not treated as equivalent; comparative evaluations show that a verifiable reference may still fail to support the associated statement [15].

The evidence-tracing taxonomy classified model-generated references, search-linked citations, corpus-constrained retrieval, document-level provenance, passage-level attribution, and claim-level alignment. The safeguard taxonomy classified source restriction, retrieval and reranking, citation validation, entailment checking, contradiction detection, uncertainty display, refusal, human verification, and escalation. Automated assessment was interpreted cautiously because large-scale medical evaluations found substantial proportions of cited statements incompletely supported or contradicted [16].

Synthesis and Visualization

Findings were summarized descriptively by task, setting, system, tracing design, hallucination measure, safeguard, and validation level. Meta-analysis was not planned because outcomes, denominators, prompts, models, reference standards, and rating procedures were not comparable. Contradictions and missing domains were retained rather than resolved by vote counting. Review-derived categories were labelled as interpretive synthesis, and numerical findings were reproduced only with original denominators, context, and reported uncertainty.

Search and Selection Results

Selection Results

The searches identified 1,284 records: 344 from MEDLINE/PubMed, 372 from Embase, 317 from Scopus, 201 from Web of Science, and 50 through supplementary searching. Removal of 318 duplicate records left 966 unique records for title and abstract screening. Of these, 901 were excluded because they did not evaluate generative medicines information, evidence tracing, hallucination, or a relevant safeguard. Sixty-five reports were sought for retrieval; three could not be obtained, leaving 62 full texts for eligibility assessment. Forty-seven reports were excluded: 13 were not medication-specific, 12 did not evaluate evidence tracing or safeguards, seven assessed non-generative systems, six lacked extractable methods, four were not peer-reviewed journal reports, three duplicated or incompletely reported another study, and two fell outside the eligibility period. Fifteen reports representing 15 distinct studies entered the

descriptive synthesis. No study entered a quantitative synthesis because outcome definitions, denominators, systems, prompts, and reference standards were not sufficiently comparable.

Model and Application Characteristics

The included evidence was recent and strongly concentrated after public release of general-purpose conversational models. Most studies used cross-sectional or retrospective comparisons in which identical drug-information questions, medication cases, or prompts were submitted to one or more commercially available models. A real-world drug-information evaluation found frequent false or partly correct responses, missing references, variable outputs across repeated prompts, and potentially harmful recommendations [17]. Clinical-pharmacy comparisons examined prescription review, medication education, adverse-reaction recognition, causality assessment, counselling, interaction detection, and medication-therapy management. Performance varied materially by task complexity and rating method [18]. Medication-therapy-management investigations similarly showed that apparently acceptable answers could coexist with incomplete reasoning, unrecognized interactions, or dependence on prompt construction [19].

The evaluated systems included successive ChatGPT versions, GPT-4-class models, Gemini or Bard, Bing or Copilot, Perplexity, and specialized retrieval interfaces. Patient-facing studies emphasized understandable drug explanations and safety risks, whereas professional-facing studies examined the correctness, completeness, reproducibility, references, and potential for harm of answers to authentic or constructed queries. No included study demonstrated sustained autonomous operation in routine pharmacy practice.

Evidence-Base Characteristics

Citation-Generation Approaches

Citation behavior fell into three categories. First, some models generated bibliographic references from parametric memory without an explicit retrieval step. These references could appear plausible while containing invented articles, incorrect authors, altered titles, nonexistent identifiers, or mismatched publication details. Comparative testing of ChatGPT and Bard showed that reference generation could produce substantial hallucination and inconsistent bibliographic accuracy [20]. Second, search-enabled systems produced links or citations after querying an external index. Third, corpus-constrained systems retrieved documents before drafting and preserved at least document-level provenance.

Bibliographic evaluation was inconsistent. Some studies verified only whether a reference existed; others compared title, author, journal, year, link, and DOI fields. Large-sample application of the Reference Hallucination Score illustrated a multidimensional method by scoring verifiability, bibliographic accuracy, identifier validity, and topical relevance, but its composite structure did not directly establish whether a source supported each generated claim [21]. SourceCheckup extended evaluation from source authenticity to statement-level relevance and supportiveness, finding that many medically referenced responses remained incompletely supported or contradicted despite containing sources [22].

Retrieval-Augmented Generation

Retrieval-augmented generation linked a model to a searchable knowledge base, guideline collection, medicine leaflet, literature corpus, or institutional resource. Implementations differed in query reformulation, embedding model, chunk size, number of retrieved passages, reranking, context-window assembly, and whether citations were displayed. Retrieval generally improved access to current or bounded sources, yet its benefit depended on whether the right document and passage were retrieved. It could also introduce failures through poor query formulation, irrelevant passages, missing documents, context truncation, or uncritical synthesis of discordant evidence.

No implementation evaluated retrieval as a single uniform intervention. Performance depended on corpus governance, indexing date, source granularity, query generation, ranking, passage selection, context limits, citation rendering, and the model's response to missing or conflicting material.

Claim–Source Alignment Methods

Only a minority of studies evaluated support at the level of individual claims. Common approaches divided an answer into statements and asked clinicians, pharmacists, or trained annotators whether a cited source supported, partly supported, contradicted, or failed to address each statement. Automated methods used natural-language-inference models, embedding similarity, model-based judges, or agent pipelines. These methods required different evidence units: an article, abstract, paragraph, retrieved passage, sentence, or structured guideline recommendation. Consequently, apparently similar alignment percentages could refer to materially different tasks.

The strongest evaluations preserved the claim, cited source, supporting passage, reviewer judgment, and disagreement resolution. Weaker evaluations inferred support from topical similarity or the presence of shared keywords. Source quality was also distinct from alignment: a claim could accurately reflect a weak source, while a high-quality source could be cited for a claim it did not support. Healthcare RAG reviews found predominantly retrospective technical evaluations, heterogeneous metrics, limited external validation, and insufficient attention to safety, bias, and clinical workflow [23]. Evidence-based question-answering benchmarks showed that providing selected literature could improve answer accuracy, although performance decreased when evidence was ambiguous or distributed across narrative

sources [24]. A medication-instruction study reported better adequacy and clarity when authoritative patient-information material was incorporated through retrieval, but human validation remained part of the proposed safety pathway [25].

Hallucination Definitions and Measurement

Hallucination was used for at least four phenomena: fabricated references, bibliographic corruption, unsupported factual content, and statements inconsistent with supplied evidence. Some studies measured the proportion of answers containing any hallucination; others used claims, citations, citation fields, or reviewer-identified errors as denominators. The review therefore did not combine hallucination rates. Each result was retained with its original unit and denominator. Measurements that counted only nonexistent references were classified as bibliographic hallucination, whereas unsupported or contradicted claims were classified as content or alignment failures. Missing clinically important information was charted separately as omission.

Verification Safeguards

Safeguards operated before, during, or after generation. Pre-generation controls restricted the permitted corpus, required structured prompts, specified the user and task, or prohibited unsupported extrapolation. Generation-time controls retrieved and reranked sources, required citations after claims, requested uncertainty statements, or instructed the model to refuse when evidence was absent. Post-generation controls checked identifiers, compared claims with passages, detected contradictions, or routed outputs to professional review.

Drug-information comparisons verified model outputs against validated sources, institutional answers, guidelines, or expert judgments. A real-world pharmacy-practice study compared two commonly used chatbots with answers produced by a health-system drug-information service and assessed correctness, completeness, references, and potential harm [26]. Hospital-pharmacy testing demonstrated that repeat submissions could yield changing advice and error profiles, making reproducibility itself a safety-relevant outcome [27]. Another drug-information-center comparison found moderate aggregate accuracy but weaker reference identification, demonstrating that answer-level scoring can conceal deficits in evidential traceability [28].

Human Review and Escalation

Human review was the most frequently proposed final control but was incompletely specified. Reviewers were pharmacists, physicians, researchers, or mixed panels; their number, specialty, blinding, source access, and adjudication procedures differed. Few studies measured review time, cognitive burden, automation bias, correction rates, or residual error after review. Escalation criteria were usually implicit rather than operationalized. Human involvement should therefore be described as an active verification and authority function, not a generic statement that a person remains "in the loop."

Evidence-Conflict Handling

Evidence conflict was the least developed domain. Retrieved documents could disagree because of publication date, population, dose, outcome definition, regulatory jurisdiction, or evidence hierarchy. Most systems silently selected or summarized retrieved content rather than exposing disagreement. Few reported rules for prioritizing current guidelines, primary evidence, product information, or patient-specific contraindications. Medication-review and reconciliation studies illustrated the need to reconcile dosing, interaction, therapeutic-monitoring, and pharmacogenomic information across several evidence types [29]. Cross-sectional comparisons with licensed pharmacists showed favourable average ratings for some model responses but did not establish that conflicting evidence could be resolved safely without professional adjudication [30]. Real-world pharmacotherapy-counselling comparisons further indicated that clinicians remained stronger at contextualizing uncertainty and integrating patient-specific considerations [31].

A review-derived conflict pathway was therefore constructed: detect incompatible claims; preserve source identity and date; characterize whether disagreement reflects population, intervention, outcome, jurisdiction, or recency; display uncertainty; and escalate when resolution could alter a medication decision. This pathway is an interpretive synthesis, not an established clinical algorithm.

Evaluation Gaps

Methodological appraisal identified small and convenience-based question sets, incomplete prompt reporting, unstable model versions, nonblinded expert ratings, locally developed rubrics, weak denominator reporting, and limited replication. Reference standards ranged from one reviewer's judgment to multidisciplinary consensus or comparison with institutional drug-information answers. Few studies used prospective patients, measured workflow consequences, assessed equity across languages or settings, or examined whether users could detect unsupported claims.

Independent replication was especially uncommon, and evaluations rarely preserved complete prompts, retrieved passages, model settings, timestamps, and reviewer corrections needed to reproduce individual evidential failures across updates.

Evaluation settings also omitted several consequential outcomes. No included report compared escalation thresholds across medication-risk levels, quantified downstream prescribing changes, or established how often a model should abstain when evidence quality, applicability, or recency was uncertain.

Discussion: Principal Findings and Interpretation

Principal Map of the Field

This review found that evidence tracing in generative medicines information is not one function but a chain of separable operations. A model may retrieve a relevant document yet cite the wrong passage, summarize the passage inaccurately, omit a limiting condition, combine incompatible sources, or present an unsupported recommendation. Conversely, an answer may contain no fabricated references but remain unsafe because evidence is incomplete, outdated, or inapplicable to the patient. Evaluation should therefore distinguish retrieval success, source identity, source quality, bibliographic correctness, claim-level support, contradiction, completeness, and decision relevance.

The evidence base nevertheless tended to compress this chain into broad labels such as accuracy, hallucination, grounding, or citation quality. RAG reviews in medicine similarly report substantial variation in knowledge bases, retrieval pipelines, models, metrics, and human evaluation, with limited prospective clinical validation [32]. Traceability requires the source and supporting passage to remain inspectable, while trustworthiness requires a justified relationship between that passage and the exact generated claim.

Figure 1 presents the original conceptual synthesis for evidence-tracing, hallucination, and safeguard taxonomy for generative medicines information, showing how the article's principal components, evidence relationships, uncertainties, and decision boundaries are connected.

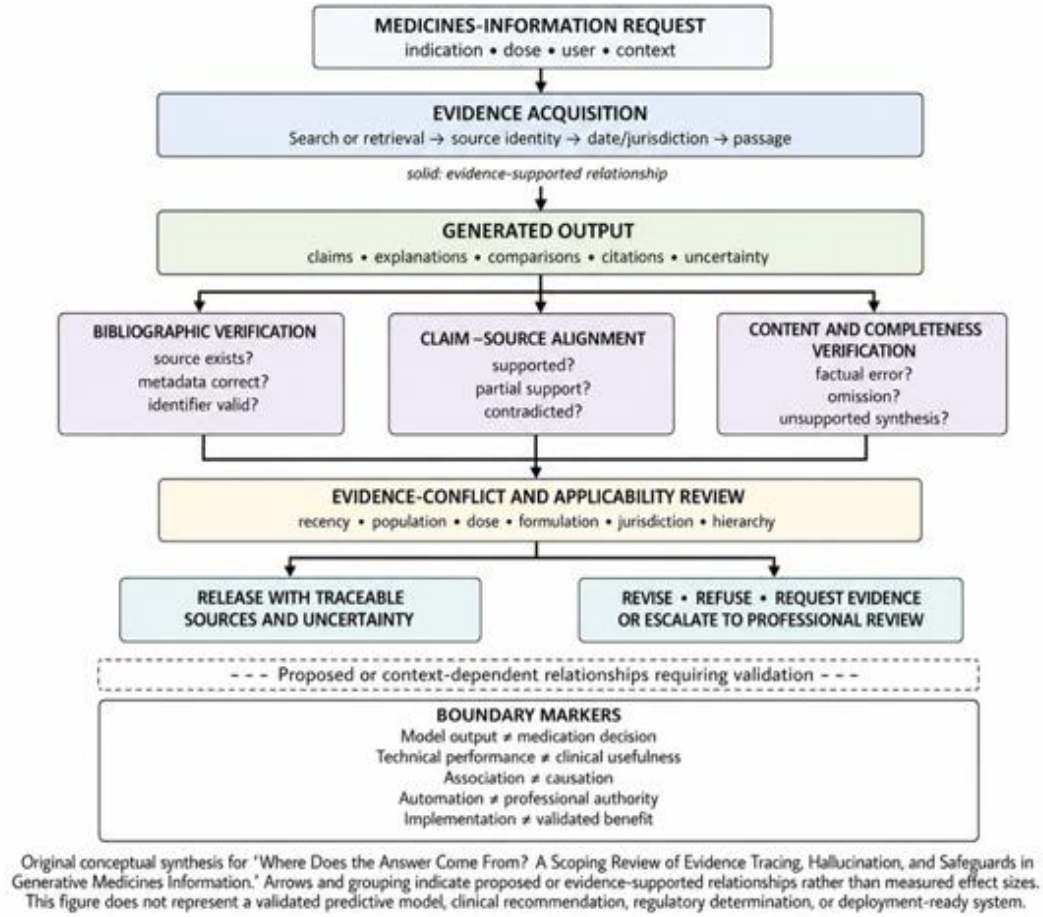

Figure 1. Evidence-Tracing, Hallucination, and Safeguard Taxonomy for Generative Medicines Information

Terminological and Measurement Inconsistency

Hallucination was the most unstable term. Studies used it for nonexistent references, incorrect metadata, unsupported claims, factual mistakes, or any disagreement with an evaluator. These phenomena have different causes and safeguards. Bibliographic fabrication can be detected through identifier verification; unsupported synthesis requires claim–source comparison; omission requires a completeness standard; and inappropriate medication advice requires contextual professional judgment. A clinical-safety framework for medical summarization likewise separates factual consistency from the potential severity of resulting error, showing why a single hallucination rate may not represent clinical risk [33].

Measurement inconsistency also arose from the unit of analysis. An answer-level measure classifies a response with one minor unsupported statement in the same way as a response dominated by fabrication. Citation-level measures may ignore uncited claims. Field-level bibliographic scores

detect corrupted metadata without determining entailment. Automated judges permit large-scale analysis but may inherit model biases and require validation against expert annotations. Strategies proposed for medical hallucination reduction consequently span retrieval, prompting, uncertainty calibration, verification, and human oversight rather than one universal correction [34].

A minimum reporting set should identify the model; prompt; evidence corpus; retrieval and reranking procedure; claim-segmentation rule; source passage; verification method; reviewer expertise; disagreement process; denominator; error severity; and whether results were reproduced after a model update. Without these details, apparently improved performance may reflect easier prompts, narrower evidence, different scoring, or selective reporting.

Safeguards Supported Only in Laboratory Settings

Most safeguards were evaluated with static question sets, curated documents, or retrospective cases. Such designs are valuable for detecting technical failure modes but cannot establish how a safeguard behaves when users reformulate questions, evidence changes, multiple medicines interact, or time pressure affects verification. Laboratory improvements may also shift work rather than remove it: retrieval can reduce open-ended fabrication while requiring users to inspect more passages, resolve contradictory documents, and determine applicability.

Human review was often treated as a dependable fallback, although the included studies rarely measured whether reviewers noticed persuasive unsupported statements or became over-reliant on cited outputs. Broader evidence on human–AI combinations shows that adding a person and an AI does not automatically outperform the stronger component; performance depends on task allocation and complementarity [35]. For medicines information, professional review must include authority to reject, revise, request better evidence, and escalate—not merely observe the output.

Implications, Strengths, and Limitations

Design and Research Implications

A defensible system should expose an evidence trail at claim level. The interface should distinguish retrieved text from generated interpretation, show source dates and jurisdictions, identify unsupported claims, and retain alternative or conflicting evidence. It should also separate technical confidence from clinical authority. The convergence of human expertise and artificial intelligence is most valuable when systems augment evidence handling while professionals retain contextual judgment and responsibility [36].

Evaluation should proceed through staged evidence maturity: controlled citation testing; external and temporally updated datasets; simulated workflows; prospective studies measuring reviewer effort and residual error; and monitored operational evaluation. High average accuracy should not compensate for rare, consequential unsupported medication claims. Prespecified escalation should apply when sources conflict, evidence is absent, recommendations depend on patient-specific information, or an error could change treatment.

Governance should treat material changes in models, retrieval corpora, interfaces, or prompting as reasons for reassessment. Commentary on medical-device oversight has emphasized that conversational systems can produce health-relevant outputs while changing rapidly and obscuring conventional product boundaries [37]. This review does not determine regulatory classification, but it supports auditable versioning, defined intended use, restricted permissions, fallback procedures, and postimplementation monitoring.

Limitations

The search was limited to peer-reviewed journal reports with sufficient English information, potentially excluding relevant technical work. Rapid model changes may have overtaken some evaluations before publication. Heterogeneity prevented pooled estimates, and appraisal reflected reporting quality as well as methodological quality. Some included studies assessed general medical citations rather than exclusively pharmacy outputs, although they contributed directly to evidence-tracing methods. Finally, the taxonomy is a review-derived synthesis requiring empirical testing across medicines-information settings.

Conclusion

Generative medicines-information systems increasingly provide fluent answers with citations, retrieved documents, or other signals of evidential grounding. This review shows that such signals cannot be interpreted as proof that a response is correct, complete, current, or suitable for a medication decision. Evidence tracing consists of distinct operations: identifying sources, preserving provenance, verifying bibliographic details, locating supporting passages, testing claim–source alignment, detecting contradiction and omission, and documenting professional adjudication.

The reviewed literature was concentrated in retrospective, cross-sectional, proof-of-concept, and simulated evaluations. Model-generated citations were vulnerable to fabrication and metadata error. Search-enabled and retrieval-augmented systems improved access to bounded evidence but remained susceptible to retrieval failure, partial support, outdated information, conflicting documents, and unsupported synthesis. Human verification was widely invoked as a safeguard, yet reviewer workload, correction performance, escalation thresholds, and residual risk were seldom evaluated prospectively.

Future evaluation should report claims, sources, passages, denominators, error severity, model versions, prompts,

retrieval settings, reviewer procedures, and conflicts transparently. Systems should distinguish retrieved evidence from generated interpretation and should escalate when evidence is absent, contradictory, temporally unstable, jurisdiction-dependent, or contingent on patient-specific information.

Research must test whether these controls reduce consequential errors without excessive verification burden.

The proposed taxonomy organizes these requirements but is an interpretive synthesis rather than a validated clinical framework. Current evidence does not establish autonomous clinical safety, universal applicability, regulatory acceptance, or deployment readiness. Generative medicines information should therefore be assessed as an evidence-handling process whose outputs remain conditional on traceable sources, explicit uncertainty, and accountable professional judgment.

ACKNOWLEDGMENTS: None
CONFLICT OF INTEREST: None
FINANCIAL SUPPORT: None
ETHICS STATEMENT: None

REFERENCES

1. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. Nat Med. 2023;29(8):1930-40. doi:10.1038/s41591-023-02448-8
2. Moor M, Banerjee O, Abad ZSH, Krumholz HM, Leskovec J, Topol EJ, et al. Foundation models for generalist medical artificial intelligence. Nature. 2023;616(7956):259-65. doi:10.1038/s41586-023-05881-4
3. Lee P, Bubeck S, Petro J. Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. N Engl J Med. 2023;388(13):1233-9. doi:10.1056/NEJMsr2214184
4. Howell MD. Generative artificial intelligence, patient safety and healthcare quality: A review. BMJ Qual Saf. 2024;33(11):748-54. doi:10.1136/bmjqs-2023-016690
5. Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, et al. PRISMA extension for scoping reviews (PRISMA-ScR): Checklist and explanation. Ann Intern Med. 2018;169(7):467-73. doi:10.7326/M18-0850
6. Peters MDJ, Marnie C, Tricco AC, Pollock D, Munn Z, Alexander L, et al. Updated methodological guidance for the conduct of scoping reviews. JBI Evid Synth. 2020;18(10):2119-26. doi:10.11124/JBIES-20-00167
7. Munn Z, Peters MDJ, Stern C, Tufanaru C, McArthur A, Aromataris E. Systematic review or scoping review? Guidance for authors when choosing between a systematic or scoping review approach. BMC Med Res Methodol. 2018;18(1):143. doi:10.1186/s12874-018-0611-x
8. Ong JCL, Chen MH, Ng N, Elangovan K, Tan NYT, Jin L, et al. A scoping review on generative AI and large language models in mitigating medication related harm. NPJ Digit Med. 2025;8(1):182. doi:10.1038/s41746-025-01565-7
9. Peters MDJ, Godfrey C, McInerney P, Khalil H, Larsen P, Marnie C, et al. Best practice guidance and reporting items for the development of scoping review protocols. JBI Evid Synth. 2022;20(4):953-68. doi:10.11124/JBIES-21-00242
10. Pollock D, Davies EL, Peters MDJ, Tricco AC, Alexander L, McInerney P, et al. Undertaking a scoping review: A practical guide for nursing and midwifery students, clinicians, researchers, and academics. J Adv Nurs. 2021;77(4):2102-13. doi:10.1111/jan.14743
11. Bramer WM, de Jonge GB, Rethlefsen ML, Mast F, Kleijnen J. A systematic approach to searching: An efficient and complete method to develop literature searches. J Med Libr Assoc. 2018;106(4):531-41. doi:10.5195/jmla.2018.283
12. Rethlefsen ML, Kirtley S, Waffenschmidt S, Ayala AP, Moher D, Page MJ, et al. PRISMA-S: An extension to the PRISMA Statement for reporting literature searches in systematic reviews. Syst Rev. 2021;10(1):39. doi:10.1186/s13643-020-01542-z
13. Pollock D, Peters MDJ, Khalil H, McInerney P, Alexander L, Tricco AC, et al. Recommendations for the extraction, analysis, and presentation of results in scoping reviews. JBI Evid Synth. 2023;21(3):520-32. doi:10.11124/JBIES-22-00123
14. Aljamaan F, Temsah MH, Altamimi I, Al-Eyadhy A, Jamal A, Alhasan K, et al. Reference Hallucination Score for medical artificial intelligence chatbots: Development and usability study. JMIR Med Inform. 2024;12. doi:10.2196/54345
15. Walters WH, Wilder EI. Fabrication and errors in the bibliographic citations generated by ChatGPT. Sci Rep. 2023;13(1):14045. doi:10.1038/s41598-023-41032-5
16. Sebo P. How accurate are the references generated by ChatGPT in internal medicine? Intern Emerg Med. 2024;19(1):247-9. doi:10.1007/s11739-023-03484-5
17. Morath B, Chiriac U, Jaszkowski E, Deiß C, Nürnberg H, Hörth K, et al. Performance and risks of ChatGPT used in drug information: An exploratory real-world analysis. Eur J Hosp Pharm. 2024;31(6):491-7. doi:10.1136/ejhpharm-2023-003750
18. Huang X, Estau D, Liu X, Yu Y, Qin J, Li Z. Evaluating the performance of ChatGPT in clinical pharmacy: A comparative study of ChatGPT and clinical pharmacists. Br J Clin Pharmacol. 2024;90(1):232-8. doi:10.1111/bcp.15896
19. Roosan D, Padua P, Khan R, Khan H, Verzosa C, Wu Y. Effectiveness of ChatGPT in clinical pharmacy and the role of artificial intelligence in medication therapy management. J Am Pharm Assoc (2003). 2024;64(2):422-8.e8. doi:10.1016/j.japh.2023.11.023
20. Chelli M, Descamps J, Lavoué V, Trojani C, Azar M, Deckert M, et al. Hallucination rates and reference accuracy of ChatGPT and Bard for systematic reviews: Comparative analysis. J Med Internet Res. 2024;26. doi:10.2196/53164
21. Taş N, Erden Y, Temel MH, Bağcıer F. A reproducible statistical evaluation framework for large-sample assessment of AI-generated medical references: Cross-platform application of the Reference Hallucination Score. BMC Med Res Methodol. 2026. Online ahead of print. doi:10.1186/s12874-026-02906-0
22. Wu K, Wu E, Wei K, Zhang A, Casasola A, Nguyen T, et al. An automated framework for assessing how well large language models cite relevant medical references. Nat Commun. 2025;16(1):3615. doi:10.1038/s41467-025-58551-6
23. Amugongo LM, Mascheroni P, Brooks S, Doering S, Seidel J. Retrieval augmented generation for large language models in healthcare: A systematic review. PLOS Digit Health. 2025;4(6). doi:10.1371/journal.pdig.0000877
24. Wang C, Chen Y. Evaluating large language models for evidence-based clinical question answering. Patterns (N Y). 2026;7(5):101519. doi:10.1016/j.patter.2026.101519
25. de Jesus DR, de Souza Júnior AP, de Albergaria ET, Pagano AS, de Oliveira IJR, Dias CDS, et al. Enhanced LLM-supported instructions for medication use through retrieval-augmented generation. Comput Biol Med. 2025;198(Pt A):111135. doi:10.1016/j.compbiomed.2025.111135
26. Frazer A, Yoon E, Durant KM. Artificial intelligence for drug information: Evaluating its use in real-world pharmacy practice. Am J Health Syst Pharm. 2026;83(12):669-76. doi:10.1093/ajhp/zxaf350
27. van Nuland M, Lobbezoo AFH, van de Garde EMW, Herbrink M, van Heijl I, Bognár T, et al. Assessing accuracy of ChatGPT in response to questions from day to day pharmaceutical care in hospitals. Explor Res Clin Soc Pharm. 2024;15:100464. doi:10.1016/j.rcsop.2024.100464
28. Triplett S, Ness-Engle GL, Behnen EM. A comparison of drug information question responses by a drug information center and by ChatGPT. Am J Health Syst Pharm. 2025;82(8):448-60. doi:10.1093/ajhp/zxae316
29. Sridharan K, Sivaramakrishnan G. Unlocking the potential of advanced large language models in medication review and reconciliation: A proof-of-concept investigation. Explor Res Clin Soc Pharm. 2024;15:100492. doi:10.1016/j.rcsop.2024.100492
30. Albogami Y, Alfakhri A, Alaqil A, Alkoraishi A, Alshammari H, Elsharawy Y, et al. Safety and quality of AI chatbots for drug-related

inquiries: A real-world comparison with licensed pharmacists. Digit Health. 2024;10:20552076241253523. doi:10.1177/20552076241253523

31. Krichevsky B, Engeli S, Bode-Böger SM, Koop F, Schulze Westhoff M, Schröder S, et al. Human vs. artificial intelligence: Physicians outperform ChatGPT in real-world pharmacotherapy counselling. Br J Clin Pharmacol. 2026;92(3):891-902. doi:10.1002/bcp.70321
32. Yang R, Wong MYH, Li H, Li X, Zhu W, Liao J, et al. Retrieval-augmented generation in medicine: A scoping review of technical implementations, clinical applications, and ethical considerations. Cell Rep Med. 2026. Online ahead of print. Article 102927. doi:10.1016/j.xcrm.2026.102927
33. Asgari E, Montaña-Brown N, Dubois M, Khalil S, Balloch J, Au Yeung J, et al. A framework to assess clinical safety and hallucination rates of LLMs for medical text summarisation. NPJ Digit Med. 2025;8(1):274. doi:10.1038/s41746-025-01670-7
34. Chen F, Li Y, Chen Y, Bian Z, Duo L, Zhou Q, et al. Strategies for the analysis and elimination of hallucinations in artificial intelligence generated medical knowledge. J Evid Based Med. 2025;18(3). doi:10.1111/jebm.70075
35. Vaccaro M, Almaatouq A, Malone T. When combinations of humans and AI are useful: A systematic review and meta-analysis. Nat Hum Behav. 2024;8(12):2293-303. doi:10.1038/s41562-024-02024-1
36. Topol EJ. High-performance medicine: The convergence of human and artificial intelligence. Nat Med. 2019;25(1):44-56. doi:10.1038/s41591-018-0300-7
37. Gilbert S, Harvey H, Melvin T, Vollebregt E, Wicks P. Large language model AI chatbots require approval as medical devices. Nat Med. 2023;29(10):2396-8. doi:10.1038/s41591-023-02412-6