Annals of Pharmacy Practice and Pharmacotherapy
ISSN: 3062-4436
2026, Volume 6, Issue 1, Page No: 51-60

Available online at: www.galaxypub.co/page/journals

Galaxy Publication

Generative AI in Pharmacotherapy: Why Evidence Traceability, Verification Burden, Uncertainty, and Human Override Define the Boundary Between Clinical Assistance and Medication-Safety Risk

Emma Wilson[1]*, Frederik De Jong[2], Lara Peters[2]

[1] Department of Generative AI in Clinical Pharmacy, Faculty of Pharmacy, University of Edinburgh, Edinburgh, United Kingdom.
[2] Department of AI Safety and Human Override, Faculty of Pharmaceutical Sciences, Utrecht University, Utrecht, Netherlands.

***E-mail** ✉ emma.wilson@ed.ac.uk

Received: 30 November 2025; **Revised:** 20 February 2026; **Accepted:** 26 February 2026

ABSTRACT

Generative artificial intelligence is increasingly being evaluated for medication information, clinical-pharmacy questions, treatment recommendations, deprescribing, documentation, and other pharmacotherapy tasks. Its apparent efficiency can obscure a medication-safety problem: fluent output may be difficult to audit, expensive to verify, insensitive to changing patient context, or persuasive even when wrong. This Current Opinion argues that the clinically relevant boundary is not whether a model can generate a plausible answer, but whether the surrounding use condition preserves evidence traceability, feasible verification, interpretable uncertainty, and meaningful human override in proportion to the consequence of error. Available literature supports useful task-level capabilities, retrieval-supported improvements, and emerging methods for uncertainty estimation, while also documenting fabricated references, omissions, variable calibration, context-sensitive failures, and adverse effects of incorrect AI advice on human decisions. We therefore propose an assistance–risk boundary architecture in which generative AI remains assistive only when its evidential basis can be inspected, checking effort is commensurate with workflow capacity, uncertainty and context limitations remain visible, and a qualified human can reject or escalate the output before medication action. This argument does not establish a validated risk threshold, autonomous-use criterion, or medication-safety benefit. Model version, retrieval corpus, clinical task, drug and patient characteristics, local guidance, workflow design, user expertise, and post-deployment change remain material boundaries. Prospective, task-specific evaluation is required before higher-consequence pharmacotherapy responsibilities can be justified.

Keywords: Generative artificial intelligence, Pharmacotherapy, Medication safety, Evidence traceability, Clinical verification, Uncertainty

How to Cite This Article: Wilson E, De Jong F, Peters L. Generative AI in Pharmacotherapy: Why Evidence Traceability, Verification Burden, Uncertainty, and Human Override Define the Boundary Between Clinical Assistance and Medication-Safety Risk. Ann Pharm Pract Pharmacother. 2026;6(1):51-60. https://doi.org/10.51847/xT3dQdT8kP

Introduction

Generative AI has moved rapidly from general conversational systems into medication-related applications, including drug-information support, clinical-pharmacy question answering, prescribing-oriented decision support, and deprescribing. Yet the evidence base remains uneven. A scoping review of generative AI and large language models for medication-related harm found heterogeneous applications and predominantly exploratory evaluation, with no prospective deployment study among the included evidence [1]. That observation should not be read as proof that generative AI is harmful or incapable of benefit. It instead limits how confidently task performance can be translated into claims about medication-safety effectiveness in real care, where a response becomes consequential only after interacting with a patient, regimen, clinician, workflow, and health-system context.

This distinction matters because medication safety is a sociotechnical property rather than an output-level property. Earlier healthcare-AI literature described plausible safety gains from detection, prediction, and decision support while also emphasizing implementation, workflow, and evaluation hazards [2]. Generative systems add a distinctive problem: they produce natural-language answers that may blend evidence, inference, omitted context, and unsupported content within a single fluent response. For pharmacotherapy, risk is therefore not confined to an obviously false fact. It can arise when a correct general statement is applied to the wrong patient, when an evidence source is outdated, when a relevant interaction or monitoring requirement is omitted, or when the cost of checking an answer exceeds the time apparently saved.

Strong model capability cannot therefore be treated as equivalent to safe assistance. Large language models can encode substantial clinical knowledge, yet realistic evaluations have identified limitations in guideline adherence, laboratory interpretation, instruction following, and robustness to information order; randomized clinician studies likewise show that access to an LLM does not automatically improve reasoning [3–5]. These findings do not imply uniform failure, nor do they establish that one model, specialty, or prompt pattern predicts another. They instead indicate that the clinically relevant unit of assessment is the model–task–user–workflow configuration, with performance interpreted against the consequence of the decision being supported.

This Current Opinion proposes that the boundary between useful assistance and medication-safety risk is defined by four interacting conditions: evidence traceability, verification burden, uncertainty, and meaningful human override. The proposal deliberately separates generation from action. A model may be useful for drafting, retrieval, option generation, or preparing questions for review while remaining unsuitable for an unreviewed medication decision. These functions can also shift over time as models, evidence sources, interfaces, and local workflows change. The central question is therefore not simply "Can the model answer?" but "Can the answer be safely inspected, challenged, contextualized, updated, and stopped before it changes therapy?"

Generative AI use cases in pharmacotherapy

Medication-related generative AI spans tasks with materially different purposes and clinical stakes. Multi-specialty evaluations have tested LLMs as medication-safety decision-support tools, illustrating potential assistance while showing that performance depends on model configuration and workflow [6]. Comparative clinical-pharmacy work similarly suggests that ChatGPT can handle some information tasks competitively but remains variable relative to pharmacists, particularly where contextual judgment is required [7]. These uses should therefore be classified by what the output is permitted to do—retrieve evidence, draft an explanation, compare options, recommend an intervention, or directly influence action—rather than by the model brand producing it. The same linguistic interface can conceal very different medication-safety exposures.

Deprescribing makes this distinction concrete. In benzodiazepine deprescribing scenarios, AI-assisted recommendations aligned with healthcare-professional judgments on some criteria but remained inconsistent or ambiguous on others [8]. Drug-information chatbots likewise produced responses of uneven accuracy, detail, and potential harm when compared with licensed pharmacists [9]. A plausible answer may consequently be useful as a candidate explanation, prompt for further review, or checklist item while still being inappropriate as an autonomous treatment instruction. The unresolved work is often precisely the clinical work: identifying indication, treatment history, interacting medicines, renal or hepatic function, competing risks, adherence capability, patient priorities, and what would happen if the recommendation were wrong.

This suggests a functional distinction among information support, synthesis support, and decision modification. Information support can still cause harm, but errors can sometimes be intercepted before they affect therapy. Decision modification places greater weight on the completeness of patient-specific context and on who has authority to accept or reject the output. Generative AI can therefore occupy several positions within the same pharmacotherapy workflow without each position warranting the same evidential standard.

Real-world pharmacotherapy counselling further challenges the assumption that conversational fluency indicates medication reliability. In a blinded comparison using authentic pharmacotherapeutic questions, physicians outperformed ChatGPT-3.5 on information quality and factual correctness [10]. The finding is model- and period-specific, not a verdict on later systems. Its more durable implication is methodological: pharmacotherapy assistance should be evaluated against qualified human work on clinically representative questions, with attention not only to answer generation but also to contextual completeness, error detectability, and the downstream work required to establish whether the answer is sufficiently trustworthy for its intended use.

Evidence traceability as a minimum condition for clinical use

Evidence grounding can improve clinical LLM performance, but grounding and traceability are not identical to safety. Retrieval-augmented or document-supported systems have improved performance in guideline interpretation, evidence-based neurology, and guideline-bound medical-fitness assessment, while the magnitude and reliability of gains vary with task, corpus, model, and residual error [11–13]. These studies support authoritative retrieval as a safety-relevant design property. They do not establish that retrieval-augmented generation is safe for autonomous pharmacotherapy. A retrieved source may be irrelevant, incomplete, outdated, incorrectly ranked, or correctly quoted but incorrectly applied. Retrieval can therefore reduce one uncertainty while introducing another: whether the retrieved evidence is the right evidence for the current clinical state.

For clinical use, traceability should mean more than displaying a citation. The clinician should be able to identify which evidence or guidance supports a material medication claim, determine whether the source is authentic and current, and inspect whether the generated interpretation matches it. This requirement is nontrivial because generative systems can fabricate, corrupt, or misattribute references [14]. A polished bibliography can consequently increase rather than reduce verification burden if each citation must first be proven to exist. Traceability is strongest when provenance is linked to inspectable source material and when the clinical claim can be reconstructed from that material; it is weakest when the user must reverse-engineer an evidential path from generated prose.

The same principle applies to evidence synthesis. Generative AI may accelerate searching, summarization, and synthesis, but trustworthy clinical evidence synthesis requires controls for provenance, validation, and unsupported inference [15]. Pharmacotherapy raises the consequence because evidence claims may be translated directly into drug selection, dose changes, discontinuation, substitution, or monitoring. We therefore propose evidence traceability as a minimum condition for action-oriented clinical use: not proof that an output is correct, and not a guarantee that cited evidence is applicable, but a prerequisite for efficient falsification. If the supporting basis cannot be inspected, a clinician cannot reliably distinguish a well-grounded recommendation from a plausible but unauditable one.

Traceability also has a temporal dimension. A verifiable source may still be clinically unsuitable if it no longer reflects current guidance, product information, resistance patterns, formulary restrictions, or the patient's present regimen. The clinically relevant provenance chain must therefore connect a generated claim not only to *a* source, but to an appropriate and sufficiently current source for the intended decision. That standard becomes more demanding as an output moves closer to medication action.

Verification burden and the hidden cost of apparent efficiency

The most visible efficiency metric for generative AI is often response time, yet the clinically relevant workload extends beyond generation. In a randomized mock medication-verification task, uncertainty-aware AI presentation changed pharmacists' reaction time and decision-making behavior [16]. The experiment was simulated and does not establish patient benefit, but it demonstrates that output presentation alters the work of verification itself. We define verification burden as the cognitive, temporal, and workflow effort required to determine whether an AI output is sufficiently correct, complete, current, and context-appropriate for its intended medication-related use. That burden includes locating evidence, reconstructing missing context, resolving discrepancies, and deciding when uncertainty warrants escalation.

Apparent time saving can coexist with substantial correction work. Physician review of AI-generated hospital-course summaries found usable outputs alongside omissions and inaccuracies requiring review, with heterogeneous perceptions of time saved [17]. Documentation is not equivalent to prescribing, and summary errors should not be relabelled as medication harms. The transferable point is narrower: automation can shift effort from production to inspection. For pharmacotherapy, that shift may be unfavorable when verification requires reconstructing the medication history, checking several sources, reconciling interacting recommendations, or detecting an omission that is harder to notice than an explicit error. Efficiency should therefore be judged on the full human–AI task, not the model's generation latency alone.

Human oversight is not costless protection. Incorrect AI advice can degrade clinician decisions, while automation bias and verification complexity can make nominal oversight ineffective when checking is difficult or the output appears authoritative [18–20]. Expertise, interface design, training, and workflow may moderate these effects, so they should not be generalized as fixed medication-specific risk estimates. Nevertheless, they support an important boundary: assistance becomes less defensible when the system produces more consequential material than the

user can realistically verify before action. In that setting, "human in the loop" may describe accountability without demonstrating practical control.

Verification burden also creates opportunity cost. Time spent checking a low-value generated answer competes with medication reconciliation, patient communication, monitoring, and other safety work. Conversely, a well-grounded output that directs attention efficiently to the exact evidence requiring review may lower total effort even if it does not eliminate checking. The appropriate comparison is therefore not AI generation versus unaided drafting, but total work to reach a defensible clinical decision.

Table 1 decomposes this hidden work across representative pharmacotherapy tasks and identifies where apparent automation gain is transferred back to the clinician as verification responsibility.

Table 1. Pharmacotherapy tasks in which apparent generative-AI efficiency transfers clinical work to verification

AI-assisted pharmacotherapy task	Apparent automation gain	Verification work	Traceability demand	Missed-error consequence	Human checkpoint
Drug-information or counselling draft	Rapid synthesis and wording	Check accuracy, completeness, currency, and patient fit	Inspectable support for material claims	Misinformation or delayed escalation	Pharmacist/prescriber review before use
Clinical-pharmacy question answering	Rapid option generation	Reconcile indication, interactions, organ function, monitoring, and local guidance	Claim-to-source path	Inappropriate recommendation or monitoring	Qualified clinician validates actionable claims
Deprescribing or treatment modification	Candidate stop, taper, or substitution plan	Confirm indication, withdrawal risk, alternatives, sequence, and follow-up	Current guidance plus patient-specific basis	Withdrawal, relapse, undertreatment, or avoidable exposure	Prescriber/pharmacist authorizes change
Clinical documentation synthesis	Reduced drafting workload	Detect omissions, distortion, and medication-context loss	Provenance to source record	Error propagation into subsequent care	Clinician reconciles against record before sign-off

Note: Task decomposition, verification demands, and checkpoints are proposed synthesis derived from the evidence discussed in the surrounding text. They are not validated risk tiers, performance thresholds, or autonomous-use criteria.

Uncertainty, hallucination, staleness, and context loss

Uncertainty is not a single failure state. Methods based on semantic entropy can identify some hallucination-prone generations by measuring instability in meaning across outputs [21]. That is useful as a screening mechanism, but it is not a truth oracle. A model can be consistently wrong, and a low-uncertainty answer may still rest on an inappropriate source, an omitted variable, or an invalid clinical assumption. In pharmacotherapy, uncertainty signaling should therefore complement evidence verification rather than substitute for it. It is most useful when it changes what the user does next—verify, abstain, seek a second opinion, or withhold medication action.

Nor is there a universal confidence threshold that can be carried from one clinical task to another. Medical LLM uncertainty proxies differ in discrimination and calibration across diagnosis and treatment settings [22]. Their usefulness depends on the model, prompt, task distribution, and decision context. A confidence display is clinically meaningful only if users understand what it estimates and whether it has been calibrated for the action being considered. Otherwise, numerical confidence can become another persuasive surface feature rather than an effective safeguard. The relevant question is not merely whether uncertainty is displayed, but whether it is interpretable enough to support safer escalation or abstention.

"Hallucination rate" is likewise too coarse for medication safety. Clinical summarization research shows the importance of separating fabricated content from omission, factual error, and clinically weighted severity [23]. The same distinction is necessary in pharmacotherapy even though error rates from summarization cannot be transferred directly. An invented contraindication, an omitted interaction, an outdated dose recommendation, and a generally correct statement detached from renal function represent different failure modes. They differ in visibility, ease of verification, clinical consequence, and the likelihood that a fluent answer will conceal them.

Guideline-based evaluations further show that LLM outputs may omit relevant guidance or introduce unsupported content in context-sensitive ways [24]. This makes staleness and context loss analytically separate from hallucination. A statement may be factually defensible in general yet inappropriate for the current version of guidance, local formulary, pregnancy status, organ function, treatment sequence, interacting regimen, or patient preference. Conversely, an answer may accurately reproduce an older source while being temporally wrong for current care. The practical safety question is therefore not merely whether the model "knows" an answer, but whether its evidence, uncertainty, temporal validity, and patient context remain visible enough for a qualified human to challenge before medication action.

Human override and responsibility for medication decisions

Human override is meaningful only if a clinician can recognize, challenge, reject, correct, or escalate an AI-supported recommendation before it becomes a medication decision. In randomized evidence, AI assistance modified physicians' clinical decisions and could transmit bias into those decisions [25]. This does not mean that all AI influence is harmful. It means that clinician presence alone cannot be treated as proof that system influence has been neutralized. Override is therefore an active capability, not a label applied after the model has shaped the decision.

Automation bias provides a plausible pathway by which apparently assistive AI may become unsafe when users defer to recommendations that are difficult to independently check [26]. In pharmacotherapy, responsibility should remain coupled to practical authority: record access, time, relevant expertise, and permission to stop or escalate. Accountability without those capabilities risks becoming ceremonial.

Uncertainty communication can support abstention or a second opinion when model confidence is limited [27]. Poorly calibrated or burdensome uncertainty displays can add work without clarifying action. Meaningful override therefore requires uncertainty information that changes a decision pathway rather than merely decorating an output.

Human–AI complementarity is also conditional. Human-computer collaboration has improved performance in some circumstances while transmitting error in others [28]. Although that evidence predates current generative systems and is outside pharmacotherapy, it supports a durable design principle: safe collaboration depends on case, model, user, and interface. A "human in the loop" is insufficient unless the human can still exercise independent judgment.

The proposed assistance–risk boundary architecture

Healthcare AI governance frameworks vary substantially and often under-specify operational oversight, evaluation, or evidence that governance mechanisms work in practice [29]. Pharmacotherapy therefore needs governance linked to the actual task. High-level principles can be translated into concrete lifecycle evaluation and deployment checkpoints [30], but such checkpoints cannot substitute for validation of the medication decision being supported.

We propose that generative AI remains on the assistance side of the boundary when four conditions remain jointly defensible: material claims are traceable to inspectable evidence; verification work is feasible within the clinical workflow; uncertainty, staleness, and missing context remain visible enough to alter reliance; and a qualified human can prevent or reverse progression to medication action. These are interacting constraints, not a numerical score; failure in one matters more as consequence rises or reversibility falls.

The boundary must also be distribution-sensitive. Bias and health-equity concerns can emerge at multiple lifecycle stages and require explicit assessment rather than reliance on average performance [31]. Thus, an apparently acceptable task can cross toward risk if evidence or performance is unreliable for a subgroup, language, institution, or access context.

Finally, the boundary is temporal. Responsible clinical AI requires evaluation, monitoring, and re-evaluation across the lifecycle [32]. A previously acceptable model–task configuration may become less defensible after model updates, retrieval-corpus drift, changed guidance, workflow redesign, or recurrent override failures. **Figure 1** visualizes this proposed, non-validated boundary logic.

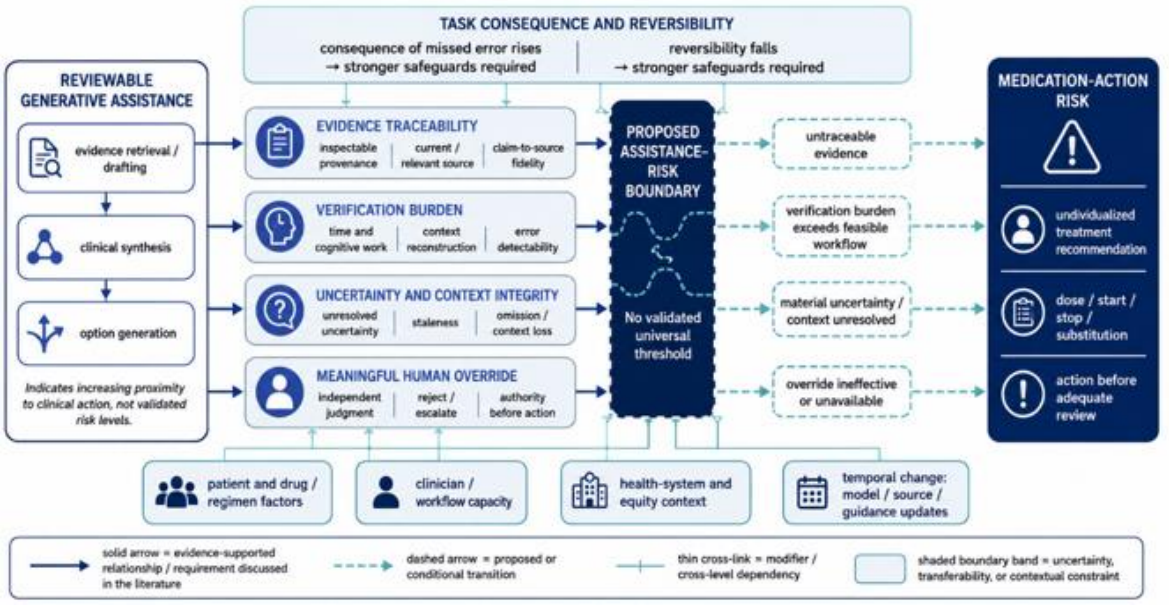

Figure 1. Traceability, verification burden, uncertainty, and override determine when generative output becomes medication-action risk

Task stratification by consequence and reversibility

A useful task taxonomy should begin with what happens if the output is wrong, not with how impressive the output appears. A pragmatic cluster-randomized trial of generative-AI clinical decision support did not significantly improve its primary treatment-failure endpoint, despite the plausibility of process support [33]. The finding is setting- and intervention-specific, but it demonstrates why process utility cannot be assumed to produce a patient-level outcome. Validation must match the consequence being claimed.

We therefore propose consequence and reversibility as two organizing dimensions. Drafting a counselling explanation that is reviewed before use differs from generating a dose change that can immediately alter exposure. Reversibility is contextual: an error corrected before administration differs from an action whose effects persist.

Average performance is insufficient when harms are unevenly distributed. LLM evaluation can surface subgroup-specific health-equity harms that aggregate metrics obscure [34]. More broadly, algorithmic inequity can arise from the proxy or target chosen even when the model performs consistently against that target [35]. This is a transferable caution, not evidence of the same mechanism in every pharmacotherapy LLM.

Clinical reasoning performance also varies across models and tasks [36]. Consequently, permissibility should attach to a defined model–task–workflow use, not to a model as a whole. **Table 2** translates the proposed consequence–reversibility logic into task-level assistance boundaries without implying validated risk weights.

Table 2. Permissible generative-AI assistance should tighten as pharmacotherapy consequences rise and reversibility falls

Task class	Decision consequence	Reversibility	Uncertainty tolerance (proposed)	Evidence requirement	Override authority	Autonomous boundary
Drafting or retrieval support	Indirect; action requires later review	Usually high before use	Some uncertainty may remain if explicit	Inspectable source for material claims	Reviewer may edit or discard	No unreviewed patient-facing instruction

Evidence synthesis or counselling support	May shape understanding or options	Moderate if intercepted before action	Unresolved uncertainty prompts verification	Current, relevant, claim-linked evidence	Pharmacist/prescriber confirms actionable content	No autonomous individualized recommendation
Treatment-option recommendation	Can alter drug choice, monitoring, or sequence	Variable	Low tolerance for unresolved material uncertainty	Patient-specific evidence and current guidance	Qualified clinician accepts, rejects, or escalates	No autonomous treatment selection
Dose, start, stop, or substitution action	Direct medication exposure or withdrawal consequence	May be difficult or time-sensitive	Material uncertainty incompatible with unreviewed action	Full clinical context plus independently verifiable basis	Authorized clinician controls execution	No autonomous medication change within the proposed architecture

Note: This permissibility register is proposed synthesis, not a validated risk scale, regulatory classification, or universal prohibition. Actual safeguards depend on the drug, patient, setting, timing, and available clinical controls.

Governance, monitoring, and post-deployment learning

Governance should treat a generative-AI pharmacotherapy use as a changing configuration rather than a one-time approved product. Healthcare governance reviews support accountable oversight and lifecycle monitoring but also show that operational completeness and effectiveness evaluation remain limited [29]. The governance object is the specific model version, retrieval source, interface, task, user group, and medication workflow.

Operational principles become useful when converted into controls that can be inspected: named ownership, version and provenance documentation, predeployment task testing, escalation routes, and defined authority to restrict use [30]. We additionally propose monitoring source freshness, verification failures, overrides, near misses, and disagreement with current local guidance. These categories are proposed safety-learning signals, not validated predictors of harm.

Predeployment testing cannot cover later change. Responsible-healthcare ML guidance supports ongoing monitoring and re-evaluation as data, performance, and workflows evolve [32]. A pharmacotherapy system should therefore be capable of being narrowed, paused, or returned to a lower-consequence role when its evidential basis or behavior becomes uncertain. Reauthorization should follow investigation rather than occur automatically after an update. **Figure 2** depicts this proposed closed-loop governance response.

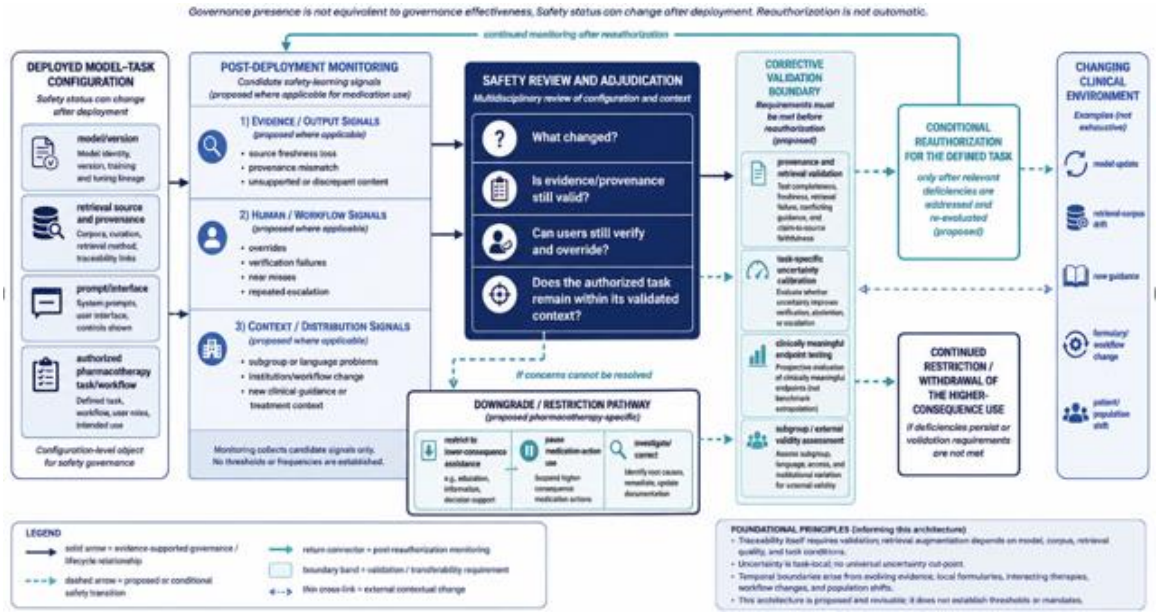

Figure 2. Post-deployment change should trigger provenance review, task restriction, and reauthorization before renewed medication use

Boundary conditions and validation requirements

Traceability itself requires validation. Retrieval augmentation can improve performance, but its value depends on model, corpus, retrieval quality, and task conditions [13]. A pharmacotherapy system should be tested for source

completeness, freshness, retrieval failure, conflicting guidance, and faithful claim-to-source linkage. A provenance display is not sufficient if the wrong evidence is retrieved reliably.

Uncertainty is similarly local. Medical LLM uncertainty proxies vary in discrimination and calibration across tasks [22]. Medication workflows need task-specific evaluation of whether an uncertainty signal meaningfully improves verification, abstention, or escalation without creating false reassurance. No universal uncertainty cut-point follows from the current evidence.

Prospective validation should also measure the endpoint the intended use claims to improve. Pragmatic trial evidence illustrates that plausible decision support does not guarantee improvement in a patient-level primary outcome [33]. Higher-consequence roles therefore require representative workflows, appropriate comparators, and clinically meaningful outcomes rather than benchmark extrapolation.

External validity must include distributional performance. Equity-focused LLM evaluation shows why subgroup harms can remain hidden behind averages [34]. Validation should therefore examine relevant population, language, access, and institutional variation while recognizing that subgroup definitions and fairness measures are context-dependent. Model updates, evolving evidence, local formularies, new interacting therapies, and workflow changes create additional temporal boundaries. The proposed architecture should be treated as testable and revisable: prospective studies must be able to refine, restrict, or reject its task classifications rather than merely confirm them.

Conclusion

Generative AI creates a pharmacotherapy safety problem precisely because useful and unsafe outputs can share the same fluent interface. The clinically important distinction is therefore not generative capability alone but whether a particular use remains inspectable and controllable before it changes medication care. This Current Opinion proposes that evidence traceability, feasible verification, visible uncertainty and context limits, and meaningful human override jointly define that boundary, with consequence and reversibility determining how demanding those conditions should become.

The proposal does not claim clinical effectiveness or implementation readiness. Evidence remains heterogeneous across models, tasks, patient populations, specialties, interfaces, and health systems. Retrieval can improve grounding without guaranteeing applicability; uncertainty measures can assist detection without establishing truth; clinician involvement can moderate risk without guaranteeing independent judgment; and governance structures can organize accountability without proving improved outcomes.

The practical implication is to authorize generative AI at the level of a specific task rather than a model-wide reputation. Low-consequence, readily reversible assistance may tolerate more residual uncertainty when evidence and review remain available. Medication changes with substantial or difficult-to-reverse consequences require stronger provenance, lower unresolved uncertainty, realistic verification capacity, and genuine authority to stop. These proposed boundaries should be prospectively challenged and revised as evidence, models, workflows, or patient contexts change.

Acknowledgments: None

Conflict of Interest: None

Financial Support: None

Ethics Statement: None

References

1. 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

2. Bates DW, Levine D, Syrowatka A, Kuznetsova M, Craig KJT, Rui A, et al. The potential of artificial intelligence to improve patient safety: A scoping review. NPJ Digit Med. 2021;4(1):54. doi:10.1038/s41746-021-00423-6
3. Hager P, Jungmann F, Holland R, Bhagat K, Hubrecht I, Knauer M, et al. Evaluation and mitigation of the limitations of large language models in clinical decision-making. Nat Med. 2024;30(9):2613-22. doi:10.1038/s41591-024-03097-1
4. Goh E, Gallo R, Hom J, Strong E, Weng Y, Kerman H, et al. Large language model influence on diagnostic reasoning: A randomized clinical trial. JAMA Netw Open. 2024;7(10):e2440969. doi:10.1001/jamanetworkopen.2024.40969
5. 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
6. Ong JCL, Jin L, Elangovan K, Lim GYS, Lim DYZ, Sng GGR, et al. Large language model as clinical decision support system augments medication safety in 16 clinical specialties. Cell Rep Med. 2025;6(10):102323. doi:10.1016/j.xcrm.2025.102323
7. 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
8. Bužančić I, Belec D, Držaić M, Kummer I, Brkić J, Fialová D, et al. Clinical decision-making in benzodiazepine deprescribing by healthcare providers vs. AI-assisted approach. Br J Clin Pharmacol. 2024;90(3):662-74. doi:10.1111/bcp.15963
9. 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
10. 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
11. Kresevic S, Giuffrè M, Ajčević M, Accardo A, Crocè LS, Shung DL. Optimization of hepatological clinical guidelines interpretation by large language models: A retrieval augmented generation-based framework. NPJ Digit Med. 2024;7(1):102. doi:10.1038/s41746-024-01091-y
12. Masanneck L, Meuth SG, Pawlitzki M. Evaluating base and retrieval augmented LLMs with document or online support for evidence based neurology. NPJ Digit Med. 2025;8(1):137. doi:10.1038/s41746-025-01536-y
13. Ke YH, Jin L, Elangovan K, Abdullah HR, Liu N, Sia ATH, et al. Retrieval augmented generation for 10 large language models and its generalizability in assessing medical fitness. NPJ Digit Med. 2025;8(1):187. doi:10.1038/s41746-025-01519-z
14. 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:e53164. doi:10.2196/53164
15. Zhang G, Jin Q, McInerney DJ, Chen Y, Wang F, Cole CL, et al. Leveraging generative AI for clinical evidence synthesis needs to ensure trustworthiness. J Biomed Inform. 2024;153:104640. doi:10.1016/j.jbi.2024.104640
16. Lester C, Rowell B, Zheng Y, Co Z, Marshall V, Kim JY, et al. Effect of uncertainty-aware AI models on pharmacists' reaction time and decision-making in a web-based mock medication verification task: Randomized controlled trial. JMIR Med Inform. 2025;13:e64902. doi:10.2196/64902
17. Grolleau F, Liang AS, Keyes T, Ma SP, Lew T, Huynh TR, et al. Physician-reported safety outcomes of AI-generated hospital course summaries. JAMA Netw Open. 2026;9(5):e2616556. doi:10.1001/jamanetworkopen.2026.16556
18. Jussupow E, Spohrer K, Heinzl A, Gawlitza J. Augmenting medical diagnosis decisions? An investigation into physicians' decision-making process with artificial intelligence. Inf Syst Res. 2021;32(3):713-35. doi:10.1287/isre.2020.0980
19. Jabbour S, Fouhey D, Shepard S, Valley TS, Kazerooni EA, Banovic N, et al. Measuring the impact of AI in the diagnosis of hospitalized patients: A randomized clinical vignette survey study. JAMA. 2023;330(23):2275-84. doi:10.1001/jama.2023.22295

20. 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
21. Farquhar S, Kossen J, Kuhn L, Gal Y. Detecting hallucinations in large language models using semantic entropy. Nature. 2024;630(8017):625-30. doi:10.1038/s41586-024-07421-0
22. 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
23. 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
24. van Kessel R, Anderson M, McMillan B, Matthews MR, Rust P, Pearcy P, et al. Omission and hallucination prevalence of clinical guidelines in diagnostic large language model outputs. BMJ Health Care Inform. 2026;33(1):e101959. doi:10.1136/bmjhci-2025-101959
25. Goh E, Bunning B, Khoong EC, Gallo RJ, Milstein A, Centola D, et al. Physician clinical decision modification and bias assessment in a randomized controlled trial of AI assistance. Commun Med (Lond). 2025;5(1):59. doi:10.1038/s43856-025-00781-2
26. Khera R, Simon MA, Ross JS. Automation bias and assistive AI: Risk of harm from AI-driven clinical decision support. JAMA. 2023;330(23):2255-7. doi:10.1001/jama.2023.22557
27. 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
28. Tschandl P, Rinner C, Apalla Z, Argenziano G, Codella N, Halpern A, et al. Human-computer collaboration for skin cancer recognition. Nat Med. 2020;26(8):1229-34. doi:10.1038/s41591-020-0942-0
29. Wang A, Freeman S, Magrabi F. Governance for safe and responsible AI in healthcare organisations: A scoping review of frameworks. NPJ Digit Med. 2026;9(1):516. doi:10.1038/s41746-026-02679-2
30. Economou-Zavlanos NJ, Bessias S, Cary MP Jr, Bedoya AD, Goldstein BA, Jelovsek JE, et al. Translating ethical and quality principles for the effective, safe and fair development, deployment and use of artificial intelligence technologies in healthcare. J Am Med Inform Assoc. 2024;31(3):705-13. doi:10.1093/jamia/ocad221
31. Abràmoff MD, Tarver ME, Loyo-Berrios N, Trujillo S, Char D, Obermeyer Z, et al. Considerations for addressing bias in artificial intelligence for health equity. NPJ Digit Med. 2023;6(1):170. doi:10.1038/s41746-023-00913-9
32. 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
33. Agweyu A, Mwaniki P, Menon V, Korom R, Isaaka L, Wanyama C, et al. Generative AI-enabled clinical decision support system in primary care: A pragmatic, cluster-randomized trial. Nat Med. 2026. doi:10.1038/s41591-026-04503-6
34. Pfohl SR, Cole-Lewis H, Sayres R, Neal D, Asiedu M, Dieng A, et al. A toolbox for surfacing health equity harms and biases in large language models. Nat Med. 2024;30(12):3590-600. doi:10.1038/s41591-024-03258-2
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. Rao AS, Esmail KP, Lee RS, Jiang S, Arraiza Carlo B, Gill J, et al. Large language model performance and clinical reasoning tasks. JAMA Netw Open. 2026;9(4):e264003. doi:10.1001/jamanetworkopen.2026.4003