Original Article

Evidence-Bound Generation for Medicines Information with Verifiable Claims and Visible Uncertainty

Luca Ferraro[1]*, Matteo Ricci[1], Giulia Moretti[2], Paolo Greco[3]

[1]Department of Medicines Information and Evidence Generation, Faculty of Pharmacy, University of Bologna, Bologna, Italy. [2]Department of Verifiable AI Claims in Pharmacy, Faculty of Pharmacy, University of Turin, Turin, Italy. [3]Department of Uncertainty Communication in Pharmacy AI, Faculty of Pharmacy, Sapienza University of Rome, Rome, Italy.

Abstract

Generative artificial intelligence can produce fluent medicines information while obscuring whether individual claims are supported, qualified, contradicted, or unsupported by current evidence. Retrieval augmentation may improve access to external information, but retrieval alone does not establish source eligibility, claim faithfulness, clinical applicability, or safe decision use. This article develops an original, non-empirical Retrieval–Evidence–Generation Architecture for evidence-bound medicines information. Evidence-bound generation is defined as a proposed form of generation in which externally verifiable claims are constrained by eligible evidence, linked to inspectable evidence units, assigned explicit support states, and accompanied by visible uncertainty, conflict, or evidence absence. The architecture separates task specification, governed retrieval, evidence construction, claim planning, bounded generation, claim-level verification, professional review, and lifecycle monitoring. It further proposes a provenance chain connecting each displayed claim to its supporting passage, source identity, version, retrieval context, and transformation history. Five qualitative claim states—supported, supported with qualification, conflicting evidence, unsupported, and evidence absent—are proposed to prevent citations or model confidence from functioning as substitutes for verification. Validation would require technical assessment of retrieval and evidence binding, pharmacist assessment of medication-content correctness and harmful omission, human-factors testing of reliance and verification burden, equity assessment, and prospective workflow evaluation. The architecture does not establish clinical benefit, medication-safety improvement, regulatory acceptance, or deployment readiness. Its original contribution is to make evidence control, uncertainty, professional authority, and governance integral to medicines-information generation rather than optional post-generation checks.

Keywords: Digital pharmacy, Artificial intelligence, Pharmacy practice, Medication safety, Human–AI collaboration, Clinical decision support

Introduction

Large language models can produce coherent medical explanations, syntheses, and conversational answers, yet their clinical promise remains inseparable from hallucination, privacy, reliability, and misuse risks [1]. In medicines information, linguistic plausibility is particularly insufficient because small errors involving dose, formulation, contraindication, interaction, monitoring, or patient context may alter the meaning and potential consequence of an answer.

Strong performance on medical question-answering benchmarks shows that large models can encode and reproduce substantial clinical knowledge, but benchmark competence does not demonstrate that a generated statement is grounded in identifiable, current, and applicable evidence [2]. Evaluators may also prefer the style, apparent completeness, or empathy of generated responses without

Address for correspondence: Luca Ferraro, Department of Medicines Information and Evidence Generation, Faculty of Pharmacy, University of Bologna, Bologna, Italy.
luca.ferraro@unibo.it
Received: 12 November 2025; **Accepted:** 19 February 2026

This is an open-access article distributed under the terms of the Creative Commons Attribution-Non Commercial-Share Alike 4.0 License, which allows others to remix, tweak, and build upon the work non commercially, as long as the author is credited and the new creations are licensed under the identical terms.

How to cite this article: Ferraro L, Ricci M, Moretti G, Greco P. Evidence-Bound Generation for Medicines Information with Verifiable Claims and Visible Uncertainty. Arch Pharm Pract. 2026;17(1):48-56. https://doi.org/10.51847/TzCg1KSqsZ

independently testing the provenance of their claims [3]. Fluency and perceived usefulness must therefore be distinguished from evidence traceability and decision fitness.

Medication-specific evaluations reinforce this distinction. Real-world drug-information responses can contain incomplete, inaccurate, or potentially unsafe content despite appearing authoritative [4]. Even in bounded clinical-text extraction, performance can vary with prompting and specialist interpretation, indicating that source access does not remove model-mediated error [5]. Alternative explanations for apparently strong answers include memorised associations, broad probabilistic patterning, permissive grading, or retrieval of information that is relevant but not sufficiently applicable.

This article proposes evidence-bound generation as a stricter conceptual standard. An evidence-bound answer is not merely accompanied by references. Each externally verifiable claim should be constrained by eligible evidence, connected to inspectable evidence units, assigned a support state, and presented with uncertainty or evidence absence that remains visible to the intended user. The contribution is an original architecture linking retrieval, evidence qualification, claim construction, provenance, conflict handling, professional review, and governance. These relationships are proposed rather than empirically validated.

Requirements for Evidence-Bound Medicines Information

Evidence-bound medicines information should first be treated as a form of clinical decision support whose intended user, task, workflow position, and consequences are explicitly defined [6]. Requirements should extend across development, validation, stakeholder involvement, monitoring, and change control rather than being reduced to model selection or prompt design [7]. A patient-facing explanation, pharmacist reference answer, formulary summary, interaction assessment, and clinician decision aid impose different evidence, language, and escalation requirements.

Explanation, citation, evidence support, and provenance are related but non-equivalent properties. Explainability is meaningful only relative to its audience and purpose; an understandable rationale may still be unfaithful, irrelevant, or unsupported [8]. A citation may be authentic but fail to entail the attached claim, while a correct claim may omit material qualifications. Provenance identifies where information came from and how it was transformed, but does not independently establish source truth or patient-level applicability.

Clinical evaluation must also examine workflow, human factors, errors, and unintended consequences rather than technical accuracy alone [9]. Medication-question evaluation further shows why plausible wording should not replace professional verification of correctness, completeness, and appropriateness [10]. The proposed minimum requirements are therefore: defined intended use; governed source eligibility; inspectable evidence units; claim-level support; explicit conflict and absence states; visible uncertainty; risk-sensitive escalation; reproducible provenance; and a boundary separating generated information from medication decisions.

Table 1 organizes the evidence, constructs, relationships, uncertainties, and boundary conditions required for definitions, components, relationships, and intended uses for the article Evidence-Bound Generation for Medicines Information with Verifiable Claims and Visible Uncertainty.

Table 1. Definitions, components, relationships, and intended uses for the article Evidence-Bound Generation for Medicines Information with Verifiable Claims and Visible Uncertainty.

Component or scientific dimension	Problem addressed	Inputs or determinants	Proposed mechanism or relationship	Expected contribution	Evidence required	Failure or uncertainty risk	Boundary statement
Intended-use specification	One answer form may be applied across incompatible tasks	User, question, workflow, urgency, available context	**Proposed:** bind system behaviour to a declared task and user	More interpretable evaluation and escalation	Workflow and user evaluation [6, 9]	Hidden change in intended use	Intended use does not establish clinical utility
Source eligibility	Retrieved material may be irrelevant, obsolete, or unauthoritative	Source type, date, jurisdiction, population, study design	**Proposed:** apply eligibility rules before generation	Restriction to inspectable candidate evidence	Source-policy validation [7]	Relevant evidence may be excluded	Eligibility does not prove source truth
Evidence unit	Long documents obscure the passage supporting a claim	Source passage, table, structured record, metadata	**Proposed:** construct the smallest inspectable support unit	Enables claim-level checking	Extraction and pharmacist adjudication [5]	Passage may lose context	Evidence units are not independent clinical recommendations
Claim–evidence binding	Answer-level citations conceal unsupported statements	Atomic claim and eligible evidence unit	**Proposed:** assign an explicit support relationship	Verifiable claim coverage	Entailment and citation testing	Citation may be authentic but non-supporting	Binding does not establish applicability

Provenance	Outputs cannot be reconstructed or corrected	Source identity, location, version, retrieval and transformation record	**Proposed:** preserve an end-to-end trace	Audit and correction capability	Reproducibility assessment [7]	Metadata may be incomplete	Provenance is not proof of correctness
Visible uncertainty	Limitations are hidden behind fluent language	Evidence quality, retrieval limits, applicability and model interpretation	**Proposed:** attach uncertainty to the affected claim	More transparent reliance conditions	User-comprehension and calibration testing [8]	Disclosure may confuse or falsely reassure	Uncertainty display is not a validated safety intervention
Conflict or absence	Systems may force consensus or speculate	Contradictory eligible sources or no eligible support	**Proposed:** preserve conflict and absence as distinct states	Reduced false certainty	Conflict-detection and abstention testing	Relevant evidence may have been missed	Absence is not evidence of no effect
Professional boundary	Generated information may be treated as a medication decision	Claim risk, patient specificity, urgency and professional scope	**Proposed:** trigger pharmacist review, qualification, withholding, or referral	Retains professional authority	Human-factors and workflow evaluation [9, 10]	Nominal review may be ineffective	The architecture does not make the medication decision

These requirements form a proposed specification for design and evaluation. Their inclusion would not establish improved medication safety, universal applicability, or readiness for routine clinical use.

Proposed Retrieval–Evidence–Generation Architecture

Retrieval augmentation can improve access to external knowledge and may increase performance on bounded medical tasks, but observed gains remain dependent on task design, corpus quality, and evaluation context [11]. The proposed architecture therefore treats retrieval as an upstream evidence-discovery function rather than as proof that a final answer is grounded.

Medical retrieval-augmented generation is sensitive to context ordering, density, and the presence of competing information [12]. Layered retrieval and summarisation can separate local evidence processing from final answer synthesis, although current proof-of-concept results do not validate such separation for medication decisions [13]. The architecture consequently places an evidence-construction layer between retrieval and generation. Retrieved records are screened for source authority, recency, jurisdiction, formulation, population, and relevance before being transformed into inspectable evidence units.

Figure 1 presents the original conceptual synthesis for retrieval–evidence–generation architecture for medicines information, showing how the article's principal components, evidence relationships, uncertainties, and decision boundaries are connected.

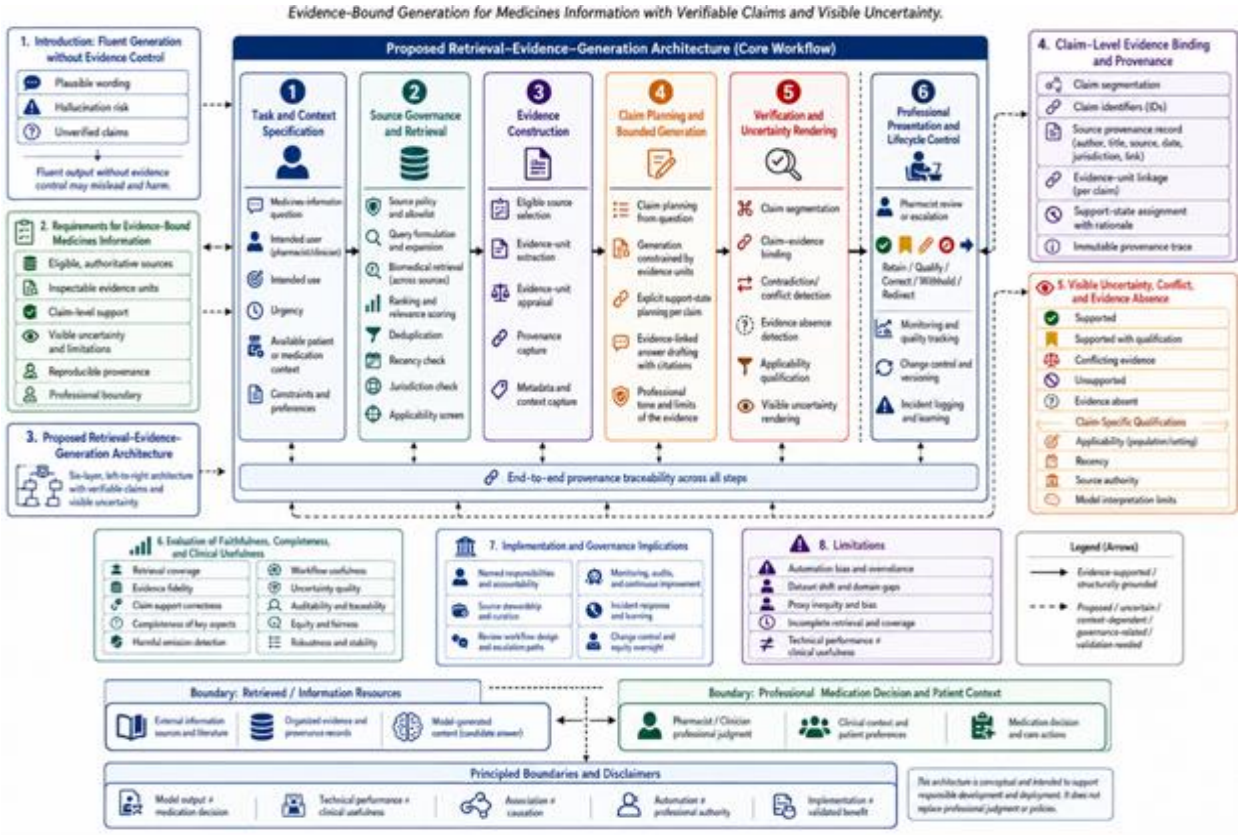

Figure 1. Retrieval–Evidence–Generation Architecture for Medicines Information

The architecture contains six proposed layers. First, task and context specification identifies the question, intended user, intended use, urgency, and available medication or patient context. Second, source governance and retrieval applies an approved source policy and generates queries. Domain-trained biomedical retrieval can improve identification of relevant literature, but ranking relevance remains distinct from evidence validity [14]. Third, evidence construction deduplicates results, checks eligibility, extracts evidence units, and records exclusions or limitations.

Fourth, claim planning and bounded generation converts the requested answer into planned propositions before prose generation. Each planned claim should be associated with eligible evidence or designated as requiring qualification, abstention, or professional escalation. Fifth, verification and uncertainty rendering reassesses generated claims for support, contradiction, omission, applicability, and provenance. Sixth, professional presentation and lifecycle control determines whether the answer can be displayed, qualified, corrected, withheld, or referred.

Online retrieval may improve access to current authoritative material, but its effects vary across tasks and models and introduce dependencies on source availability and version control [15]. Dynamic retrieval should therefore trigger renewed verification rather than silent source replacement. Feedback from incidents, evidence updates, user behaviour, and detected shift should return to source policy, evaluation, and change control. These feedback relations are proposed and require prospective validation.

Table 2 organizes the evidence, constructs, relationships, uncertainties, and boundary conditions required for evaluation criteria, evidence requirements, and failure modes for the article Evidence-Bound Generation for Medicines Information with Verifiable Claims and Visible Uncertainty.

Table 2. Evaluation criteria, evidence requirements, and failure modes for the article Evidence-Bound Generation for Medicines Information with Verifiable Claims and Visible Uncertainty.

Architecture layer or process stage	Core function	Information flow	Interaction with other components	Evaluation criterion	Potential failure mode	Human or organizational responsibility	Readiness boundary

Task specification	Define intended use and available context	Question to structured task record	Constrains retrieval, generation, and escalation	Completeness and correctness of task representation [6, 9]	Missing formulation, population, urgency, or purpose	Clinical owner and product owner	A defined task is not a validated use
Source governance	Determine eligible source classes	Source policy to retrieval filter	Sets authority, recency, and jurisdiction rules	Policy consistency and expert agreement [7]	Low-authority or outdated material admitted	Evidence-governance lead	Approved sources may still contain weak evidence
Biomedical retrieval	Identify candidate records	Query to ranked documents	Supplies evidence-construction layer	Retrieval relevance, coverage, and robustness [11, 14, 15]	Relevant evidence omitted or poorly ranked	Technical and information-specialist team	Retrieval performance is not answer faithfulness
Evidence construction	Produce inspectable evidence units	Documents to passages and metadata	Feeds claim planning and provenance	Extraction fidelity and context preservation [5, 12]	Passage misinterpretation or loss of qualification	Pharmacy-content reviewer	Extracted text is not automatically applicable
Claim planning	Specify verifiable propositions	Evidence units to planned claims	Constrains generation	Claim coverage and unsupported-claim detection	Important claims omitted or generated without support	Clinical and technical design teams	Planned claims require empirical testing
Bounded generation	Produce prose within evidence constraints	Claim plan to draft answer	Receives evidence links and restrictions	Adherence to claim plan and instruction robustness [13]	Unsupported elaboration or distorted synthesis	Model owner	Prompt compliance is not a safety guarantee
Verification and uncertainty	Assign support, conflict, absence, and applicability states	Draft claims to verified claim record	Controls display and escalation	Claim support, contradiction, omission, and uncertainty quality	Authentic citation attached to an unsupported claim	Pharmacist reviewer and evaluation team	Verification methods require external validation
Presentation and escalation	Display, qualify, withhold, or refer	Verified record to user-facing answer	Preserves decision boundary	Comprehension, reliance, review burden, and recovery from error [9, 10]	Automation bias or unusable qualification	Pharmacy service and governance body	Usability does not establish clinical benefit
Monitoring and change control	Detect deterioration and manage updates	Runtime records to governance feedback	Revises sources, prompts, models, and policies	Traceability of changes and repeat evaluation [7]	Silent performance decay or unreconstructable updates	Named lifecycle owner	Monitoring cannot guarantee detection of every failure

The architecture is intended to organise functions and responsibilities, not to prescribe one technical implementation. Its clinical and organisational value remains conditional on validation within defined medicines-information tasks and workflows.

Claim-Level Evidence Binding and Provenance

The primary verification unit should be the externally verifiable claim, not the answer as a whole. A response can contain an authentic reference while still misrepresenting what the source states, attaching the source to the wrong proposition, or generating additional unsupported detail. Comparative analyses of generated literature references demonstrate that citation existence and citation accuracy require separate assessment [16].

Claim-level control begins by segmenting the answer into propositions that can be checked independently. Citation-retrieval systems can then be evaluated separately for source identification, citation correctness, and narrative synthesis rather than receiving one aggregate score [17]. Evidence resources that are continuously updated similarly depend on structured ingestion, classification, and traceable revision processes [18]. The proposed architecture extends these principles by assigning every claim a stable identifier and linking it to one or more evidence-unit identifiers.

A provenance record should include source identity, source type, version or date, passage location, retrieval time, jurisdiction, eligibility decision, and any transformation applied before generation. Transparent recording of inputs, methods, and analytical decisions is central to reproducibility, even when hosted models or changing retrieval environments prevent exact deterministic replication [19]. The intended standard is therefore reconstructability: an auditor should be able to determine why a claim appeared, which evidence was used, which qualifications were preserved or lost, and which system version produced it.

Each claim should receive one proposed support state: supported, supported with qualification, conflicting evidence, unsupported, or evidence absent. These states describe the relationship between a claim and the available eligible evidence; they do not measure clinical importance or

establish source truth. High-risk or patient-specific claims should additionally carry a review requirement independent of apparent support.

Plausible explanations cannot be assumed to reveal faithful model reasoning or justify clinical reliance [20]. Explanation, citation accuracy, and provenance are distinct controls and none alone demonstrates faithful evidence support [8, 16, 20]. Claim-level binding can make the evidence relationship inspectable, but correctness still depends on retrieval coverage, source quality, accurate interpretation, preserved context, and appropriate professional judgement. The proposed mechanism must therefore be validated against pharmacist-adjudicated claims, deliberate source conflicts, missing-evidence conditions, misleading but authentic citations, and changes in the evidence base. It does not convert generation into autonomous medication decision making.

Visible Uncertainty, Conflict, and Evidence Absence

Evidence-bound generation requires uncertainty to be represented as a property of specific claims and evidence relationships rather than hidden behind fluent prose. Available large-language-model uncertainty proxies differ in their ability to discriminate correct from incorrect medical outputs and may remain poorly calibrated [21]. The architecture therefore does not treat verbal confidence, token probability, answer consistency, or model self-assessment as interchangeable measures of clinical certainty.

Uncertainty quantification and uncertainty communication serve different functions. Communication must help the intended user understand what is uncertain, why it is uncertain, and what action is appropriate without creating false reassurance or unusable ambiguity [22]. Quantification may support internal testing, but high-stakes machine-assisted decisions require explicit recognition of data limitations, model limitations, and uncertainty sources [23]. Numerical confidence should not be displayed as a probability unless its meaning and calibration have been established for the defined task and setting [24].

Five proposed support states provide a qualitative alternative: supported, supported with qualification, conflicting evidence, unsupported, and evidence absent. Conflict means that eligible sources support materially different conclusions; evidence absence means that no eligible source supports the requested proposition. These states should remain distinct because disagreement, incomplete retrieval, and a genuinely sparse evidence base require different responses.

Visible uncertainty should identify whether the limitation originates in evidence quality, retrieval coverage, applicability, source conflict, or model interpretation. The display may include a claim-specific qualification, conflicting-source summary, statement of missing evidence, restricted answer, or escalation to a pharmacist. Such displays require user-comprehension and reliance testing. They are proposed information controls, not validated medication-safety interventions.

Evaluation of Faithfulness, Completeness, and Clinical Usefulness

Evaluation should distinguish faithfulness, completeness, and clinical usefulness. Faithfulness asks whether each claim is supported by its linked evidence. Completeness asks whether material benefits, harms, contraindications, alternatives, and qualifications have been omitted. Clinical usefulness asks whether the answer is understandable, timely, applicable, and appropriately integrated into pharmacy work. Existing health-care large-language-model evaluations remain heterogeneous and frequently underrepresent realistic workflows, safety consequences, and clinically meaningful outcomes [25].

Minimum reporting should specify intended use, data and source selection, model configuration, validation design, reference standards, generalisability, and limitations [26]. Generative systems additionally require documentation of prompting, open-ended output handling, bias, security, reproducibility, and human evaluation [27]. Checklist compliance supports transparency but does not establish correctness or clinical benefit.

Comparisons with clinicians require appropriate reference standards, representative cases, independent adjudication, and evaluation of human–AI teams rather than unsupported claims of replacement or superiority [28]. Uncertainty discrimination, calibration, and multidimensional clinical evaluation should be assessed separately [21, 24, 25]. Proposed evaluation should include retrieval coverage, evidence-unit fidelity, claim segmentation, citation entailment, harmful omission, conflict detection, abstention, pharmacist verification burden, subgroup performance, and recovery from error. No single metric should determine readiness.

Implementation and Governance Implications

Evidence-bound generation distributes responsibility across evidence selection, model behaviour, pharmacy review, information governance, implementation, and monitoring. Governance should assign named authority for fairness, transparency, trustworthiness, accountability, and incident response [29]. Generative-AI adoption should be treated as a sociotechnical implementation process rather than installation of an isolated model [30].

Figure 2 presents the original conceptual synthesis for claim-level evidence binding, provenance, conflict, and visible uncertainty, showing how the article's principal components, evidence relationships, uncertainties, and decision boundaries are connected.

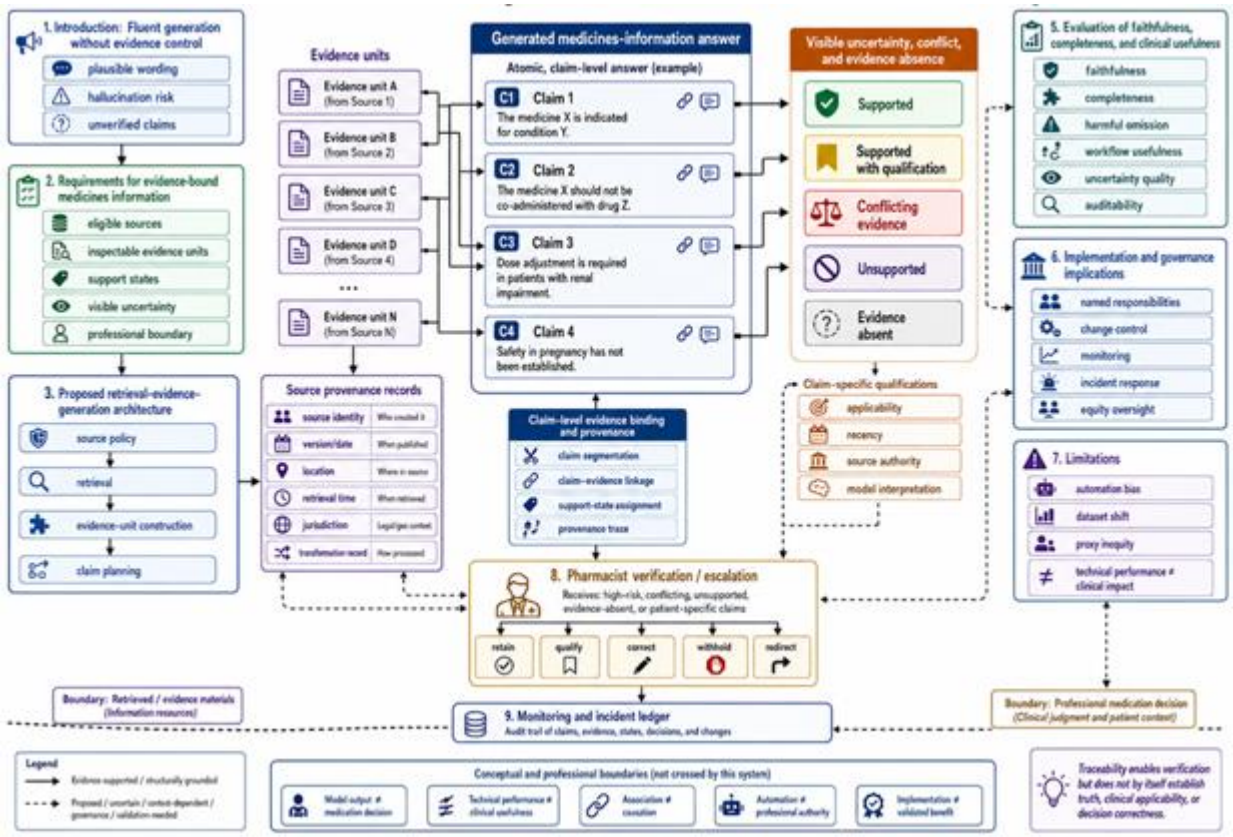

Figure 2. Claim-Level Evidence Binding, Provenance, Conflict, and Visible Uncertainty

Implementation conditions vary with the technology, users, organisation, and wider care system; nonadoption, abandonment, scale-up, and sustainability cannot be inferred from technical performance alone [31]. Relevant determinants also arise across the innovation, inner and outer settings, individuals, and implementation process [32]. Lifecycle governance must therefore connect responsible development, generative-AI implementation, and context-sensitive continuing evaluation [7, 30, 32].

Table 3 organizes the evidence, constructs, relationships, uncertainties, and boundary conditions required for governance responsibilities, implementation conditions, and monitoring requirements for the article Evidence-Bound Generation for Medicines Information with Verifiable Claims and Visible Uncertainty.

Table 3. Governance responsibilities, implementation conditions, and monitoring requirements for the article Evidence-Bound Generation for Medicines Information with Verifiable Claims and Visible Uncertainty.

Governance domain	Responsible actor	Required authority	Documentation requirement	Monitoring indicator	Escalation trigger	Corrective action	Accountability boundary
Intended use	Clinical and product owners	Approve defined task and users	Use case, exclusions, decision boundary	Use outside declared scope	Material scope expansion	Suspend or revalidate use	Approval does not establish benefit [29]
Evidence policy	Evidence-governance lead	Approve eligible source classes	Source, jurisdiction, recency, and exclusion rules	Source change or unavailable evidence	Loss of authoritative source	Replace source and repeat verification	Source approval does not prove truth
Pharmacy content	Named pharmacist owner	Adjudicate medication-content standards	Review criteria and escalation rules	Harmful omission or recurrent correction	High-risk unsupported claim	Correct, withhold, or refer	Pharmacist review may still fail
Technical system	Model and retrieval owners	Control configurations and releases	Model, prompt, retrieval, and version record	Behavioural or retrieval change	Unexplained performance deterioration	Roll back and investigate	Technical ownership is not clinical authority

Human oversight	Pharmacy service lead	Design feasible verification	Reviewer role, workload, and override process	Review burden or reliance error	Verification becomes ineffective	Redesign workflow or restrict use	Human presence is not effective oversight
Equity	Governance and equity leads	Require subgroup assessment	Language, subgroup, and access evaluation	Differential error or abstention	Material disparity	Restrict, redesign, and reassess	Aggregate performance may conceal harm
Implementation	Multidisciplinary implementation team	Adapt workflow and training	Context, resources, training, and change plan	Abandonment or workarounds	Unsafe or unsustainable use	Modify or withdraw implementation [31, 32]	Adoption is not validated benefit
Lifecycle monitoring	Named lifecycle owner	Initiate correction, suspension, or withdrawal	Incidents, updates, audits, and decisions	Drift, source conflict, repeated error	Loss of defined performance or traceability	Revalidate, suspend, or withdraw [30]	Monitoring cannot detect every failure

These responsibilities are proposed functional allocations. Their legal and professional distribution must be adapted to jurisdiction, pharmacy scope, information systems, and organisational capacity.

Limitations

Human review is not an infallible safeguard because automation bias and verification complexity can reduce effective oversight [33]. The architecture may also lose validity when populations, evidence, documentation, practice patterns, or source distributions change [34]. Source-selection rules and apparently neutral proxies may reproduce structural inequities or unequal access to authoritative evidence [35]. Finally, technical performance remains distinct from workflow fit, clinical utility, generalisability, regulation, and patient impact [36].

The framework does not determine optimal thresholds, staffing, interfaces, source hierarchies, or jurisdiction-specific responsibilities. Its proposed support states and escalation rules require empirical testing across medicines-information tasks, languages, populations, and pharmacy settings.

Conclusion

Evidence-bound generation reframes medicines-information generation as a controlled relationship among retrieval, eligible evidence, verifiable claims, visible uncertainty, professional authority, and lifecycle governance. The proposed architecture places generation downstream of source qualification and evidence-unit construction, then requires claim-level binding, provenance, conflict preservation, evidence-absence handling, and risk-sensitive review.

Its contribution is conceptual rather than empirical. It provides a structured basis for testing whether generated medicines information is faithful, sufficiently complete, clinically usable, equitable, auditable, and governable. It also makes explicit that citations do not establish truth, confidence does not establish calibration, retrieval does not establish applicability, and human participation does not automatically constitute effective oversight.

Future validation should examine technical performance, pharmacist adjudication, user comprehension, verification workload, subgroup effects, workflow consequences, and continuing validity under evidence and system change. Until such validation is completed, the architecture should be interpreted as a proposed scientific and governance framework. It does not establish improved medication safety, autonomous medication decision making, regulatory acceptance, universal applicability, or deployment readiness.

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

References

1. 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
2. Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, et al. Large language models encode clinical knowledge. Nature. 2023;620(7972):172-80. doi:10.1038/s41586-023-06291-2
3. Ayers JW, Poliak A, Dredze M, Leas EC, Zhu Z, Kelley JB, et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. JAMA Intern Med. 2023;183(6):589-96. doi:10.1001/jamainternmed.2023.1838
4. 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
5. Huang J, Yang DM, Rong R, Nezafati K, Treager C, Chi Z, et al. A critical assessment of using ChatGPT for extracting structured data from clinical notes. NPJ Digit Med. 2024;7(1):106. doi:10.1038/s41746-024-01079-8
6. Sutton RT, Pincock D, Baumgart DC, Sadowski DC, Fedorak RN, Kroeker KI. An overview of clinical decision support systems: Benefits, risks, and strategies for success. NPJ Digit Med. 2020;3:17. doi:10.1038/s41746-020-0221-y
7. Wiens J, Saria S, Sendak M, Ghassemi M, Liu VX, Doshi-Velez F, et al. Do no harm: A roadmap for responsible machine learning for health care. Nat Med. 2019;25(9):1337-40. doi:10.1038/s41591-019-0548-6
8. Amann J, Blasimme A, Vayena E, Frey D, Madai VI. Explainability for artificial intelligence in healthcare: A multidisciplinary perspective. BMC Med Inform Decis Mak. 2020;20(1):310. doi:10.1186/s12911-020-01332-6
9. Vasey B, Nagendran M, Campbell B, Clifton DA, Collins GS, Denaxas S, et al. Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. Nat Med. 2022;28(5):924-33. doi:10.1038/s41591-022-01772-9

10. Grossman S, Zerilli T, Nathan JP. Appropriateness of ChatGPT as a resource for medication-related questions. Br J Clin Pharmacol. 2024;90(10):2691-5. doi:10.1111/bcp.16212
11. Bhayana R, Fawzy A, Deng Y, Bleakney RR, Krishna S. Retrieval-augmented generation for large language models in radiology: Another leap forward in board examination performance. Radiology. 2024;313(1). doi:10.1148/radiol.241489
12. Zhang G, Xu Z, Jin Q, Chen F, Fang Y, Liu Y, et al. Leveraging long context in retrieval augmented language models for medical question answering. NPJ Digit Med. 2025;8(1):239. doi:10.1038/s41746-025-01651-w
13. Das S, Ge Y, Guo Y, Rajwal S, Hairston J, Powell J, et al. Two-layer retrieval-augmented generation framework for low-resource medical question answering using Reddit data: Proof-of-concept study. J Med Internet Res. 2025;27. doi:10.2196/66220
14. Jin Q, Kim W, Chen Q, Comeau DC, Yeganova L, Wilbur WJ, et al. MedCPT: Contrastive pre-trained transformers with large-scale PubMed search logs for zero-shot biomedical information retrieval. Bioinformatics. 2023;39(11). doi:10.1093/bioinformatics/btad651
15. Tayebi Arasteh S, Lotfinia M, Bressem K, Siepmann R, Adams LC, Ferber D, et al. RadioRAG: Online retrieval-augmented generation for radiology question answering. Radiol Artif Intell. 2025;7(4). doi:10.1148/ryai.240476
16. 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
17. Gorenshtein A, Shihada K, Sorka M, Aran D, Shelly S. LITERAS: Biomedical literature review and citation retrieval agents. Comput Biol Med. 2025;192(Pt B):110363. doi:10.1016/j.compbiomed.2025.110363
18. Marshall IJ, Nye B, Kuiper J, Noel-Storr A, Marshall R, Maclean R, et al. Trialstreamer: A living, automatically updated database of clinical trial reports. J Am Med Inform Assoc. 2020;27(12):1903-12. doi:10.1093/jamia/ocaa163
19. Munafò MR, Nosek BA, Bishop DVM, Button KS, Chambers CD, Percie du Sert N, et al. A manifesto for reproducible science. Nat Hum Behav. 2017;1(1):0021. doi:10.1038/s41562-016-0021
20. Ghassemi M, Oakden-Rayner L, Beam AL. The false hope of current approaches to explainable artificial intelligence in health care. Lancet Digit Health. 2021;3(11). doi:10.1016/S2589-7500(21)00208-9
21. Savage T, Wang J, Gallo R, Boukil A, Patel V, Safavi-Naini SAA, et al. Large language model uncertainty proxies: Discrimination and calibration for medical diagnosis and treatment. J Am Med Inform Assoc. 2025;32(1):139-49. doi:10.1093/jamia/ocae254
22. Kompa B, Snoek J, Beam AL. Second opinion needed: Communicating uncertainty in medical machine learning. NPJ Digit Med. 2021;4(1):4. doi:10.1038/s41746-020-00367-3
23. Begoli E, Bhattacharya T, Kusnezov D. The need for uncertainty quantification in machine-assisted medical decision making. Nat Mach Intell. 2019;1(1):20-3. doi:10.1038/s42256-018-0004-1
24. Van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW; Topic Group "Evaluating diagnostic tests and prediction models" of the STRATOS initiative. Calibration: The Achilles heel of predictive analytics. BMC Med. 2019;17(1):230. doi:10.1186/s12916-019-1466-7
25. Bedi S, Liu Y, Orr-Ewing L, Dash D, Koyejo S, Callahan A, et al. Testing and evaluation of health care applications of large language models: A systematic review. JAMA. 2025;333(4):319-28. doi:10.1001/jama.2024.21700
26. Norgeot B, Quer G, Beaulieu-Jones BK, Torkamani A, Dias R, Gianfrancesco M, et al. Minimum information about clinical artificial intelligence modeling: The MI-CLAIM checklist. Nat Med. 2020;26(9):1320-4. doi:10.1038/s41591-020-1041-y
27. Miao BY, Chen IY, Williams CYK, Davidson J, Garcia-Agundez A, Sun S, et al. The MI-CLAIM-GEN checklist for generative artificial intelligence in health. Nat Med. 2025;31(5):1394-8. doi:10.1038/s41591-024-03470-0
28. Nagendran M, Chen Y, Lovejoy CA, Gordon AC, Komorowski M, Harvey H, et al. Artificial intelligence versus clinicians: Systematic review of design, reporting standards, and claims of deep learning studies. BMJ. 2020;368. doi:10.1136/bmj.m689
29. Reddy S, Allan S, Coghlan S, Cooper P. A governance model for the application of AI in health care. J Am Med Inform Assoc. 2020;27(3):491-7. doi:10.1093/jamia/ocz192
30. Reddy S. Generative AI in healthcare: An implementation science informed translational path on application, integration and governance. Implement Sci. 2024;19(1):27. doi:10.1186/s13012-024-01357-9
31. Greenhalgh T, Wherton J, Papoutsi C, Lynch J, Hughes G, A'Court C, et al. Beyond adoption: A new framework for theorizing and evaluating nonadoption, abandonment, and challenges to the scale-up, spread, and sustainability of health and care technologies. J Med Internet Res. 2017;19(11). doi:10.2196/jmir.8775
32. Damschroder LJ, Reardon CM, Widerquist MAO, Lowery J. The updated Consolidated Framework for Implementation Research based on user feedback. Implement Sci. 2022;17(1):75. doi:10.1186/s13012-022-01245-0
33. Lyell D, Coiera E. Automation bias and verification complexity: A systematic review. J Am Med Inform Assoc. 2017;24(2):423-31. doi:10.1093/jamia/ocw105
34. Subbaswamy A, Saria S. From development to deployment: Dataset shift, causality, and shift-stable models in health AI. Biostatistics. 2020;21(2):345-52. doi:10.1093/biostatistics/kxz041
35. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. Science. 2019;366(6464):447-53. doi:10.1126/science.aax2342
36. Kelly CJ, Karthikesalingam A, Suleyman M, Corrado G, King D. Key challenges for delivering clinical impact with artificial intelligence. BMC Med. 2019;17(1):195. doi:10.1186/s12916-019-1426-2