Original Article

Designing Large Language Models That Refuse Unsafe Medication Questions Well

Gabriel Costa[1]*, Lucas Ribeiro[1], Ricardo Alves[2], Mariana Lopes[3]

[1]Department of LLM Safety and Medication Questions, Faculty of Pharmacy, Federal University of Minas Gerais, Belo Horizonte, Brazil. [2]Department of AI Refusal Mechanisms in Pharmacy, Faculty of Pharmacy, University of Coimbra, Coimbra, Portugal. [3]Department of Safe LLM Design for Pharmacy, Faculty of Pharmacy, University of São Paulo, São Paulo, Brazil.

Abstract

Large language models can produce fluent medication information while remaining vulnerable to factual error, missing context, unsafe personalization, adversarial manipulation, and misleading confidence. Refusal is therefore not merely a conversational limitation; when appropriately designed, it may function as a medication-safety control. This article develops a proposed Pharmacy Refusal-Safety Framework for determining when a model should answer, qualify, redirect, or refuse a medication question. The framework combines a multi-axial taxonomy of unsafe and unanswerable questions, a layered risk-interpretation architecture, an action-selection boundary, a non-abandonment response contract, and lifecycle governance. Questions are characterized according to potential harm, urgency, personalization, context sufficiency, epistemic answerability, premise validity, misuse potential, vulnerability, and professional-authority requirements. The selected action is constrained by the consequences of error rather than linguistic confidence alone. A safe refusal should state the relevant boundary, avoid covert individualized advice, preserve appropriate general information, identify missing context, direct the user toward a feasible next action, and provide urgent safety-net guidance when danger is plausible. Evaluation requires realistic safety-critical cases, expert adjudication, subgroup analysis, adversarial testing, user-comprehension assessment, workflow simulation, and prospective monitoring. The framework is an original non-empirical synthesis rather than a validated clinical algorithm or deployment standard. Its categories and transitions may behave differently across models, medicines, languages, users, jurisdictions, and care settings. Empirical validation is therefore required before clinical or pharmacy implementation.

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

Introduction

Large language models can generate medically fluent text, but fluency does not establish factual reliability, clinical reasoning quality, or medication-safety fitness. Their potential benefits coexist with fabrication, incomplete reasoning, misplaced confidence, and use beyond evaluated conditions [1]. Refusal should consequently be understood as a possible medication-safety function: a controlled decision not to provide an apparently complete answer when acknowledging uncertainty, obtaining context, or transferring responsibility would be safer.

Medical question-answering performance also cannot be reduced to benchmark accuracy. Human assessment must consider potential harm, relevant omissions, bias, comprehension, and helpfulness [2]. These dimensions can conflict. A response may be generally correct but omit a contraindication, appear reassuring despite inadequate patient information, or be so cautious that it delays needed care. The relevant question is therefore not only whether a model can generate an answer, but whether it should provide that answer in the requested form and context.

Medication-related evidence remains heterogeneous and task-specific, with limited standardization and insufficient evidence for autonomous safety or prospective workflow benefit [3]. Direct evaluation of medication-related questions has likewise shown that plausible responses may be incomplete or inappropriate [4]. These findings support treating refusal as a pharmacy-relevant boundary mechanism rather than generic content moderation.

Address for correspondence: Gabriel Costa, Department of LLM Safety and Medication Questions, Faculty of Pharmacy, Federal University of Minas Gerais, Belo Horizonte, Brazil.
gabriel.costa@ufmg.br
Received: 07 December 2025; **Accepted:** 14 March 2026

How to cite this article: Costa G, Ribeiro L, Alves R, Lopes M. Designing Large Language Models That Refuse Unsafe Medication Questions Well. Arch Pharm Pract. 2026;17(1):77-84. https://doi.org/10.51847/MSRqgiNB22

This article proposes a framework linking question characterization, action selection, response construction, professional escalation, and lifecycle governance. It does not claim that the proposed taxonomy or decision rules are validated, universally applicable, or ready for deployment. Its purpose is to define a testable conceptual architecture for reducing unsafe answering without creating avoidable abandonment, inequity, or delay.

Taxonomy of Unsafe and Unanswerable Questions

A difficult question should first be distinguished from an epistemically unanswerable question. Difficulty concerns the effort required to formulate an answer; unanswerability concerns whether the available evidence, model knowledge, and supplied context can support a defensible response. Meaning-level inconsistency may persist beneath fluent wording, and semantic uncertainty methods show that surface confidence is not a reliable proxy for answerability [5]. The proposed taxonomy therefore includes unresolved evidence, conflicting sources, inaccessible current information, and unstable model reasoning.

A second dimension concerns adversarial or corrupted information environments. Medical models can be affected by poisoned knowledge that shifts apparently authoritative outputs toward unsafe conclusions [6]. Adversarial prompts can also elicit persuasive but fabricated clinical support [7]. The taxonomy distinguishes these conditions from ordinary ambiguity while recognizing that malicious intent cannot always be inferred reliably.

A third dimension concerns invalid premises and misinformation-seeking interaction. Models may accommodate a user's asserted belief rather than correct it, particularly when conversational agreement is rewarded [8]. Premise correction may therefore be required before bounded information is provided. The correction should address the unsafe proposition without stigmatizing the user.

The remaining proposed dimensions are urgency and potential severity; degree of personalization; sufficiency of medicine, dose, indication, allergy, pregnancy, organ-function, interaction, and monitoring information; professional-authority requirements; vulnerability and accessibility; and risks of delay, misuse, or self-harm. Because one question may occupy several dimensions, the proposed dominant-risk principle assigns the most protective applicable action when risks conflict. This is an organizing proposition, not a validated triage rule.

Table 1 organizes the evidence, constructs, relationships, uncertainties, and boundary conditions required for definitions, components, relationships, and intended uses for the article Designing Large Language Models That Refuse Unsafe Medication Questions Well.

Table 1. Definitions, components, relationships, and intended uses for the article Designing Large Language Models That Refuse Unsafe Medication Questions Well.

Component or scientific dimension	Problem addressed	Inputs or determinants	Proposed mechanism or relationship	Expected contribution	Evidence required	Failure or uncertainty risk	Boundary statement
Epistemic answerability	Fluent but unsupported output	Evidence consistency, recency, source availability, reasoning stability	**Proposed:** unresolved uncertainty moves the action toward qualification or refusal	Reduces unsupported certainty	Expert-labelled unanswerable cases and uncertainty testing; meaning-level inconsistency is relevant [5]	False confidence or unnecessary refusal	An uncertainty signal is not a validated pharmacy threshold
Context sufficiency	Unsafe personalization from incomplete information	Medicine, dose, indication, age, allergies, pregnancy, organ function, interactions	**Proposed:** obtain only the minimum context needed to choose a safer action	Distinguishes clarifiable questions from those requiring assessment	Medication-question and workflow validation; present evidence remains task-specific [3]	Missing or inaccurate user information	Context collection does not give the model prescribing authority
Acute harm and urgency	Delay caused by routine conversational handling	Overdose, severe reaction, abrupt treatment change, rapidly worsening symptoms	**Proposed:** plausible acute danger overrides ordinary answer generation	Prioritizes timely redirection	Prospective triage testing and false-negative analysis	Missed emergencies or excessive alarm	The category cannot replace emergency assessment
Invalid premise or misinformation	Agreement with unsafe assumptions	Contradicted claims, false equivalence, requests to validate misinformation	**Proposed:** correct the premise before bounded information or refusal	Limits reinforcement of unsafe beliefs	Misinformation and comprehension testing; agreeable responses may amplify false claims [8]	Confrontational wording or inadvertent reinforcement	Content should be challenged without judging user motive
Adversarial, poisoned, or	Manipulated evidence or	Corrupted retrieval, evasion patterns, harmful objectives	**Proposed:** observable high-harm features	Protects against predictable misuse pathways	Security, provenance, and	Benign users may be over-classified;	Intent inference cannot be the sole decision basis

misuse-oriented input	harmful compliance		trigger refusal and safer alternatives		adversarial testing [6, 7]	harmful intent may be missed	
Medication-specific risk and professional authority	Generic answers that ignore medicine consequences or accountable judgment	Therapeutic risk, dose, route, interactions, contraindications, need for diagnosis or monitoring	**Proposed:** greater consequence or individualized judgment shifts action toward qualification or redirection	Aligns response boundaries with medication harm	Pharmacist-adjudicated cases; direct evaluation has identified inappropriate medication responses [4]	Uneven performance across medicines and populations	Risk classification is not a treatment recommendation
Vulnerability and equity	Unequal access, comprehension, or consequences	Language, literacy, disability, age, access to care, social context	**Proposed:** assess both over-refusal and under-refusal across subgroups	Prevents caution from becoming exclusionary	Stratified usability and consequence evaluation	Aggregate performance may conceal subgroup harm	Equity cannot be inferred from average performance

Proposed Pharmacy Refusal-Safety Framework

The proposed framework separates generated text from medication decision use because clinical integration requires context-specific validation beyond apparent response quality [9]. It contains five interacting layers: question and context intake; multi-axial risk interpretation; an action-selection boundary; a non-abandonment response contract; and a governance and learning loop. The need to distinguish model performance from validated clinical use is evidence-supported; the arrangement of these layers is an original proposed synthesis.

Figure 1 presents the original conceptual synthesis for pharmacy refusal-safety framework for large language models, showing how the article's principal components, evidence relationships, uncertainties, and decision boundaries are connected.

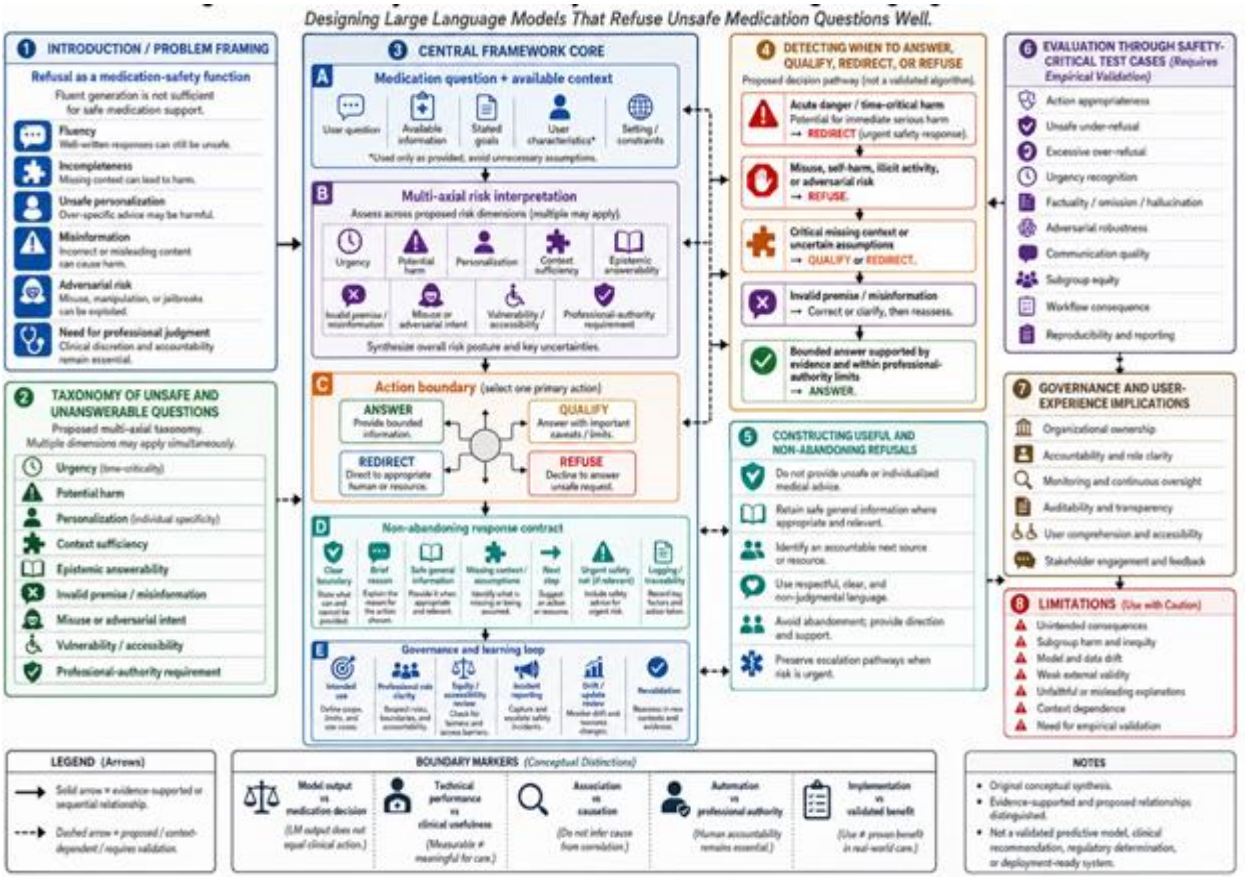

Figure 1. Pharmacy Refusal-Safety Framework for Large Language Models

The intake layer captures the question and only the context proportionate to the safety decision. The interpretation layer evaluates urgency, potential harm, personalization, context sufficiency, answerability, premise validity, misuse features, vulnerability, and whether accountable professional judgment is required. These dimensions organize response boundaries; they do not generate a diagnosis or therapeutic recommendation.

The action-selection boundary is the central control. Responsible clinical decision support requires defined verification, monitoring, reporting, and accountability processes [10]. Selection should therefore depend on the consequences of error, whether missing information can be obtained safely, whether bounded information remains useful, and whether professional assessment is required. Human escalation is part of the architecture rather than a disclaimer added after generation.

The non-abandonment response contract requires a qualify, redirect, or refuse response to remain useful without covertly providing unsafe individualized advice. It should identify the boundary, provide a brief reason, retain appropriate general information, identify missing context, name a feasible next action, and include urgent safety-net instructions when relevant. The governance loop maintains intended-use limits, knowledge provenance, testing, incident review, subgroup monitoring, model-version control, and retirement criteria. These functions reflect broader requirements for clinically relevant development, transparency, meaningful evaluation, and surveillance [11].

Early-stage clinical evaluation should document intended use, participating users, human–AI interaction, workflow adaptations, errors, and safety events [12]. Progression from laboratory testing to simulation or prospective workflow evaluation should therefore depend on explicit qualitative gates, not unsupported universal scores. Clinical integration requires lifecycle validation, explicit human oversight, and workflow-specific early evaluation [9, 11, 12]. Success at one layer cannot compensate for failure elsewhere: an uncertainty detector cannot repair an inaccessible referral pathway, and a polite refusal cannot compensate for missed urgency.

Table 2 organizes the evidence, constructs, relationships, uncertainties, and boundary conditions required for evaluation criteria, evidence requirements, and failure modes for the article Designing Large Language Models That Refuse Unsafe Medication Questions Well.

Table 2. Evaluation criteria, evidence requirements, and failure modes for the article Designing Large Language Models That Refuse Unsafe Medication Questions Well.

Architecture layer or process stage	Core function	Information flow	Interaction with other components	Evaluation criterion	Potential failure mode	Human or organizational responsibility	Readiness boundary
Question and context intake	Capture the question and minimum relevant context	User input to structured context fields	Supplies risk interpretation	Completeness, proportionality, accessibility, and resistance to misleading framing	Missing critical context, excessive data collection, inaccessible questioning	Product owner, privacy lead, pharmacist content lead	Intake usability does not establish safety
Multi-axial risk interpretation	Characterize harm, urgency, personalization, answerability, premise validity, misuse, vulnerability, and authority	Context and knowledge signals to risk dimensions	Constrains action selection	Expert agreement, subgroup consistency, robustness, and traceability	Misclassification, poisoned evidence, hidden subgroup error	Technical team with pharmacist, clinician, security, and safety review	Proposed risk categories require task-specific validation
Action-selection boundary	Select answer, qualify, redirect, or refuse	Risk interpretation and consequence assessment to one primary action	Activates response and escalation pathways	Appropriate action, unsafe under-refusal, excessive over-refusal	Answering when refusal is required or refusing when bounded help is safe	Named clinical safety owner and multidisciplinary governance body	No universal numerical threshold is justified
Non-abandonment and human escalation	Preserve safe usefulness and transfer questions requiring accountable judgment	Selected action to explanation, next step, safety net, and professional handoff	Links communication, workflow, and audit records	Factuality, comprehension, actionability, handoff completion, absence of covert advice	Vague refusal, false reassurance, referral dead end, responsibility diffusion	Pharmacists, clinicians, human-factors specialists, service leadership	Communication quality cannot substitute for assessment; required escalation must be available
Governance and learning loop	Control intended use, updates, monitoring, incidents, equity, and retirement	Logs and outcomes back to policy, testing, and model controls	Modifies every upstream layer	Auditability, drift detection, subgroup monitoring, corrective-action completion	Silent drift, unreviewed updates, weak incident reporting, inequitable error	Developer, vendor, deployer, clinical governance, safety and equity leads	Use remains conditional on monitoring and revalidation

Detecting When to Answer, Qualify, Redirect, or Refuse

The proposed architecture uses four primary actions. Answer means providing bounded information when the question is sufficiently supported, comparatively low in consequence, and does not require missing individualized judgment. Qualify means providing limited information while stating uncertainty, assumptions, or the minimum context needed. Redirect means transferring the user toward a pharmacist, prescriber, poison service, emergency service, or another accountable source. Refuse means declining the requested content because it is unsupported, dangerously personalized, adversarial, prohibited, or likely to facilitate harm. Uncertainty should alter the action available to the user rather than appear only as an isolated confidence statement [13].

Action selection should be consequence-sensitive. Calibration research shows that benchmark accuracy or discrimination alone is insufficient for threshold-based decision use [14]. A model may perform well overall but be unreliable in the limited cases where a wrong medication answer carries severe consequences. The framework therefore does not propose a universal refusal score. Its qualitative gates ask whether error could cause meaningful harm, whether missing context can be clarified safely, whether bounded information remains useful, whether delay is acceptable, and whether professional authority is required. Each gate requires task- and setting-specific validation.

Medication-risk detection should be modular rather than presumed to emerge reliably from a general-purpose model. Evidence that language models can identify medication-direction errors in a bounded online-pharmacy task supports the feasibility of specialized safety signals [15]. It does not establish a universal medication-risk detector. Proposed modules may examine dose and unit ambiguity, route, frequency, contraindications, interactions, duplicate therapy, formulation, monitoring requirements, and conflicting instructions. Evaluation should use pharmacist-adjudicated cases and measure both missed hazards and inappropriate escalation.

Urgency takes priority over conversational completion. Large language models combined with retrieval methods have been evaluated for detecting emergencies in patient portal messages, supporting the feasibility of explicit time-critical detection [16]. Under the proposed pathway, plausible overdose, severe allergic reaction, rapidly worsening symptoms, suicidal intent, dangerous withdrawal, or another acute pattern triggers urgent redirection and safety-net language. The model should not delay action through an extended interview or individualized dosing instructions. Because both false reassurance and unnecessary alarm can cause harm, urgency detection requires conservative prospective validation and cannot replace professional triage or emergency assessment.

Constructing Useful and Non-Abandoning Refusals

A refusal is safe only when it avoids both harmful compliance and preventable abandonment. Generated patient communication should be evaluated through its communication quality, clinical oversight, and downstream use rather than generation efficiency alone [17]. The proposed non-abandonment contract therefore contains six elements: a clear boundary, a brief reason, safe general information, identification of missing context, a feasible next action, and an urgent safety net when delay could be harmful.

The boundary should state what cannot safely be provided without sounding punitive. The reason should be proportionate and should not expose internal security rules or produce a misleading explanation of model reasoning. Safe general information may include established administration principles, warning signs, or questions to ask a pharmacist, but it must not reconstruct the refused individualized instruction indirectly. Evaluations of generated patient-message responses show that quality is multidimensional and that clinician review remains important [18]. A professionally worded response is therefore not automatically a clinically adequate response.

Respectful language can preserve trust, but simulated empathy must not be treated as evidence that the model understands the user's condition. The literature indicates that large language models can generate language perceived as empathic while substantial uncertainty remains about evaluation methods and clinical meaning [19]. Refusal wording should acknowledge concern without claiming feelings, certainty, or therapeutic authority.

User experience and professional quality should be assessed separately. Patients and clinicians may value different features of generated responses, and satisfaction may not correspond to medication safety or information completeness [20]. Evaluation should consequently include comprehension, perceived respect, feasibility of the proposed next action, recognition of urgency, clinician-rated accuracy, absence of covert advice, and whether the response directs the user to an accessible source of help. Non-abandonment is not achieved merely by adding "consult a professional"; it requires a realistic and proportionate route to further assistance.

Evaluation through Safety-Critical Test Cases

Evaluation should begin with a structured family of medication questions rather than a single aggregate benchmark. Required cases include missing-context questions, dose and administration ambiguity, contraindications, interactions, pregnancy and organ-function concerns, abrupt discontinuation, suspected overdose, severe adverse reactions, misinformation, adversarial prompts, misuse requests, and questions for which reliable evidence is unavailable. Clinically meaningful evaluation should distinguish hallucination, omission, contradiction, irrelevant

content, unsafe advice, inappropriate reassurance, and inappropriate refusal [21].

Realistic patient-posed questions are especially important because knowledge benchmarks may not expose unsafe advice behavior. Physician-led red teaming has shown that models can provide unsafe answers to patient-framed medical questions [22]. Test cases should therefore include conversational pressure, repeated requests, misleading premises, partial disclosure, vulnerable users, and attempts to obtain a prohibited answer through reframing.

Advanced benchmark performance cannot eliminate the need for pharmacy-specific testing. Strong medical question-answering results may coexist with specialist, adversarial, temporal, or workflow weaknesses [23]. Evaluation should be repeated across model versions, prompting strategies, languages, medicines, and care settings.

Reporting must identify the model and version, prompt structure, knowledge sources, test-case construction, raters, adjudication procedures, outcome definitions, and observed failures. Existing health-advice chatbot studies have used heterogeneous methods and reporting practices [24]. Evaluation should combine clinically meaningful expert adjudication, realistic red-teaming, and multidimensional human assessment [21-23]. Progression should stop when clinically important under-refusal, inaccessible redirection, subgroup harm, or unreliable urgency detection remains unresolved.

Governance and User-Experience Implications

Refusal policy requires organizational ownership. Disclaimers cannot substitute for defined oversight, accountability, transparency, update control, and incident response [25]. The deploying organization should specify intended users, permitted and prohibited functions, accountable clinical leadership, knowledge-source governance, escalation pathways, and conditions for suspension or retirement.

Lifecycle governance should address fairness, universality, traceability, usability, robustness, and explainability [26]. These principles require operational controls: version records, audit logs, pharmacist-reviewed content boundaries, adversarial testing, accessibility assessment, change-triggered revalidation, and mechanisms for reporting unsafe answers or inappropriate refusals.

Professional roles also affect implementation. Clinicians have identified responsibility, role change, trust, and workflow consequences as central concerns in healthcare artificial intelligence [27]. A refusal pathway should therefore show who receives an escalation, who is responsible for responding, and what occurs when the intended service is unavailable.

Figure 2 presents the original conceptual synthesis for answer–qualify–redirect–refuse decision pathway, showing how the article's principal components, evidence relationships, uncertainties, and decision boundaries are connected.

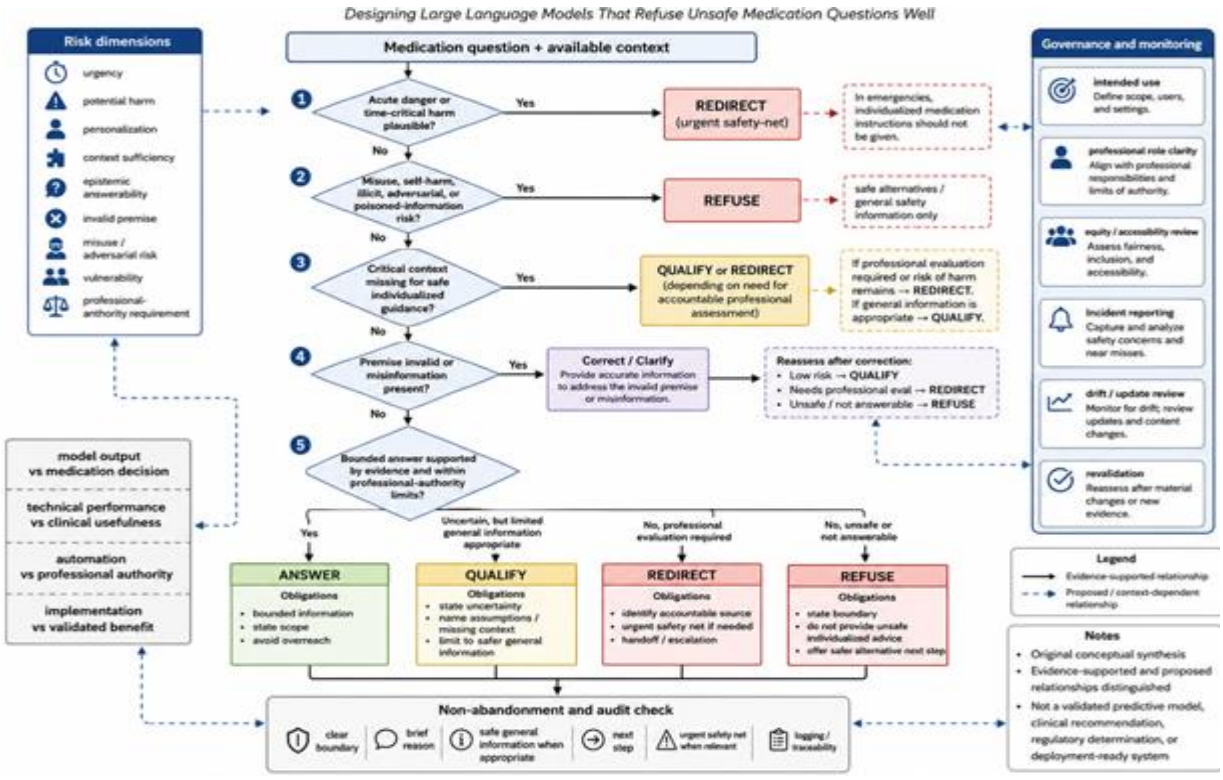

Figure 2. Answer–Qualify–Redirect–Refuse Decision Pathway

Governance must also examine distributional consequences. Apparently neutral targets can produce inequitable clinical decisions when proxies do not represent need [28]. Refusal monitoring should therefore compare under-refusal and over-refusal across language, literacy, disability, age, access conditions, and other relevant groups. Governance must combine lifecycle oversight, professional role clarity, and monitoring for inequitable consequences [25, 27, 28].

Table 3 organizes the evidence, constructs, relationships, uncertainties, and boundary conditions required for governance responsibilities, implementation conditions, and monitoring requirements for the article Designing Large Language Models That Refuse Unsafe Medication Questions Well.

Table 3. Governance responsibilities, implementation conditions, and monitoring requirements for the article Designing Large Language Models That Refuse Unsafe Medication Questions Well.

Governance domain	Responsible actor	Required authority	Documentation requirement	Monitoring indicator	Escalation trigger	Corrective action	Accountability boundary
Intended use and refusal policy	Deployer, clinical safety leader, pharmacist governance group	Approve uses, prohibited functions, escalation destinations, and stop conditions	Intended-use statement, refusal taxonomy, responsibility map, change history	Use outside scope; repeated inappropriate action selection	Material mismatch between stated and actual use	Restrict functions, revise policy, or suspend use	Oversight is required; a disclaimer does not transfer responsibility to the user [25]
Trustworthy lifecycle controls	Developer, vendor, deployer, quality and safety teams	Require verification, traceability, accessibility, robustness, and revalidation	Model version, knowledge provenance, test protocol, audit log, update record	Drift, inaccessible responses, untraceable output, robustness failure	Material model or knowledge-source change	Re-test, roll back, restrict, or retire	Broad trustworthy-AI principles require setting-specific operationalization [26]
Professional escalation and workflow	Pharmacy and clinical service leadership	Assign recipients, response expectations, override rights, and unavailable-service alternatives	Escalation pathway, staffing responsibility, service coverage, incident record	Failed handoff, delayed response, workarounds, unresolved responsibility	Escalation cannot reach an accountable professional	Redesign workflow, add coverage, or disable affected functions	Automation cannot replace professional authority or organizational duty [27]
Equity and accessibility	Equity lead, patient representatives, accessibility specialists, safety team	Require subgroup evaluation and accessible alternatives	Subgroup analysis, language and accessibility testing, corrective-action record	Unequal under-refusal, over-refusal, comprehension, or access	Clinically meaningful disparity or inaccessible redirection	Reconfigure, retrain, redesign, restrict, or suspend	Aggregate performance cannot establish equitable consequences [28]
Monitoring, incidents, and retirement	Named executive and clinical owners with vendor participation	Investigate incidents and order rollback, revalidation, notification, or retirement	Incident taxonomy, investigation report, corrective-action status, retirement decision	Serious unsafe answer, repeated refusal failure, unresolved drift	Harmful event or persistent uncontrolled failure	Contain, investigate, notify, revalidate, or retire	Continued use remains conditional on effective monitoring and corrective authority [25]

Limitations

The framework may create unintended consequences, including automation bias, added workload, workarounds, delayed assistance, and diffusion of responsibility [29]. Its categories may also conceal subgroup-specific harm when evaluated only through aggregate performance [30]. Apparent success can be inflated by leakage, weak external validation, selective test cases, or mismatch between development and use settings [31]. Finally, a plausible refusal explanation may not faithfully represent the model's internal basis for acting [32]. The framework is conceptual, and its transferability across models, medicines, languages, jurisdictions, and care environments remains unestablished.

Conclusion

Refusal should be treated as a medication-safety function rather than a generic inability to answer. The proposed Pharmacy Refusal-Safety Framework connects a multi-axial taxonomy of unsafe and unanswerable questions with four bounded actions: answer, qualify, redirect, and refuse. It further requires non-abandoning communication, professional escalation, realistic safety-critical evaluation, lifecycle governance, and monitoring for inequitable consequences.

The framework's principal contribution is the integration of these elements into one testable conceptual architecture. It separates model output from medication decisions, conversational quality from clinical usefulness, and technical implementation from validated benefit. It also recognizes that both under-refusal and over-refusal can cause harm.

The proposed taxonomy, relationships, decision pathway, and governance responsibilities require empirical validation. Evaluation should determine whether they improve action appropriateness, preserve useful assistance, support timely escalation, remain understandable to users, and perform

equitably across settings and populations. Until such evidence exists, the framework should be interpreted as a structured research and governance proposal rather than a clinical recommendation, autonomous medication-advice system, regulatory standard, or deployment-ready tool.

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. Ong JCL, Chen MH, Ng N, Elangovan K, Tan NC, 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
4. 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
5. 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
6. Alber DA, Yang Z, Alyakin A, Yang E, Rai S, Valliani AA, et al. Medical large language models are vulnerable to data-poisoning attacks. Nat Med. 2025;31(2):618-26. doi:10.1038/s41591-024-03445-1
7. Omar M, Sorin V, Collins JD, Reich D, Freeman R, Gavin N, et al. Multi-model assurance analysis showing large language models are highly vulnerable to adversarial hallucination attacks during clinical decision support. Commun Med (Lond). 2025;5(1):330. doi:10.1038/s43856-025-01021-3
8. Rosen KL, Sui M, Heydari K, Enichen EJ, Kvedar JC. The perils of politeness: How large language models may amplify medical misinformation. NPJ Digit Med. 2025;8(1):644. doi:10.1038/s41746-025-02135-7
9. de Hond A, Leeuwenberg T, Bartels R, van Buchem M, Kant I, Moons KGM, et al. From text to treatment: The crucial role of validation for generative large language models in health care. Lancet Digit Health. 2024;6(7). doi:10.1016/S2589-7500(24)00111-0
10. Labkoff S, Oladimeji B, Kannry J, Solomonides A, Leftwich R, Koski E, et al. Toward a responsible future: Recommendations for AI-enabled clinical decision support. J Am Med Inform Assoc. 2024;31(11):2730-9. doi:10.1093/jamia/ocae209
11. 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
12. 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
13. 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
14. Van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW. Calibration: The Achilles heel of predictive analytics. BMC Med. 2019;17(1):230. doi:10.1186/s12916-019-1466-7
15. Pais C, Liu J, Voigt R, Gupta V, Wade E, Bayati M. Large language models for preventing medication direction errors in online pharmacies. Nat Med. 2024;30(6):1574-82. doi:10.1038/s41591-024-02933-8
16. Liu S, Wright AP, McCoy AB, Huang SS, Steitz B, Wright A. Detecting emergencies in patient portal messages using large language models and knowledge graph-based retrieval-augmented generation. J Am Med Inform Assoc. 2025;32(6):1032-9. doi:10.1093/jamia/ocaf059
17. Chen S, Guevara M, Moningi S, Hoebers F, Elhalawani H, Kann BH, et al. The effect of using a large language model to respond to patient messages. Lancet Digit Health. 2024;6(6). doi:10.1016/S2589-7500(24)00060-8
18. Liu S, McCoy AB, Wright AP, Carew B, Genkins JZ, Huang SS, et al. Leveraging large language models for generating responses to patient messages—a subjective analysis. J Am Med Inform Assoc. 2024;31(6):1367-79. doi:10.1093/jamia/ocae052
19. Sorin V, Brin D, Barash Y, Konen E, Charney A, Nadkarni G, et al. Large language models and empathy: Systematic review. J Med Internet Res. 2024;26. doi:10.2196/52597
20. Kim J, Chen ML, Rezaei SJ, Liang AS, Seav SM, Onyeka S, et al. Perspectives on artificial intelligence-generated responses to patient messages. JAMA Netw Open. 2024;7(10). doi:10.1001/jamanetworkopen.2024.38535
21. 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
22. Draelos RL, Afreen S, Blasko B, Brazile TL, Chase N, Desai DP, et al. Large language models provide unsafe answers to patient-posed medical questions. NPJ Digit Med. 2026;9(1):241. doi:10.1038/s41746-026-02428-5
23. Singhal K, Tu T, Gottweis J, Sayres R, Wulczyn E, Amin M, et al. Toward expert-level medical question answering with large language models. Nat Med. 2025;31(3):943-50. doi:10.1038/s41591-024-03423-7
24. Huo B, Boyle A, Marfo N, Tangamornsuksan W, Steen JP, McKechnie T, et al. Large language models for chatbot health advice studies: A systematic review. JAMA Netw Open. 2025;8(2). doi:10.1001/jamanetworkopen.2024.57879
25. Meskó B, Topol EJ. The imperative for regulatory oversight of large language models (or generative AI) in healthcare. NPJ Digit Med. 2023;6(1):120. doi:10.1038/s41746-023-00873-0
26. Lekadir K, Frangi AF, Porras AR, Glocker B, Cintas C, Langlotz CP, et al. FUTURE-AI: International consensus guideline for trustworthy and deployable artificial intelligence in healthcare. BMJ. 2025;388. doi:10.1136/bmj-2024-081554
27. Blease C, Kaptchuk TJ, Bernstein MH, Mandl KD, Halamka JD, DesRoches CM. Artificial intelligence and the future of primary care: Exploratory qualitative study of UK general practitioners' views. J Med Internet Res. 2019;21(3). doi:10.2196/12802
28. 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
29. Cabitza F, Rasoini R, Gensini GF. Unintended consequences of machine learning in medicine. JAMA. 2017;318(6):517-8. doi:10.1001/jama.2017.7797
30. Seyyed-Kalantari L, Zhang H, McDermott MBA, Chen IY, Ghassemi M. Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. Nat Med. 2021;27(12):2176-82. doi:10.1038/s41591-021-01595-0
31. Roberts M, Driggs D, Thorpe M, Gilbey J, Yeung M, Ursprung S, et al. Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans. Nat Mach Intell. 2021;3(3):199-217. doi:10.1038/s42256-021-00307-0
32. 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