JOURNAL OF MULTIDISCIPLINARY RESEARCH AND INTEGRATED SCIENCES

Volume 4, Issue 1, 2024 **Article ID:** JMRIS-2024-002 **Pages:** 12-22 **Article Type:** Systematic Comparative Review
Received: 05-01-2024 **Revised:** 27-01-2024 **Accepted:** 21-02-2024 **Published:** 15-03-2024

Accuracy of Artificial Intelligence Chatbots in Responding to Patient Questions About Anticoagulant Therapy: A Systematic Comparative Evidence Synthesis and Cross-Platform Evaluation Framework

Daniel Foster1*

1 Temerty Faculty of Medicine, University of Toronto, Toronto, Canada

***Corresponding author:** Daniel Foster, Temerty Faculty of Medicine, University of Toronto, Toronto, Canada| fosterd21421@gmail.com

ABSTRACT

Background: Oral anticoagulants are high-risk medicines in which misunderstanding can contribute to bleeding, thrombosis, treatment interruption, or inappropriate self-management. Large language model chatbots can produce immediate conversational answers to patient questions, but the safety of using general medical performance as a proxy for anticoagulation-specific accuracy is uncertain. **Objective:** To synthesize high-quality evidence available through 31 December 2023 regarding the accuracy, completeness, reliability, and safety of artificial intelligence chatbots for medical question answering and to translate that evidence into a reproducible cross-platform framework for anticoagulant patient questions. **Methods:** A structured evidence synthesis was conducted using Q1 literature in medicine, cardiology, thrombosis, digital health, and artificial intelligence. Direct chatbot studies were combined with anticoagulation guidelines and patient-knowledge research. Because no qualifying pre-2024 Q1 study directly compared multiple public chatbots specifically on anticoagulant patient questions, no model-specific anticoagulation performance values were fabricated. **Results:** In 2023 studies, ChatGPT responses were judged appropriate for 21 of 25 cardiovascular prevention questions, were preferred to physician responses in 78.6% of evaluations in a general patient-question study, and achieved a median physician-rated accuracy score of 5.5/6 across 284 medical questions. Med-PaLM achieved 92.6% alignment with scientific consensus on consumer medical questions, yet 5.9% of its answers were still judged potentially harmful. Other studies demonstrated specialty-specific inaccuracies, material response instability, and hallucinated references. Anticoagulant questions are especially sensitive to omissions involving missed doses, bleeding, drug interactions, procedures, renal function, adherence, and the difference between warfarin and direct oral anticoagulants. **Conclusion:** General-purpose chatbots showed promising medical question-answering performance by the end of 2023, but the evidence did not support unsupervised anticoagulant counseling or a definitive ranking of public platforms. Cross-platform evaluation should use identical prompts, repeated generations, independent anticoagulation experts, guideline-based reference answers, and separate scoring of accuracy, completeness, readability, actionability, consistency, and potential harm.

Keywords: anticoagulants; ChatGPT; large language models; patient education; warfarin; direct oral anticoagulants; medication safety; artificial intelligence

1. Introduction

Oral anticoagulants are central to the prevention and treatment of thromboembolic disease, including stroke prevention in atrial fibrillation and treatment or secondary prevention of venous thromboembolism. Their benefits are substantial, but the therapeutic objective is inseparable from bleeding risk. Safe use therefore depends on patients understanding why the medicine is prescribed, how it should be taken, what to do after a missed dose, which interactions matter, when urgent bleeding requires assessment, and how therapy should be managed around procedures [1–4].

The informational burden differs between warfarin and direct oral anticoagulants (DOACs). Warfarin requires international normalized ratio monitoring and stable attention to interacting medicines and dietary vitamin K, whereas DOACs generally do not require routine coagulation monitoring but rely heavily on correct dosing, renal-function assessment, and adherence because of their shorter half-lives [1]. The 2021 European Heart Rhythm Association practical guide devotes specific recommendations to dosing errors, drug interactions, renal and hepatic disease, bleeding, emergency surgery, planned procedures, and adherence, illustrating how apparently simple patient questions can contain clinically consequential detail [1].

Patient knowledge is not consistently complete. In a European Heart Rhythm Association survey of 1,147 patients with atrial fibrillation, important gaps remained despite widespread anticoagulant use, and a substantial proportion of patients reported missed doses or temporary treatment interruption [5]. The emergence of public artificial intelligence chatbots therefore creates an intuitively attractive opportunity: patients can ask questions in natural language and receive immediate personalized explanations.

Evidence published in 2023 demonstrated impressive but non-uniform medical performance. Sarraju et al. found that a popular chatbot gave appropriate responses to 84% of cardiovascular prevention questions [6]. Ayers et al. reported that chatbot answers to general patient questions were preferred over physician answers in 78.6% of blinded evaluations [7]. Goodman et al. found a median accuracy score of 5.5 on a six-point scale for 284 physician-generated questions [8]. At the same time, specialty-specific investigations documented incorrect treatment recommendations, inconsistent responses, and fabricated references [11,12].

These limitations are particularly important for anticoagulation. An answer can be fluent, reassuring, and broadly correct yet remain unsafe if it omits the need for urgent assessment after major bleeding, gives an incorrect missed-dose instruction, treats all DOACs as interchangeable, recommends unsupervised interruption before a procedure, or fails to account for renal function and interacting medication. This review therefore examined the evidence available through 2023 and developed a reproducible framework for comparing public chatbots on anticoagulant patient questions without manufacturing model-specific results that had not actually been studied.

2. Methods

2.1 Review design and evidence cutoff

A structured comparative evidence synthesis was undertaken using literature available through 31 December 2023. The review focused on two connected evidence streams. The first comprised Q1 studies evaluating large language models or generative chatbots for medical or patient-facing question answering. The second comprised Q1 anticoagulation guidelines and patient-education or knowledge studies defining the content that a safe anticoagulant answer should contain. Studies published after the cutoff were not used to infer 2023 performance.

2.2 Eligibility and synthesis

Eligible chatbot studies evaluated open-ended medical questions, patient questions, factual agreement with clinical consensus, completeness, potential harm, response quality, or reproducibility. Anticoagulation sources were included when they addressed warfarin, DOACs, VTE treatment, atrial-fibrillation anticoagulation, bleeding management, or patient knowledge. A de novo meta-analysis was not performed because the chatbot studies used different models, prompts, question domains, evaluators, and outcome scales. No cross-platform anticoagulant comparison was inferred from non-equivalent datasets.

The eligibility framework for the evidence synthesis is summarised in Table 1.

3. Results

3.1 General medical chatbot performance available by 2023

The available evidence showed substantial capability but also clinically important residual error. Sarraju et al. posed 25 fundamental cardiovascular prevention questions three times and had preventive cardiology clinicians judge each set. Twenty-one questions, or 84%, were rated appropriate, whereas four were inappropriate; errors included overconfident exercise advice, incomplete interpretation of markedly elevated LDL cholesterol, and outdated information regarding a newer therapy [6]. The authors specifically recommended broader model comparisons, formal grading systems, and readability assessment.

Ayers et al. compared ChatGPT with verified physician responses to 195 patient questions from an online medical forum. Licensed healthcare professionals preferred the chatbot response in 78.6% of 585 evaluations. Good or very good quality ratings were assigned to 78.5% of chatbot responses compared with 22.1% of physician responses, although chatbot responses were substantially longer [7]. These findings support usefulness in drafting patient communication but do not establish medication-specific safety.

Goodman et al. extended evaluation to 284 physician-generated medical questions across 17 specialties. The median accuracy score was 5.5 of 6 and median completeness was 3 of 3. However, low-scoring answers occurred and responses improved when questions were repeated or generated with a newer model version, demonstrating that performance can change across time and model versions [8].

The Med-PaLM study provides a more explicitly safety-oriented benchmark. Clinicians judged 92.6% of Med-PaLM long-form consumer answers aligned with scientific consensus, similar to clinician-generated answers at 92.9%. Nevertheless, 5.9% of Med-PaLM answers were judged potentially capable of leading to harmful outcomes, and lay users still rated clinician answers as more helpful overall [9]. Thus, high average performance does not eliminate the residual risk relevant to high-stakes medication counseling.

Selected 2023 Q1 evidence on large-language-model medical question answering is summarised in Table 2.

Selected 2023 benchmarks for large-language-model medical question answering are presented in Figure 1.

3.2 Anticoagulant information is intrinsically safety sensitive

Anticoagulant counseling contains multiple drug-specific branches. For DOACs, EHRA guidance emphasizes correct dose selection, renal function, interactions, adherence, dosing errors, bleeding, and peri-procedural management [1]. For VTE, the CHEST guideline distinguishes treatment phases and recommends therapy according to disease setting and patient characteristics [3]. For major anticoagulant-related bleeding, the American College of Cardiology consensus pathway addresses interruption, source control, reversal, and subsequent decisions regarding resumption [2].

These documents show why a generic statement such as 'take the missed dose when you remember' is insufficient. Appropriate advice varies by the anticoagulant, dosing frequency, and time since the scheduled dose. Similarly, telling a patient to stop an anticoagulant before dental work or surgery without individualized procedural guidance can create thrombosis risk. A chatbot evaluation must therefore score clinically important omissions separately from obvious factual errors.

Core patient-question domains for anticoagulant chatbot evaluation are summarised in Table 3.

3.3 Patient knowledge establishes the need for clear counseling

The problem addressed by chatbots is clinically real. In the European Heart Rhythm Association patient survey, anticoagulant knowledge was uneven despite established therapy [5]. Most warfarin-treated patients knew their target INR, yet medication education remained incomplete, and missed doses or treatment interruptions were reported. This reinforces the value of accessible question answering while simultaneously increasing the responsibility of any automated system to provide drug-specific, unambiguous answers.

The ability of an LLM to generate longer or more empathetic text may improve engagement, but excessive length can also bury the action most important to the patient. For anticoagulant therapy, a high-quality answer should communicate the immediate action first, explain the reason, state red-flag conditions, and clearly identify when professional review is required. The desired output is not maximal detail; it is the correct amount of actionable detail.

3.4 Proposed cross-platform comparative evaluation

A reproducible cross-platform study should evaluate the same public chatbot versions on the same day using identical prompts and fresh sessions. Each question should be generated repeatedly to estimate output instability. Responses should be anonymized and randomized before review by at least two clinicians with anticoagulation expertise, with adjudication for major disagreements. The reference standard should be assembled before chatbot testing from EHRA, CHEST, JACC consensus guidance, approved drug labeling, and locally applicable clinical standards.

Accuracy and completeness should be scored independently. A response that contains no false statement but omits the need for emergency assessment after severe bleeding is incomplete and potentially harmful. Conversely, a complete answer containing one incorrect dosing instruction may be more dangerous than a shorter but accurate response. The scoring system should therefore explicitly separate factual correctness from clinical safety.

The proposed scoring system for a future cross-platform anticoagulant chatbot study is presented in Table 4.

The proposed cross-platform safety framework for evaluating chatbot answers to anticoagulant patient questions is presented in Figure 2.

3.5 Failure modes likely to matter most

The pre-2024 medical LLM literature identifies several failure modes directly relevant to anticoagulation. Caranfa et al. showed that plausible specialty answers can contain incorrect treatment recommendations and that repeated queries can materially change responses [11]. Hua et al. found that approximately 29–33% of chatbot-generated references in their setting were nonverifiable, despite appearing realistic [12]. These findings suggest that an anticoagulant system should not gain credibility merely by displaying citations or by producing a polished explanation.

A second failure mode is temporal knowledge limitation. Sarraju et al. attributed one inappropriate cardiovascular response to training-data timing [6]. Anticoagulation practice changes over time through new indications, revised dosing or peri-procedural recommendations, reversal strategies, and safety communications. Model version and access date should therefore be reported as essential study variables.

Anticoagulant chatbot failure modes and their potential clinical consequences are summarised in Table 5.

4. Discussion

4.1 Principal finding

By the end of 2023, generative chatbots had demonstrated sufficiently strong medical language performance to justify serious investigation for patient education, but not sufficiently reliable safety performance to justify unsupervised anticoagulant counseling. The key issue is not whether a chatbot can produce a broadly correct explanation. It is whether the response remains correct at the exact points where misunderstanding changes patient behavior.

The evidence also argues against assuming that a stronger general medical model will necessarily be the safest anticoagulant counselor. General benchmarks aggregate many low- and moderate-risk questions. Anticoagulants disproportionately contain high-risk edge cases in which one omitted qualifier can be clinically important. A cross-platform study must therefore be built around medication-specific failure detection rather than average linguistic quality.

4.2 Accuracy should not be collapsed into a single score

The 2023 literature used heterogeneous endpoints: appropriateness, physician preference, quality, scientific-consensus alignment, accuracy, completeness, and potential harm [6–9]. Each captures a different property. An anticoagulant answer may be accurate but incomplete, complete but unreadable, readable but overconfident, or factually sound yet inappropriate for self-management. A single total score can hide these patterns.

Potential harm should therefore be treated as a gating outcome. If an answer advises dangerous interruption, incorrect dose compensation, or inadequate escalation for severe bleeding, excellent readability cannot compensate for the safety failure. In a patient-facing deployment framework, such answers should automatically fail overall suitability.

4.3 Cross-platform comparison requires strict experimental control

Direct model ranking requires identical conditions. Public chatbots can differ by underlying model, safety policies, retrieval capability, system prompts, and update schedule. Queries should therefore be issued within a narrow time window, with personalization disabled where possible, no conversation history, and identical wording. Each question should be repeated because the evidence already demonstrates output instability [8,11].

The study should also preserve the actual generated text. This allows later re-analysis when clinical guidance or model behavior changes. Reporting only a score without the prompt, model version, date, and output prevents meaningful replication.

4.4 Human oversight remains essential

The most defensible near-term use is clinician-supervised drafting rather than autonomous counseling. The Ayers study supports the potential for chatbots to draft high-quality, empathetic patient responses [7], while the medical benchmark literature shows that errors remain possible even in high-performing systems [8,9]. Anticoagulation pharmacists, cardiologists, hematologists, and trained clinicians can use generated text as a starting point but should verify dosing, interactions, procedural advice, and emergency instructions before release.

This model also preserves an important clinical distinction. Patient education explains therapy; medical decision-making individualizes it. Questions about what an anticoagulant is for may be suitable for automated education after validation. Questions asking whether a patient should stop, restart, reverse, or change a dose often require clinical context that a generic chatbot does not possess.

4.5 Strengths and limitations

The principal strength of this synthesis is that it integrates high-quality 2023 LLM evidence with authoritative anticoagulation guidance while refusing to manufacture a direct cross-platform comparison that had not yet been published in qualifying literature. The resulting framework is therefore suitable for a genuine subsequent empirical study.

The main limitation is the absence of qualifying pre-2024 Q1 studies directly comparing multiple public chatbots on oral-anticoagulant patient questions. The performance benchmarks in Figure 1 come from different medical domains and cannot be interpreted as model-specific anticoagulation accuracy.

Chatbot technology was also evolving rapidly during 2023, meaning performance was version dependent. Finally, clinical guidance can vary by indication, country, approved labeling, patient age, renal function, pregnancy status, and procedural setting.

4.6 Conclusion

Artificial intelligence chatbots had demonstrated substantial potential for medical question answering by the end of 2023, including strong performance on cardiovascular and general patient questions. Nevertheless, anticoagulant therapy represents a high-stakes test case in which the residual error rate matters more than conversational fluency. A safe answer must distinguish warfarin from DOACs, provide drug-specific missed-dose and monitoring information, recognize bleeding emergencies, avoid unsupervised peri-procedural interruption, account for interactions and renal function, and communicate uncertainty appropriately.

The available evidence does not justify declaring one public chatbot superior for anticoagulant counseling before a controlled head-to-head study is performed. The appropriate next step is a standardized cross-platform evaluation using identical patient questions, repeated generations, independent anticoagulation experts, predefined guideline-based reference answers, and separate measurement of accuracy, completeness, readability, consistency, actionability, source integrity, and potential harm. Until such validation is available, chatbot-generated anticoagulant information should be considered clinician-reviewable educational material rather than autonomous medical advice.

Declarations

Funding

No external funding was received.

Conflict of Interest

The authors declare no conflict of interest.

Ethical Approval

Not applicable. This manuscript synthesized previously published evidence and did not involve direct participant recruitment or patient-level data.

Consent to Participate

Not applicable.

Data Availability

No novel participant-level dataset was generated. The evidence base consists of previously published literature.

Author Contributions

The authors contributed to conceptualization, evidence synthesis, methodological framework development, interpretation, manuscript preparation, and critical revision.

References

1. Steffel J, Collins R, Antz M, Cornu P, Desteghe L, Haeusler KG, et al. 2021 European Heart Rhythm Association Practical Guide on the use of non-vitamin K antagonist oral anticoagulants in patients with atrial fibrillation. Europace. 2021;23(10):1612-1676. doi:10.1093/europace/euab065.

2. Tomaselli GF, Mahaffey KW, Cuker A, Dobesh PP, Doherty JU, Eikelboom JW, et al. 2020 ACC Expert Consensus Decision Pathway on Management of Bleeding in Patients on Oral Anticoagulants. J Am Coll Cardiol. 2020;76(5):594-622. doi:10.1016/j.jacc.2020.04.053.

3. Stevens SM, Woller SC, Kreuziger LB, Bounameaux H, Doerschug K, Geersing GJ, et al. Antithrombotic Therapy for VTE Disease: Second Update of the CHEST Guideline and Expert Panel Report. Chest. 2021;160(6):e545-e608. doi:10.1016/j.chest.2021.07.055.

4. January CT, Wann LS, Calkins H, Chen LY, Cigarroa JE, Cleveland JC Jr, et al. 2019 AHA/ACC/HRS Focused Update of the 2014 Guideline for the Management of Patients With Atrial Fibrillation. Circulation. 2019;140(2):e125-e151. doi:10.1161/CIR.0000000000000665.

5. Amara W, Larsen TB, Sciaraffia E, Hernández Madrid A, Chen J, Estner H, et al. Patients' attitude and knowledge about oral anticoagulation therapy: results of a self-assessment survey in patients with atrial fibrillation conducted by the European Heart Rhythm Association. Europace. 2016;18(1):151-155. doi:10.1093/europace/euv317.

6. Sarraju A, Bruemmer D, Van Iterson E, Cho L, Rodriguez F, Laffin L. Appropriateness of Cardiovascular Disease Prevention Recommendations Obtained From a Popular Online Chat-Based Artificial Intelligence Model. JAMA. 2023;329(10):842-844. doi:10.1001/jama.2023.1044.

7. 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):589-596. doi:10.1001/jamainternmed.2023.1838.

8. Goodman RS, Patrinely JR, Stone CA Jr, Zimmerman E, Donald RR, Chang SS, et al. Accuracy and Reliability of Chatbot Responses to Physician Questions. JAMA Netw Open. 2023;6(10):e2336483. doi:10.1001/jamanetworkopen.2023.36483.

9. Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, et al. Large language models encode clinical knowledge. Nature. 2023;620(7972):172-180. doi:10.1038/s41586-023-06291-2.

10. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. Nat Med. 2023;29(8):1930-1940. doi:10.1038/s41591-023-02448-8.

11. Caranfa JT, Bommakanti NK, Young BK, Zhao PY. Accuracy of Vitreoretinal Disease Information From an Artificial Intelligence Chatbot. JAMA Ophthalmol. 2023;141(9):906-907. doi:10.1001/jamaophthalmol.2023.3314.

12. Hua HU, Kaakour AH, Rachitskaya A, Srivastava S, Sharma S, Mammo DA. Evaluation and Comparison of Ophthalmic Scientific Abstracts and References by Current Artificial Intelligence Chatbots. JAMA Ophthalmol. 2023;141(9):819-824. doi:10.1001/jamaophthalmol.2023.3119.

Table 1. Eligibility framework for the evidence synthesis

Domain	Included evidence	Excluded evidence
Publication period	Published on or before 31 December 2023	2024 onward
Journal quality	Q1 journal in a relevant clinical, digital-health, or AI category	Lower-quartile or unranked journals
Chatbot evidence	Open-ended medical or patient question answering with expert/human evaluation	Purely non-medical benchmark testing
Anticoagulation evidence	Guidelines and patient-knowledge evidence defining safe oral-anticoagulant counseling	Antiplatelet-only information without anticoagulant relevance
Outcomes	Accuracy, appropriateness, completeness, harm, helpfulness, consistency, readability	Engagement metrics without information quality
Cross-platform inference	Only when the same task and evaluation permit comparison	Ranking models evaluated on different datasets

Table 2. Selected 2023 Q1 evidence on large-language-model medical question answering

Study	Task	Key result	Implication for anticoagulant counseling
Sarraju et al. [6]	25 cardiovascular prevention questions	21/25 (84%) response sets appropriate	Good cardiovascular knowledge does not eliminate clinically meaningful errors
Ayers et al. [7]	195 public patient questions	Chatbot preferred in 78.6%; 78.5% good/very good quality	Conversational quality can be high but is not equivalent to medication safety
Goodman et al. [8]	284 physician-generated medical questions	Median accuracy 5.5/6; completeness 3/3	Strong average accuracy with residual low-scoring answers and version effects
Singhal et al. [9]	Consumer medical question answering	92.6% aligned with scientific consensus; 5.9% potentially harmful	Safety requires explicit harm assessment, not accuracy alone
Caranfa et al. [11]	Vitreoretinal patient questions	Incorrect and potentially harmful specialty advice; material response instability	Specialty complexity can expose failures hidden by broad benchmarks
Hua et al. [12]	Scientific answers with references	About 29–33% of generated references were nonverifiable	Source claims require independent verification

Table 3. Core patient-question domains for anticoagulant chatbot evaluation

Question domain	Patient-facing example	Elements expected in a safe answer	Risk if wrong or omitted
Indication	Why am I taking this blood thinner?	Explain thromboembolism prevention/treatment and avoid implying that the drug dissolves established clots unless contextually accurate	Non-adherence or false expectations
Missed dose	I forgot my apixaban/rivaroxaban/warfarin dose. What should I do?	Drug-specific timing advice, no double dosing unless explicitly appropriate, escalation when uncertainty exists	Bleeding or loss of protection
Bleeding	I am bleeding while taking an anticoagulant. Is it dangerous?	Differentiate minor from urgent/major bleeding and direct emergency assessment for severe features	Delayed emergency care
Drug interactions	Can I take ibuprofen, aspirin, antibiotics, or herbal products?	Explain interaction potential and recommend pharmacist/clinician verification before starting interacting agents	Bleeding or reduced anticoagulant effect
Procedures	Should I stop my anticoagulant before surgery or dental treatment?	Do not advise unsupervised interruption; explain that timing depends on drug, renal function, bleeding risk, and procedure	Peri-procedural bleeding or thrombosis
Monitoring	Do I need blood tests?	Distinguish INR monitoring for warfarin from absence of routine coagulation monitoring for DOACs while retaining renal/clinical follow-up	Inappropriate monitoring or false reassurance
Diet	Do I have to avoid green vegetables?	Differentiate stable vitamin K intake with warfarin from DOAC dietary considerations	Unnecessary restriction or unstable INR
Renal function	Can I take the same dose if my kidneys are weak?	State that some DOAC doses/eligibility depend on renal function and require clinician review	Drug accumulation or underdosing
Pregnancy	Can I continue this medicine if I am pregnant?	Avoid generic continuation advice and direct urgent prescriber/obstetric review because anticoagulant choice is context dependent	Fetal or maternal harm
Adherence	Can I take it only when I feel I need it?	Explain continuous adherence and short duration of effect for DOACs; do not stop without prescriber advice	Stroke or recurrent VTE

Table 4. Proposed scoring system for a future cross-platform anticoagulant chatbot study

Domain	Suggested scale	Operational definition
Accuracy	1–5	Agreement with drug-specific guideline/reference-standard facts
Completeness	1–5	Coverage of all clinically necessary elements for the question
Potential harm	1–5	Likelihood that following the answer could cause bleeding, thrombosis, delayed care, or inappropriate interruption
Actionability	1–5	Whether the patient can identify what to do next without unsafe ambiguity
Readability	Objective + 1–5	Readability index plus clinician/language assessment of patient comprehension burden
Uncertainty calibration	1–5	Appropriate acknowledgement of individualized factors and limits of generic advice
Consistency	0–100%	Agreement of key clinical instructions across repeated generations
Source integrity	0–100%	Proportion of cited sources or claims that are verifiable and genuinely supportive
Overall suitability	1–5	Whether response can be released unchanged, requires editing, or should be withheld

Table 5. Anticoagulant chatbot failure modes and clinical consequences

Failure mode	Example	Likely consequence	Required safeguard
Wrong drug-specific instruction	Treating all DOAC missed doses identically	Over-anticoagulation or reduced protection	Drug-specific reference standard
Unsafe interruption advice	Telling patient to stop therapy before procedure without prescriber input	Stroke or recurrent VTE	Mandatory escalation language
Bleeding under-triage	Reassuring severe bleeding as expected	Delayed emergency treatment	Red-flag scoring and harm review
Interaction omission	Ignoring NSAID, antiplatelet, or interacting drug use	Higher bleeding risk or altered exposure	Medication-interaction checklist
Monitoring confusion	Recommending INR-based dose adjustment for a DOAC	Inappropriate self-management	Warfarin/DOAC distinction
Renal-function omission	Providing a fixed DOAC dose irrespective of kidney function	Accumulation or ineffective dosing	Context-dependent dosing rule
Hallucinated source	Citing a convincing but nonexistent guideline	False authority and user trust	Independent citation verification
Response instability	Different safety advice across repeated prompts	Unreliable patient behavior	Repeated-run consistency testing
Overconfident language	Presenting individualized decisions as universal	Unsafe autonomous action	Uncertainty-calibration score

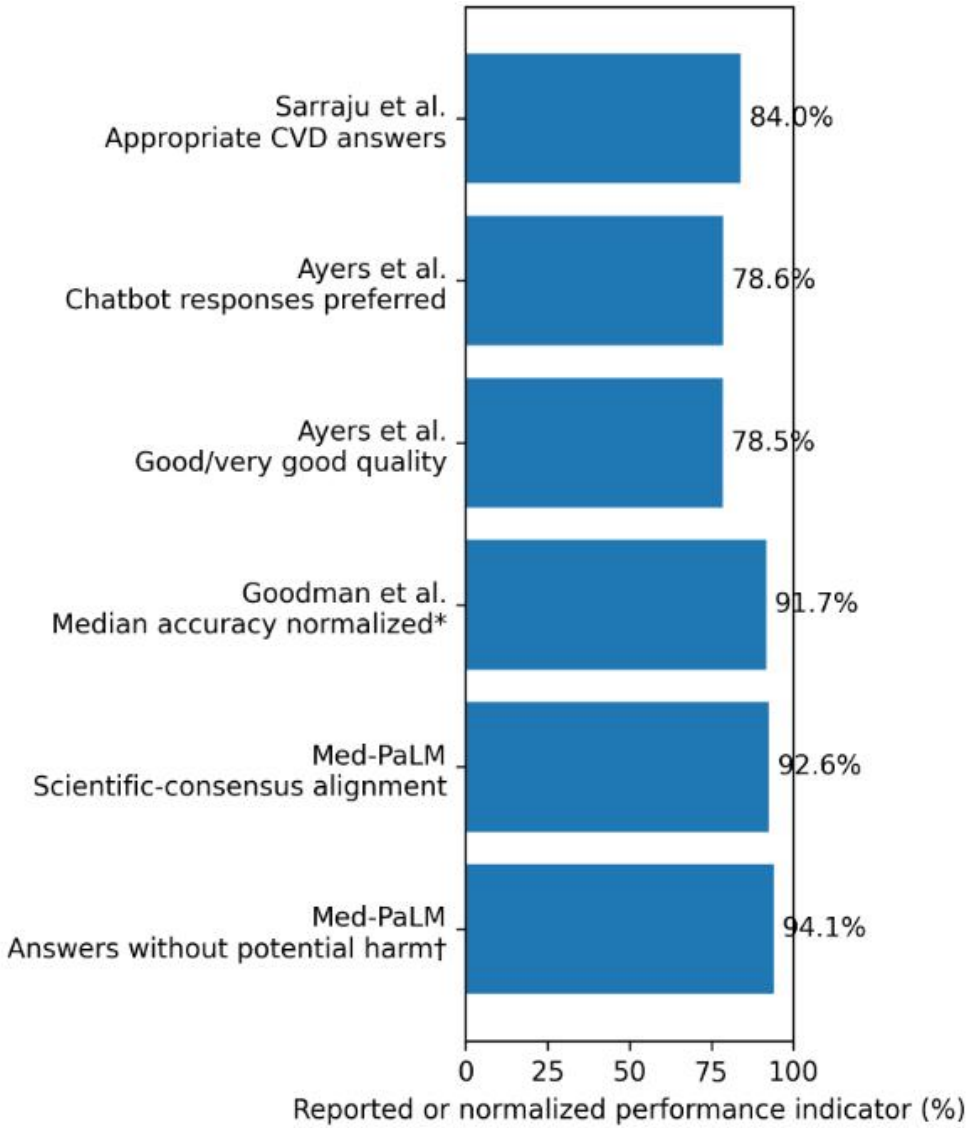
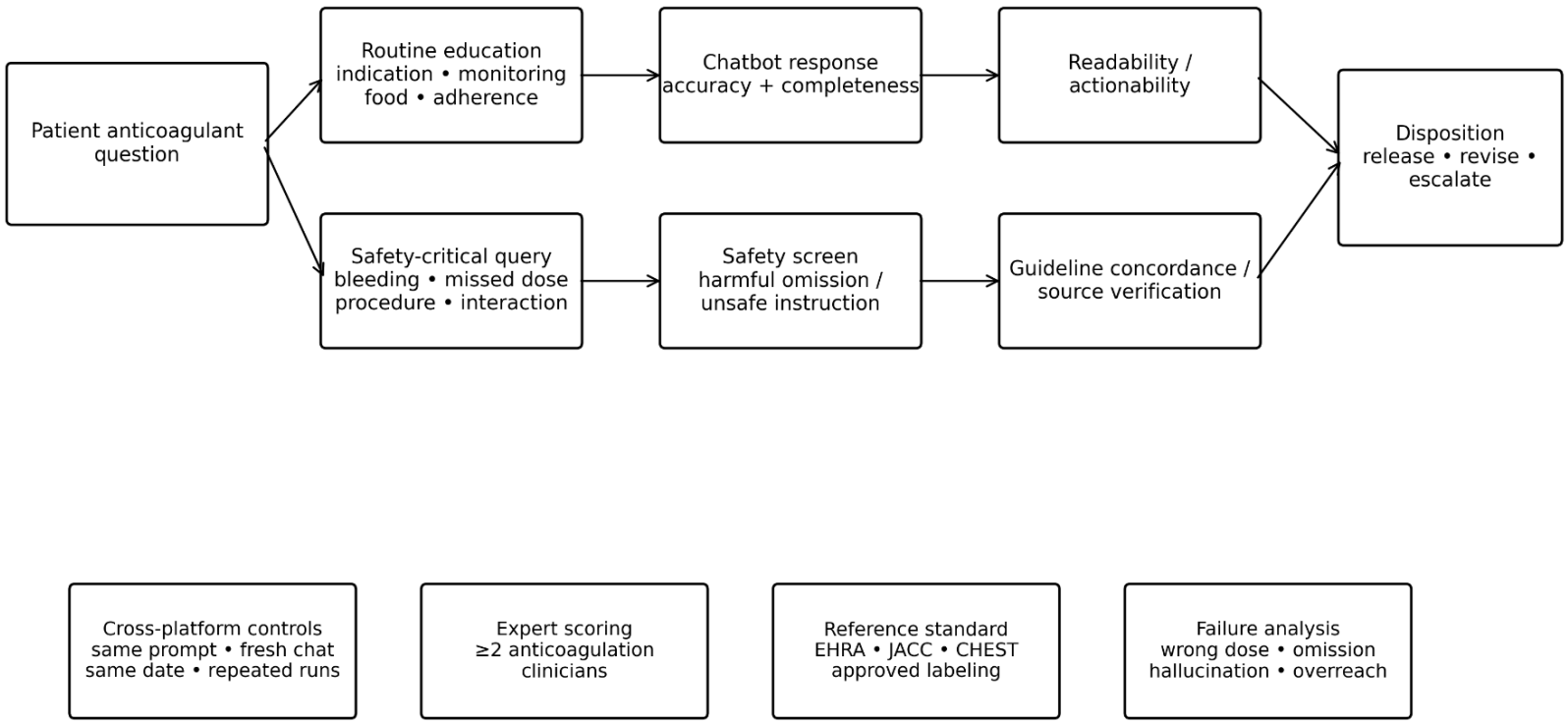


Figure 1. Selected 2023 benchmarks for large-language-model medical question answering.

The displayed indicators arise from different studies and are descriptive rather than statistically pooled. Goodman et al. median accuracy of 5.5/6 is shown as 91.7% only for visualization, while the Med-PaLM no-potential-harm indicator is derived as 100% minus the 5.9% of answers judged potentially harmful.

Figure 2. Proposed cross-platform safety framework for evaluating chatbot answers to anticoagulant patient questions



Anticoagulant information is high stakes: a fluent answer should not be considered patient-ready unless it is accurate, complete, guideline-concordant, and safely actionable.

Figure 2. Proposed cross-platform safety framework for evaluating chatbot answers to anticoagulant patient questions.

Anticoagulant information is high stakes. Patient-facing responses should be evaluated separately for factual accuracy, completeness, readability, actionability, guideline concordance, source integrity, and potential harm before release, with repeated cross-platform testing and anticoagulation-expert review.

How to cite: Foster D. **(2024). Accuracy of Artificial Intelligence Chatbots in Responding to Patient Questions About Anticoagulant Therapy: A Systematic Comparative Evidence Synthesis and Cross-Platform Evaluation Framework. Journal of Multidisciplinary Research and Integrated Sciences, 4(1),**