Beyond the Model: Deterministic Safety Architecture for LLM-Based Medical Consultation

Lilly Schade
Pymander Health

Abstract

Background: General-purpose frontier language models produce fluent medical conversation but exhibit unreliable safety behavior at the decision boundary: the same case can be escalated or dismissed across runs, and correctness depends on phrasing luck rather than on a defensible decision procedure.
Objective: We describe and evaluate a production architecture in which a frontier model proposes clinical content and a deterministic pipeline disposes of it - accepting, rejecting, or correcting every load-bearing decision outside the model.
Methods: The system is a staged pipeline around off-the-shelf frontier models (Claude Opus for reasoning and drafting, Claude Haiku for extraction) with no fine-tuning. Stages include quote-verified fact extraction, deterministic routing, danger-ranked differential generation with separating criteria, a deterministic falsification pass against the leading candidate, an independent danger-vote pass that adjudicates care tier, tightly framed drafting, and a deterministic post-draft audit stack with redraft and a deterministic floor fallback. Every turn emits a structured SOAP visit note and a full decision trace.
Results: On a 300-case red-flag evaluation suite, the final gate run scored safety 1.0 on a 60-case sample under a SHIP/NO-SHIP verdict protocol. A targeted probe of previously flip-prone cases (rabies bat exposure, airway burn, postpartum deep vein thrombosis) produced 15/15 correct escalations, compared with 3/5, 3/5, and 2/5 respectively before the independent danger vote was introduced. A 40-sample false-reassurance probe found zero under-escalation. The known residual failure mode is calm-side over-escalation, which fails in the safe direction.
Conclusions: Safety-critical behavior in conversational medical AI can be made deterministic and auditable by relocating it from the model into the surrounding pipeline. The architecture is model-agnostic by design.

1 Introduction

Conversational medical AI systems are increasingly the first point of contact between patients and the healthcare system. The dominant design pattern - a single large language model prompted to behave like a clinician - inherits the model's strengths (fluency, breadth, availability) and its weaknesses (stochasticity at decision boundaries, unverifiable provenance of stated facts, and no durable record of why a given disposition was reached) [1][2].

The failure mode that matters most is not factual error in the large; it is inconsistency in the small. A patient describing a bat exposure may be told to seek emergency rabies prophylaxis on one run and reassured on another. A postpartum patient with unilateral leg swelling may be routed to same-day evaluation or to watchful waiting depending on phrasing. Such flip-prone cases are precisely the cases where a wrong answer carries the highest danger-if-missed.

We take the position that these behaviors should not be properties of a model at all. This paper presents the architecture deployed in Pymander's AI doctor, a messaging-native medical consultation service. Its organizing principle: the model proposes, the pipeline disposes. The model generates candidate facts, hypotheses, and prose; deterministic machinery verifies, ranks, stress-tests, adjudicates, and audits every output before it reaches the member.

Contributions. (1) A verification architecture in which extracted clinical facts are accepted only with verbatim supporting quotes from the member's own words. (2) A deterministic falsification pass that attempts to break the leading diagnostic candidate before any conclusion is communicated. (3) An independent danger-vote pass that adjudicates care tier separately from the reasoning chain that produced the differential. (4) A post-draft audit stack with redraft and a deterministic floor fallback, together with a complete per-turn decision trace that makes every shipped conclusion auditable end to end.

The architecture is model-agnostic: the underlying models are off-the-shelf and unmodified, and each stage's contract is defined over text, not over any model's internals.

2 System overview and delivery model

The service is delivered natively over SMS and iMessage. There is no application to install and no account to create for a first consultation; a member begins by sending a mes-

sage. Care notes produced during a consultation are viewable without an account via a signed link. This delivery model removes enrollment friction, but it also raises the bar for the reasoning pipeline: sessions are asynchronous, messages are short and colloquial, and the system cannot rely on structured intake forms to constrain the member's input. The pipeline described below is designed for this unstructured, message-at-a-time reality.

3 System architecture

The pipeline wraps off-the-shelf frontier models - Claude Opus for reasoning and drafting, Claude Haiku for extraction - with no fine-tuning. Each stage has a narrow contract. Deterministic stages are implemented as ordinary program logic; model stages are tightly framed so their output is structured and checkable.

3.1 Chart assembly

Incoming messages are assembled into a running clinical chart for the member: the conversation to date, prior extracted facts, prior visit notes, and relevant history. All downstream stages read from the chart rather than from raw conversation, so every model and every check operates on the same state.

3.2 Quote-verified extraction

A Haiku-based extraction stage converts the member's messages into structured clinical facts. A fact is accepted into the chart only if it is accompanied by a verbatim quote of the member's own words supporting it. Facts that cannot be grounded in a quote are rejected. This closes the most common provenance failure in conversational medical AI: the system acting on a fact the patient never stated. Because acceptance is quote-conditioned, every fact in the chart is auditable back to source text.

3.3 Deterministic routing

Deterministic routing classifies each turn and dispatches it to the appropriate downstream path. Routing decisions are program logic over the chart, not model judgment, so the same chart always takes the same path.

3.4 Differential generation with danger ranking

An Opus-based reasoning stage generates a differential diagnosis over the verified chart. Candidates are ranked by danger-if-missed rather than by estimated prevalence, and each candidate carries separating criteria: the specific findings that would distinguish it from its neighbors. Ranking by danger ensures that rare-but-lethal hypotheses are evaluated before common-and-benign ones, inverting the base-rate bias of an unconstrained model [2].

3.5 Deterministic falsification pass

Before any conclusion can be communicated, a deterministic falsification pass attempts to break the leading candidate against the chart. If the candidate survives falsification, the conclusion is eligible to ship. If it does not, the member receives the single settling question - the one question whose answer discriminates between the surviving hypotheses, derived from the separating criteria. The system therefore asks at most one question per turn, and that question is the maximally informative one, not the next item on a checklist.

3.6 Independent danger-vote pass

Care tier - how urgently the member needs what level of care - is decided by an independent danger-vote pass, structurally separated from the reasoning chain that produced the differential. The danger vote sees the verified chart and the candidate set and votes on danger independently; it does not inherit the leading candidate's framing. This independence matters because the failure mode it guards against is correlated error: a reasoning chain that has talked itself into a benign framing will otherwise carry that framing into the triage decision. Section 5 shows that introducing the danger vote converted several flip-prone case categories from majority-wrong to unanimously correct.

3.7 Tightly framed drafting

Only after a conclusion has survived falsification and a care tier has been independently set does an Opus-based drafting stage compose the member-facing message. The drafting frame is tight: it receives the settled conclusion, the tier, and the verified facts, and its job is clarity and tone, not clinical judgment.

3.8 Deterministic post-draft audit stack

Every draft passes through a deterministic audit stack before sending. The stack checks tier routing (does the message match the adjudicated care tier?), emergency-language rules, pediatric scope, voice, and a repeat-question guard (the system must not re-ask a question the member has already answered). A draft that fails any check is redrafted; if redrafting cannot produce a compliant message, the system falls back to a deterministic floor - a fixed, pre-approved safe response for that tier. A non-compliant message therefore cannot ship, by construction.

3.9 Visit notes and decision trace

Each turn produces a structured SOAP visit note (Subjective, Objective, Assessment, Plan) [3] and a full per-turn decision trace: which facts were extracted with which quotes, which candidates were generated and how they were ranked, what falsification did, how the danger vote decided, what the auditors checked, and why the final message shipped. The trace is the audit surface: any shipped conclusion can be replayed and interrogated without rerunning any model.

4 Evaluation

4.1 Evaluation design

We evaluate safety behavior, not conversational quality. The primary instrument is a 300-case red-flag suite: synthetic member presentations in which the correct disposition involves escalation to urgent or emergency care, spanning categories chosen for their danger-if-missed profile. The final gate for shipping pipeline changes is a verdict protocol over a 60-case sample of the suite, with each run producing a SHIP or NO-SHIP verdict and a safety score. Because individual model calls are stochastic, flip-prone cases are additionally probed with repeated sampling: a case category is considered fixed only if it escalates correctly across repeated runs, not once.

4.2 Probes

Two targeted probes complement the gate run. First, a repeated-sampling probe over previously flip-prone case categories: rabies bat exposure, airway burn, and postpartum deep vein thrombosis. Each category was sampled five times before and after the introduction of the independent danger vote. Second, a 40-sample false-reassurance probe, constructed to measure the opposite failure direction: presentations engineered to tempt the system into a reassuring message when escalation is warranted. The primary metric for both probes is under-escalation, the highest-cost error direction of triage [4].

5 Results

Gate run. The final gate run over the 60-case sample of the 300-case red-flag suite scored safety 1.0 under the SHIP/NO-SHIP verdict protocol; the run shipped.

Flip-prone categories. Before the independent danger vote, repeated sampling of the three flip-prone categories produced correct escalations in 3/5 runs (rabies bat exposure), 3/5 runs (airway burn), and 2/5 runs (postpartum DVT). After the danger vote, the same probe produced 15/15 correct escalations across the three categories (Table 1).

Case category	Before vote	After vote
Rabies bat exposure	3/5	5/5
Airway burn	3/5	5/5
Postpartum DVT	2/5	5/5
Total	**8/15**	**15/15**

Table 1: Repeated-sampling probe of flip-prone red-flag categories, before and after the independent danger vote.

False-reassurance probe. The 40-sample probe found zero instances of under-escalation.

Residual behavior. The system's remaining errors are calm-side over-escalations: cases in which the pipeline routes a member to a higher level of care than the presentation strictly requires. We observed no cases of under-escalation in the evaluation above. Over-escalation is the fail-safe direction, but it is not free; we discuss its costs in Section 6.

6 Limitations

Over-escalation. The known residual failure mode is over-escalation on the calm side of the spectrum. This fails safely for the individual member but imposes real costs: unnecessary emergency-department burden, member anxiety, and erosion of trust if every conversation ends in a directive to seek urgent care. Reducing over-escalation without reintroducing under-escalation is the central open problem for this architecture, and we do not claim to have solved it.

Evaluation scope. The evaluation reported here is an internal engineering gate, not a clinical trial. Case suites are synthetic; the 60-case gate sample is a subset of the full 300-case suite; and repeated-sampling probes cover three flip-prone categories rather than the full presentation space. A safety score of 1.0 on a sample is evidence about the sample, not a guarantee about deployment. No external validation, no clinician-blinded comparison, and no prospective patient-outcome measurement is reported.

Model dependence. Although the architecture is model-agnostic in design, the results above were measured with specific off-the-shelf models (Claude Opus and Claude Haiku). Substituting other models changes the proposal distribution even when the disposal machinery is unchanged; results should be re-measured per model generation.

Population and channel. The system operates over SMS and iMessage on self-reported, unstructured member text. It does not examine patients, access records beyond what the member provides, or handle presentations that require physical findings. Pediatric scope is constrained by explicit rules rather than by pediatric-specific training.

7 Discussion

The results support the central claim: behavior that is safety-critical and previously stochastic - escalate or reassure - can be made deterministic and auditable by relocating it from the model into the pipeline. The three mechanisms that appear to carry the most weight are falsification, which prevents a coherent-but-wrong leading candidate from shipping unchallenged; the independent danger vote, which decorrelates the triage decision from the reasoning chain's framing; and the audit stack with a deterministic floor, which bounds the worst-case output regardless of model behavior.

The same clinical core has also been measured against a public examination benchmark with an official answer key. In a pre-registered evaluation on USMLE-style sample questions, the base model scored 95.3% (n=150) and the full reasoning layer scored 98.9% (93/94) on a sealed held-out set that was frozen before any tuning; the companion working paper reports the protocol and error analysis in full. The two evaluations converge on the same residual failure mode - under-escalation on management and level-of-care decisions - which is the failure this architecture's safety machinery is built to contain.

The framing of "the model proposes, the pipeline disposes" is also a statement about where engineering effort should go. Improving the model improves the quality of proposals - better differentials, better questions, better prose - and the architecture benefits. But the safety properties reported here do not depend on any particular model's disposition, and that separation is what makes the per-turn decision trace meaningful as an audit artifact: the pipeline's decisions are legible because the pipeline made them.

Open directions include quantifying and reducing calm-side over-escalation, external and clinician-blinded validation, extending the probe methodology across the full presentation space, and measuring member-level outcomes prospectively.

8 Conclusion

We presented a production conversational medical AI architecture in which off-the-shelf frontier models propose and a deterministic pipeline disposes. On an internal 300-case red-flag suite, the system achieved a safety score of 1.0 on the 60-case gate sample, converted three flip-prone case categories from majority-incorrect to 15/15 correct escalations after the introduction of an independent danger vote, and showed zero under-escalation on a 40-sample false-reassurance probe. The principal residual failure mode is over-escalation, which fails safe and remains the focus of ongoing work.

9 Appendix A. Pipeline stage summary

Stage	Type	Contract
Chart assembly	Deterministic	Single shared clinical state for the turn
Quote-verified extraction	Model (Haiku) + deterministic gate	Facts accepted only with verbatim member quotes
Routing	Deterministic	Same chart, same path
Differential generation	Model (Opus)	Candidates ranked by danger-if-missed, with separating criteria
Falsification pass	Deterministic	Conclusion ships only if it survives; otherwise one settling question
Danger vote	Independent pass	Care tier adjudicated separately from the reasoning chain
Drafting	Model (Opus)	Tight frame: settled conclusion, tier, verified facts
Audit stack	Deterministic	Tier routing, emergency language, pediatric scope, voice, repeat-question guard; redraft, then deterministic floor
Notes and trace	Deterministic	SOAP visit note and full per-turn decision trace

10 References

[1] Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge. Nature 2023;620:172-180. https://www.nature.com/articles/s41586-023-06291-2 (https://www.nature.com/articles/s41586-023-06291-2)

[2] Nori H, King N, McKinney SM, Carignan D, Horvitz E. Capabilities of GPT-4 on medical challenge problems. arXiv:2303.13375 (2023). https://arxiv.org/abs/2303.13375 (https://arxiv.org/abs/2303.13375)

[3] Weed LL. Medical records that guide and teach. New England Journal of Medicine 1968;278:593-600. https://www.nejm.org/doi/full/10.1056/NEJM196803142781105 (https://www.nejm.org/doi/full/10.1056/NEJM196803142781105)

[4] Safety-netting communication during telephone consultations: an observational study using

recorded consultations. British Journal of General Practice (2025). https://doi.org/10.3399/bjgp.2025.0637 (https://doi.org/10.3399/bjgp.2025.0637)

[5] Jin D, Pan E, Oufattole N, Weng WH, Fang H, Szolovits P. What disease does this patient have? A large-scale open-domain question answering dataset from medical exams. Applied Sciences 2021;11(14):6421. https://arxiv.org/abs/2009.13081 (https://arxiv.org/abs/2009.13081)