

Standard guardrails impose a safety-critical omission tax on clinical large language models that a downstream verification layer reverses: a retrospective multi-cohort evaluation

Sanjay Basu

san.jay.basu@ucsf.edu

University of California, San Francisco

Sadiq Patel

University of Pennsylvania

Parth Sheth

University of Pennsylvania

John Morgan

Virginia Commonwealth University

Rajaie Batniji

Waymark

Article

Keywords:

Posted Date: June 28th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10153003/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License. [Read Full License](#)

Additional Declarations: Competing interest reported. The authors are employees of Waymark, a public benefit organization that provides free social and medical services for patients receiving Medicaid.

Abstract

Large language models used in clinical decision support make two kinds of error: commission, the confident assertion of unsupported information, and omission, the failure to mention safety-critical care. Because hallucination is structural and does not disappear as models become more accurate, health systems add guardrails, which are usually assessed for commission alone. In a retrospective, pre-specified evaluation of three frontier large language models (GPT-5.5, Gemini 3.1 Pro, Claude Opus 4.7) on a contamination-excluded corpus of 1,363 clinical cases, a standard four-component guardrail stack lowered commission but raised safety-critical omission by 12.5 percentage points across 921 paired case-runs (17.6% to 30.1%). A downstream verification layer that modified the response before it was returned lowered omission to the bare-model level (paired reduction 16.46 percentage points; Newcombe 95% confidence interval 10.93 to 21.86; McNemar exact $P = 1.32 \times 10^{-23}$), detected 49.0% of the omissions the guardrails missed, and missed none that the guardrails caught (per-model 39.8% to 73.3%; $P < 10^{-4}$). The same layer added only as advisory text raised commission (paired difference 4.44 percentage points; $P = 0.0015$). Whether both errors fell together depended on where the verification layer acted, not on which model was used.

Introduction

Large language models embedded in clinical workflows for documentation, in-basket triage, and decision support produce at least two widely recognized categories of safety failure^{1–3}. Commission errors, commonly called hallucinations, are confident assertions of clinical information not supported by the evidence base. Safety-critical omission errors are recommendations that fail to mention indicated workup, contraindications, escalation, or higher-acuity differentials.

Representative clinical examples include co-prescribing trimethoprim-sulfamethoxazole in a warfarin-treated patient, which is associated with a roughly four-fold increase in haemorrhage-related hospitalisation⁴; recommending ibuprofen for a patient whose creatinine and estimated glomerular filtration rate are stated mid-paragraph, a lost-in-the-middle long-context failure⁵; and affirming a fabricated guideline citation with a fabricated class-of-recommendation grade⁶. Hallucinated or mis-attributed citations have been reported in 50–90% of medical large language model responses⁶, drug-dose and interaction errors in 20–40% of adversarial probes^{7,8}, and most evaluated frontier large language models are more likely to hallucinate than to answer correctly on broad factual-knowledge benchmarks⁹.

To mitigate these failure modes, health systems operationalise overlapping guardrail components, typically a clinical-safety system prompt², a content-moderation filter^{10–12}, retrieval-augmented generation with citation grounding⁶, and a learned hazard classifier, following common coalition frameworks^{13–15}. The empirical effect of these components on the safety-critical omission rate is not well characterized. Physician training may not substitute: a randomised trial showed a 14-percentage-point reduction in physician diagnostic accuracy after exposure to erroneous large language model recommendations despite completion of a 20-hour artificial-intelligence-literacy programme¹⁶.

Formal results from learning theory¹⁷, autoregressive sampling structure¹⁸, and reward- and auction-theoretic analyses^{19,20} indicate that hallucination is a structural property of transformer-based models: it cannot be removed by any guardrail placed on top of the model, and it does not disappear as models become more accurate, because safety and accuracy follow different scaling laws in clinical large language models²¹. A safety property that must hold on every output therefore has to come from a layer outside the model. Concept-decomposition frameworks^{22,23}, which can incorporate deterministic Logical Neural Network certificates²⁴ over an expert-curated clinical-rule library, have been proposed to prevent both commission and safety-critical omission downstream of the generator. Whether such a downstream verification layer simultaneously reduces both error modes in clinical use has not been empirically tested. We evaluated one such framework, comprising a Logical Neural Network certificate, six specialist verification agents, and a concept-decomposition module that emits a residual-attribution-gap statement on every output (Fig. 1), in a three-stage paired comparison (bare model; standard guardrails; standard guardrails plus the verification framework) across three frontier model families.

Results

Cohort Enrolment and Analytic Samples

All cohorts described in the Methods were fully enrolled, and no case was excluded for missing data. Cohort 1a contributed 921 paired case-runs, of which 468 were evaluable on the comprehensiveness axis after the pre-specified 50% domain-scoring sampling; Cohort 1b contributed 1,200 paired case-runs; Cohort 2 contributed 1,080 cells across four documented commission failure-mode classes; Cohort 3 contributed 1,080 cells across six guardrail-component configurations; and Cohort 4 contributed 150 stratified cases. The three supplementary cohorts (adversarial, $n = 50$; fabricated-medical-condition, $n = 189$ cells; specialist-agent leave-one-out, $n = 200$) were fully enrolled (Supplementary Table 1).

Primary Endpoint: Paired Safety-Critical Omission

The bare large language model omitted at least one safety-critical item in 17.6% of case-runs on Cohort 1a (Wilson 95% confidence interval 14.7 to 20.9). Adding the standard guardrail stack increased the safety-critical omission rate to 30.1% (95% confidence interval 26.6 to 33.9), a safety-critical omission tax of approximately 12.5 percentage points. Adding the verification framework returned the omission rate to 17.1% (95% confidence interval 14.0 to 20.8). The paired absolute reduction in safety-critical omission between the guardrail-only and verification-framework configurations was 16.46 percentage points (Newcombe 95% confidence interval 10.93 to 21.86; paired McNemar exact $P = 1.32 \times 10^{-23}$). The verification framework returned safety-critical omission to a rate not statistically distinguishable from the bare large language model on the same cases (Fig. 2a).

Catch Rate and Reverse-Miss Rate

On the 468 paired-evaluable case-runs of Cohort 1a, the verification framework caught 77 of 157 safety-critical omissions missed by the standard guardrail stack (pooled catch rate 49.04%; Wilson 95% confidence interval 41.3 to 56.8). Per-model catch rates were 73.3% for Claude Opus 4.7 (paired difference 14.10 percentage points; $P = 4.77 \times 10^{-7}$), 55.2% for GPT-5.5 (paired difference 10.26 percentage points; $P = 3.05 \times 10^{-5}$), and 39.8% for Gemini 3.1 Pro (paired difference 25.00 percentage points; $P = 3.64 \times 10^{-12}$); the larger absolute reduction on Gemini reflects its higher baseline omission rate under standard guardrails (Table 1; Supplementary Table 4). No safety-critical omission caught by the standard guardrail stack was missed by the verification framework on any model, so the framework’s behaviour on the catch axis is strictly additive over the guardrail stack.

Table 1

Per-Model Diagnostic-Test Characterisation on the Primary Safety-Critical Omission Endpoint Cohort 1a, n = 156 paired-evaluable per model and n = 468 pooled. Per-model paired sensitivity on the comprehensiveness axis, paired absolute reduction (verification framework minus standard guardrails) in percentage points, catch rate of safety-critical omissions missed by the standard guardrail stack, and paired McNemar exact P. The paired absolute reduction is positive across all three models and across the pooled cohort, with per-model 95% confidence intervals (Supplementary Table 4) not overlapping zero.

Model	Standard guardrails sensitivity (%)	Verification framework sensitivity (%)	Paired difference (pp)	Catch rate of guardrail-missed (%)	McNemar exact P
GPT-5.5	81.41	91.67	+ 10.26	55.17	3.05×10^{-5}
Gemini 3.1 Pro	37.18	62.18	+ 25.00	39.80	3.64×10^{-12}
Claude Opus 4.7	80.77	94.87	+ 14.10	73.33	4.77×10^{-7}
Pooled (Cohort 1a)	66.45	82.91	+ 16.46	49.04	1.32×10^{-23}

Per-Model Generalisability

Per-model 95% confidence intervals on the catch rate did not overlap zero across all three model families (Supplementary Table 4), supporting cross-vendor generalisability.

Triage Safety Sensitivity

On Cohort 1b (n = 1,200), the bare large language model identified 9.15% of triage-acuity-trigger cases (Wilson 95% confidence interval 6.41 to 12.91), the standard guardrail stack identified 70.59% (65.25 to 75.41), and the verification framework identified 72.22% (66.95 to 76.94). The incremental gain of the verification framework over the standard guardrail stack on this axis was 1.63 percentage points and was not significant ($P > 0.99$ after Benjamini–Hochberg adjustment). The standard guardrail stack therefore carried most of the triage-sensitivity gain, while the verification framework carried most of the comprehensiveness and differential-diagnosis gain.

Per-Axis Safety-Critical Omission

The verification framework reduced safety-critical omission on the overall comprehensiveness axis (paired difference 16.45 percentage points; McNemar exact $P = 1.32 \times 10^{-23}$) and on the differential-diagnosis axis (paired difference 8.12 percentage points; $P = 2.57 \times 10^{-7}$; Benjamini–Hochberg-adjusted $P < 0.01$). Per-axis differences on the workup (1.28 percentage points; $P = 0.15$), treatment (0.64 percentage points; $P = 0.45$), and triage axes (0.21 percentage points; $P > 0.99$) were not significant after Benjamini–Hochberg adjustment (Fig. 3).

Primary Commission Endpoint and Per-Failure-Mode Analysis

Only two of the four documented Cohort 2 commission failure-mode classes, hallucinated citation and wrong treatment recommendation, elicited non-zero commission on the bare large language model, whereas the drug–drug interaction and lost-in-the-middle laboratory-misquote classes elicited 0% commission across all configurations and could not be used to test load-bearing roles (Supplementary Tables 5 and 6). The commission contrasts reported below are therefore effectively powered on the hallucinated-citation and wrong-treatment classes.

On Cohort 3 (per-component leave-one-out), the pooled commission rate fell from 7.8% (bare large language model) to 2.8% (full guardrail stack). Removing the retrieval-augmented-generation citation-grounding component raised the pooled rate to 8.3%, driven by hallucinated-citation commission rising from 6.7% to 26.7% (paired difference + 20.0 percentage points; McNemar exact $P = 0.035$). No other single-component removal showed a load-bearing role on the classes tested.

Advisory-Only Deployment Increased Commission

On Cohort 2, an advisory-only operationalisation of the verification framework, in which a verification advisory was appended to the bare large language model output without altering response generation, did not reduce commission and instead showed a pooled paired absolute increase of 4.44 percentage points (McNemar exact $P = 0.0015$), driven by hallucinated citation (paired difference + 10.0 percentage points; $P = 0.035$) and wrong treatment recommendation (paired difference + 7.78 percentage points; $P = 0.039$). This contrasts with the full framework on Cohort 1a and 1b, which modifies the response-generation pipeline.

Component Ablation

A pre-specified five-arm ablation pooled across models on Cohort 1b ($n = 591$ paired-evaluable case-runs; Supplementary Table 8) characterized each component's contribution to comprehensiveness. The Logical Neural Network certificate alone (paired difference + 3.21 percentage points; $P = 3.81 \times 10^{-6}$) and the specialist-agent panel alone (paired difference + 6.26 percentage points; $P = 1.46 \times 10^{-11}$) each improved comprehensiveness over the standard guardrail stack. The full framework improved on the certificate-only configuration by 4.57 percentage points ($P = 7.43 \times 10^{-6}$) and on the agents-only configuration by 1.52 percentage points ($P = 0.0225$). Both components were individually load-bearing, and the full framework strictly dominated either component alone.

Secondary Analyses

The verification framework emitted a residual-attribution-gap statement on 100% of outputs (Supplementary Note 12). Orthogonal safety dimensions are summarized in Supplementary Fig. 1, including a bare-to-framework adversarial-detection progression of 0.0% to 100.0% on the supplementary adversarial cohort (Supplementary Note 3). The Logical Neural Network certificate's diagnostic-test characteristics on the marginal-value subsample ($n = 374$), the cross-model self-review and three-vendor cross-review analyses, and the large language model judge-ensemble calibration against three blinded board-certified internal-medicine physicians (Cohen's kappa 0.701 to 0.889; pre-specified threshold kappa ≥ 0.70) appear in Supplementary Notes 8, 11, and 12. Three worked clinical cases (Table 2) and seven further cases (Supplementary Table 7) illustrate the stage-to-stage progression across configurations.

Table 2

Three Worked Clinical Cases Drawn from the Retrospective Corpus Each case illustrates a representative failure mode addressed by the verification framework. Cases are paraphrased to remove identifying information. Seven additional worked cases appear in Supplementary Table 7.

Case	Clinical scenario	Bare LLM output	Standard guardrails output	Verification framework output	Specialist agent that flagged	LNN certificate state	Concept-decomposition class
1 — Acute aortic dissection	68-year-old with subacute back pain; anti-coagulant history.	Outpatient analgesics recommended.	In-clinic evaluation recommended; emergency-department referral not mentioned.	Emergency-department referral with dissection on the differential.	Differential-diagnosis breadth and Logical Neural Network anti-coagulant–imaging contraindication	Rule-violation-flagged	T1 (known-traceable)
2 — Methotrexate frequency error	54-year-old newly diagnosed with inflammatory arthritis; methotrexate prescription requested.	Daily methotrexate recommended.	Methotrexate recommended without specifying frequency.	Weekly methotrexate explicitly recommended; daily frequency flagged as hazard.	Drug-interaction (daily-versus-weekly methotrexate is a documented prescribing-cascade hazard)	Rule-violation-flagged	T1 (known-traceable)
3 — Fabricated medical condition	Adversarial query about a fabricated blue-light-related eyelid-discoloration syndrome.	Accepts the fabricated premise and provides advice.	Accepts the premise and produces a disclaimed description.	Refuses to confirm the entity and emits an epsilon-disclosed residual-attribution-gap statement.	Citation grounding (PubMed E-utilities returns zero records)	Unknown (out-of-scope)	T2 (epsilon-disclosed)

Discussion

The four-component guardrail stack reduced commission errors but introduced the 12.5-percentage-point safety-critical omission tax relative to the bare large language model. A downstream verification framework operating by response-pipeline modification reversed the safety-critical omission tax (paired absolute reduction 16.46 percentage points), caught 49.0% of the safety-critical omissions the guardrails failed to detect, left none of the guardrail-detected omissions undetected, and improved specificity on the comprehensiveness axis from 26.4% to 31.6% (Supplementary Table 4). On the highest-acuity dimension, the bare large language model returned an acuity-flagged recommendation in approximately one of every eleven hazard-present cases, and the verification framework in approximately seven of every ten. The standard guardrail stack carried most of the triage gain, and the verification framework carried most of the comprehensiveness and differential-diagnosis gain. We frame these results architecturally rather than at the level of any single model, because the pattern held across three independently developed model families.

Appending text to a model's output is a simpler way to add a safety layer than changing how the response is generated, but the two were not equivalent. The full framework, which modified the response-generation pipeline, reduced safety-critical omission by 16.46 percentage points on Cohorts 1a and 1b. The same framework added only as advisory text increased commission by 4.44 percentage points relative to standard guardrails on Cohort 2 ($P = 0.0015$), driven by hallucinated citation (+ 10.0 percentage points) and wrong treatment recommendation (+ 7.78 percentage points). Advisory text did not retract or override the model's original assertion. These data support modifying the response-generation pipeline, rather than appending advisory text, when adding a verification layer.

Formal impossibility results^{17–20} predict that hallucination cannot be eliminated by any guardrail layered on top of a transformer-based large language model, and consistent with that prediction no intra-model composition tested here reduced commission to zero. Earlier empirical work has shown fine-tuning, retrieval-augmented generation, and other intra-model modifications insufficient on safety-critical content^{6,8}. The safety-critical omission tax was not attributable to any single guardrail component: retrieval-augmented generation with citation grounding was load-bearing against hallucinated-citation commission (Cohort 3 leave-one-out: paired difference + 20.0 percentage points; $P = 0.035$), so removing it would reintroduce commission. The safety-critical omission tax therefore appears to be an emergent property of the four-component composition rather than an artefact of any single component, consistent with auction- and reward-theoretic analyses^{19,20} holding that any per-output safety property must come from a layer external to the generative model.

A pre-specified five-arm ablation pooled across models on Cohort 1b ($n = 591$ paired-evaluable case-runs; Supplementary Table 8) characterized each architectural component's contribution to comprehensiveness. The certificate-only configuration (paired difference + 3.21 percentage points; $P = 3.81 \times 10^{-6}$) and the agents-only configuration (paired difference + 6.26 percentage points; $P = 1.46 \times 10^{-11}$) each improved comprehensiveness over the standard guardrail stack, the full framework improved on the certificate-only configuration by 4.57 percentage points ($P = 7.43 \times 10^{-6}$) and on the agents-only configuration by 1.52 percentage points ($P = 0.0225$). The certificate provides deterministic specificity on rule-resolvable cases, the agents provide sensitivity on rule-residual cases, and both contribute non-redundant signal. Same-model self-review reliability ranged from 6.8% to 69.8% across the three models, and a three-model cross-review committee added zero incremental detections at approximately 4.5-fold per-output cost, supporting an architectural rather than model-level approach to clinical large language model safety. State-aware multimodal reasoning²⁵ and evidence that distinct internal circuits yield identical task performance²⁶ support grounding the audit externally rather than in model internals.

This study has several limitations. Architectural findings, namely the safety-critical omission tax, the operative role of response-pipeline modification, and the strictly additive catch behaviour, reflect deployment architecture and are likely to transfer across clinical settings, but point estimates such as the + 16.46-percentage-point pooled reduction, the per-model catch rates in Table 1, and the per-axis effects in Fig. 3 require replication on other provider populations and specialties before transferring quantitatively. The de-identified clinical-encounter subset was drawn from one multi-state United States provider organisation of which all authors are employees, with a Medicaid case mix and a single-provider care model. Partial contamination from training-data variants outside the indexed pre-training corpora cannot be fully excluded. The large language model judge ensemble was calibrated against three blinded internal-medicine physicians (kappa 0.701 to 0.889; pre-specified threshold ≥ 0.70), but large language model scorers evaluating large language model outputs remains a methodological compromise. The Cohort 2 probes are intentionally clinically typical rather than adversarial. The evaluated models reflect three frontier model families available at the time of evaluation and may not generalise to other families, to non-English languages, or to specialties outside internal medicine. The residual-attribution coverage on the third concept tier was 0.544, indicating that GPT-5.5 reasoning is not fully decomposable by the current concept template. Finally, the evaluation is retrospective, and commission and safety-critical omission rates are surrogates for the clinically meaningful endpoint of avoided patient harm, which is measurable only prospectively.

A prospective randomized controlled trial comparing the standard guardrail stack with and without the verification framework on a clinically meaningful patient-outcome endpoint is the necessary next step before routine deployment, following SPIRIT-AI 2020²⁷ and CONSORT-AI 2020²⁸ reporting standards. The present retrospective surrogate-endpoint evaluation is the precursor to that registered trial. Architectural rather than model-level approaches can reduce both commission and safety-critical omission when the verification framework operates downstream of the guardrail stack and modifies the response-generation pipeline rather than appending advisory text.

Methods

Study Design and Reporting

This was a retrospective, pre-specified, multi-cohort evaluation of three frontier large language models under three configurations: bare large language model; large language model plus a standard institutional guardrail stack; and large language model plus the standard guardrail stack plus a downstream verification framework. The statistical analysis plan was locked before result inspection. The evaluation followed the TRIPOD-LLM 2025 reporting guideline²⁹ (checklist in Supplementary Note 1). Anthropic's Claude Code (an AI coding assistant) was used to assist with analysis-code development and manuscript drafting; the authors reviewed and verified all code, analyses, and text, confirmed every reported value against the canonical source outputs, and take full responsibility for the work.

Cohorts

The evaluation drew on four main cohorts and three supplementary cohorts. Each cohort is a set of clinical cases, probes, or scenarios assembled to answer one analytic question, not a set of distinct patients. Cohort 1 was the parent retrospective clinical-case corpus with two nested analytic samples, and Cohorts 2, 3, and 4 were independently assembled. The full cohort architecture, source-benchmark composition, and analysis-to-cohort mapping appear in Supplementary Table 1 and Supplementary Note 2.

Cohort 1 was a 1,363-case contamination-excluded clinical corpus assembled from public clinical-question benchmarks³⁰, open-access peer-reviewed case reports (BMJ Case Reports, NEJM Clinical Problem-Solving, Mayo Clinic Proceedings), and de-identified clinical encounters from a multi-state provider organisation³¹. An 8-gram overlap filter³² against the models' reported pre-training corpora and against the OpenWebText, RedPajama, and CommonCrawl indexes excluded any case with an 8-gram match. Two nested analytic samples were drawn. Cohort 1a (hazard-stratified, $n = 307$ cases $\times$ 3 models = 921 paired case-runs) was the primary-endpoint sample, stratified to enrich for Beers 2023 potentially inappropriate medication signals³³ and CMS hospital-acquired condition signals; rule-check stratum assignment agreed with manual adjudication at Cohen's kappa 0.94 on a 50-case subset. Cohort 1b (comprehensive-matrix, $n = 400$ cases per model = 1,200 paired

case-runs) supported per-axis characterisation across triage, differential diagnosis, workup, and treatment. Detailed cohort-assembly criteria appear in Supplementary Note 2.

Cohort 2 ($n = 120 \text{ probes} \times 3 \text{ models} \times 3 \text{ configurations} = 1,080 \text{ cells}$) was the commission-probe cohort, with 30 probes per class across four documented failure-mode classes: drug–drug interaction commission⁴, lost-in-the-middle wrong-laboratory-retrieval commission⁵, wrong treatment recommendation³⁴, and hallucinated citation to a non-existent guideline^{6,35,36}. Each probe was authored by three board-certified internal-medicine physicians with a pre-specified correct answer and known-wrong answer, with 100% inter-physician agreement on the known-wrong answer (Supplementary Note 2). Cohort 3 ($n = 60 \text{ probes} \times 3 \text{ models} \times 6 \text{ configurations} = 1,080 \text{ cells}$) was the guardrail per-component leave-one-out cohort, a subset of Cohort 2 evaluated under bare, full guardrails, and four single-component leave-one-out variants. Cohort 4 ($n = 150$) was the large language model judge-calibration cohort. Three supplementary cohorts supported sensitivity analyses (adversarial, $n = 50$; fabricated-medical-condition, $n = 189 \text{ cells}$; specialist-agent leave-one-out, $n = 200$; Supplementary Notes 3, 4, and 5).

Models and Configurations

Three frontier models were evaluated (GPT-5.5, Gemini 3.1 Pro, Claude Opus 4.7), selected for public availability of a clinical-deployment endpoint at the evaluation lock (29 April 2026) and for representing three independent model families (Supplementary Table 2). Each was queried at temperature 0 with a maximum of 8,000 output tokens via the OpenAI Responses, Google Gemini, and Anthropic Messages application programming interfaces, with the model identifiers and evaluation dates listed in Supplementary Table 2. The bare configuration issued the case prompt directly to the model; the standard-guardrails configuration applied the four-component stack described below; and the verification-framework configuration applied the stack and then routed the model's output through the downstream verification framework.

Standard Guardrail Stack

The standard guardrail stack is the authors' operational synthesis drawing on the Coalition for Health AI assurance-laboratory framework^{14,15} and the Duke Institute for Health Innovation–led Health AI Partnership community of practice¹³. It comprised (i) a clinical-safety system prompt²; (ii) the model provider's default content-moderation pre-screen; (iii) a retrieval-augmented generation layer retrieving up to ten passages from MedlinePlus, ACP Choosing Wisely, USPSTF A and B recommendations, and NIH Office of Dietary Supplements interaction notices, with explicit citation grounding⁶; and (iv) a learned hazard-classifier pre-screen (a frozen DeBERTa-v3-large encoder³⁷ fine-tuned on the MIMIC-IV hazard-event subset³⁸, held-out area under the receiver-operating-characteristic curve 0.81). Content-moderation behaviour followed each provider's published usage policy^{10–12}. The verbatim system prompt, filter configuration, retrieval index, and hazard-classifier training detail appear in Supplementary Note 7.

Verification Framework

The verification framework comprised three components applied to the output of the standard-guardrail-stack configuration (Fig. 1). The first was a deterministic certificate generated by a Logical Neural Network²⁴ over a 3,162-rule expert-curated clinical-rule library (Beers 2023³³, STOPP/START v3³⁹, Lexicomp severity X/D, CMS hospital-acquired conditions, and the ACC/AHA, IDSA, ADA, and ACEP guideline corpora), emitting a three-state certificate (rule-consistent, rule-violation-flagged, unknown). The second was a panel of six concurrent specialist verification agents (drug-interaction, laboratory-result interpretation, guideline compliance, citation grounding⁶, safety-net, and differential-diagnosis breadth³), each emitting a structured JSON header (flagged, named hazard, confidence, evidence quote), with a consensus gate requiring at least two agents to emit the same named hazard before a triage upgrade. The third was a concept-decomposition module^{22,23} that partitions each recommendation into traceable, epsilon-disclosed, and residual concepts and emits a residual-attribution-gap statement on every output. The six-agent count was the minimum covering every documented failure mode and the maximum passing empirical calibration gates; five-agent leave-one-out and nine-agent expansion experiments are reported in Supplementary Notes 5 and 6. The full rule schema, Logical Neural Network training procedure, JSON-header agent specification, consensus logic, and concept-decomposition template appear in Supplementary Note 8.

Outcome Measures

The primary safety endpoint was the paired safety-critical omission rate, defined as a response lacking at least one item from a pre-specified four-axis rubric (workup, triage, differential, treatment), scored as a binary flag per case-run; the full rubric appears in Supplementary Note 9. The primary commission endpoint was the paired commission rate on engaged cells of Cohort 2, where an engaged cell is a model–probe pair producing a clinically substantive response rather than a content-policy refusal (refusal-handling rule and per-model refusal rates in Supplementary Note 10). Secondary outcomes included per-axis safety-critical omission rate, paired triage hazard sensitivity, catch rate of

guardrail-missed safety-critical omissions, reverse-miss rate, and per-component leave-one-out commission rates. The rubric was operationalised by a cross-family large language model judge ensemble (GPT-5.5, Gemini 3.1 Pro, Claude Opus 4.7), calibrated against three blinded board-certified internal-medicine physicians on Cohort 4 against a pre-specified criterion of Cohen's kappa ≥ 0.70 (observed per-axis kappa 0.701 to 0.889); the calibration procedure and per-axis and per-model results appear in Supplementary Note 11.

Statistical Analysis

The primary safety-critical omission endpoint (Cohort 1a; $n = 468$ paired-evaluable case-runs; verification framework versus standard guardrails) was tested by a two-tailed paired McNemar exact test, justified by the paired binary structure of the omission flag on the same case under two configurations, at $\alpha = 0.05$, with a Newcombe 95% confidence interval on the paired difference⁴⁰. The observed paired absolute reduction was 16.46 percentage points ($P = 1.32 \times 10^{-23}$). Per-model paired McNemar exact tests ($n = 156$ each; GPT-5.5 $P = 3.05 \times 10^{-5}$, Gemini 3.1 Pro $P = 3.64 \times 10^{-12}$, Claude Opus 4.7 $P = 4.77 \times 10^{-7}$) were pre-specified. Catch rates were reported with Wilson 95% confidence intervals⁴¹, and triage hazard sensitivity (Cohort 1b; $n = 1,200$) used exact Clopper–Pearson 95% confidence intervals. The four per-axis safety-critical omission comparisons (Cohort 1b; two-tailed paired McNemar exact tests) were corrected for family-wise error by the Benjamini–Hochberg procedure⁴² at a false-discovery rate of 0.05; the comprehensiveness ($P = 1.32 \times 10^{-23}$) and differential-diagnosis ($P = 2.57 \times 10^{-7}$) axes were significant after correction, whereas the workup ($P = 0.15$), treatment ($P = 0.45$), and triage ($P > 0.99$) axes were not. The primary commission endpoint (Cohort 2; advisory-only versus standard guardrails; $n = 360$ engaged pairs) was tested by a two-tailed paired McNemar exact test, justified by the paired binary commission outcome, at $\alpha = 0.05$; the observed pooled paired increase was 4.44 percentage points ($P = 0.0015$), with significant per-class increases for hallucinated citation ($P = 0.035$) and wrong treatment recommendation ($P = 0.039$). The pre-specified five-arm component ablation (Cohort 1b; $n = 591$ paired-evaluable case-runs) used two-tailed paired McNemar exact tests per contrast (certificate-only versus guardrails $P = 3.81 \times 10^{-6}$; agents-only versus guardrails $P = 1.46 \times 10^{-11}$; full versus certificate-only $P = 7.43 \times 10^{-6}$; full versus agents-only $P = 0.0225$). The guardrail per-component leave-one-out tests (Cohort 3) used two-tailed paired McNemar exact tests with Benjamini–Hochberg correction across the four ablations per failure-mode class; only removing retrieval-augmented generation was significant on hallucinated citation ($P = 0.035$). Per-cohort sample sizes provided 0.80 power or greater to detect the pre-specified effects at $\alpha = 0.05$ (Supplementary Table 3). Wilson score intervals were used for proportions⁴¹, and analyses were conducted in Python 3.11 with a canonical aggregator script; every reported number traces to that script's output JSON via the provenance registry deposited with the code.

Ethics

The retrospective corpus was assembled from public benchmarks, open-access peer-reviewed case reports, and de-identified clinical encounters from a multi-state United States provider organisation serving Medicaid-enrolled patients. The work was reviewed and approved by WCG IRB (Protocol 20261776), in accordance with the Declaration of Helsinki, with a waiver of informed consent for the de-identified clinical encounters. Patient-level demographic data are not reported by the public benchmark sources from which the retrospective cohort was assembled.

Declarations

Competing Interests

The authors are employees of Waymark, a public benefit organization that provides free social and medical services for patients receiving Medicaid.

Author Contribution

S.B. conceived and designed the study, curated the data, performed the formal analysis, and drafted the manuscript. S.P., P.S., J.M., and R.B. contributed to study conception, the evaluation framework, clinical interpretation, and critical revision of the manuscript. All authors read and approved the final manuscript.

Acknowledgements

This study received no funding.

Data Availability

Code and public-use data are available at <<https://github.com/sanjaybasu/anchor-public>>. Individual identifiable protected health information is not disclosed under US HIPAA regulations.

References

1. Singhal, K. et al. Large language models encode clinical knowledge. *Nature* **620**, 172–180 (2023).
2. Goh, E. et al. Large language model influence on diagnostic reasoning: a randomized clinical trial. *JAMA Netw. Open.* **7**, e2440969 (2024).
3. Brodeur, P. G. et al. Performance of a large language model on the reasoning tasks of a physician. *Science* **392**, 524–527. 10.1126/science.adz4433 (2026).
4. Fischer, H. D. et al. Hemorrhage during warfarin therapy associated with cotrimoxazole and other urinary tract anti-infective agents: a population-based study. *Arch. Intern. Med.* **170**, 617–621. 10.1001/archinternmed.2010.37 (2010).
5. Liu, N. F. et al. Lost in the middle: how language models use long contexts. *Trans. Assoc. Comput. Linguist.* **12**, 157–173 (2024).
6. Wu, K. et al. An automated framework for assessing how well LLMs cite relevant medical references. *Nat. Commun.* **16**, 3615 (2025).
7. Chen, S. et al. Use of artificial intelligence chatbots for cancer treatment information. *JAMA Oncol.* **9**, 1459–1462 (2023).
8. Henry, K. et al. Large language models management of medications: three performance analyses. Preprint at (2025). <https://arxiv.org/abs/2509.22926>
9. Artificial Analysis. AA-Omniscience: a benchmark of foundation model factual knowledge and hallucination. Artificial Analysis <https://artificialanalysis.ai/omniscience>. (2025).
10. OpenAI & Usage policies. OpenAI (2024). <https://openai.com/policies/usage-policies>
11. Anthropic. Usage policy. Anthropic (2024). <https://www.anthropic.com/legal/aup>
12. Google Generative AI prohibited use policy. Google (2024). <https://policies.google.com/terms/generative-ai/use-policy>
13. Sendak, M. P. et al. Empowering US healthcare delivery organizations: cultivating a community of practice to harness AI and advance health equity. *PLOS Digit. Health.* **3**, e0000513 (2024).
14. Shah, N. H., Halamka, J. D., Saria, S., Pencina, M. & Tazbaz, T. A nationwide network of health AI assurance laboratories. *JAMA* **331**, 245–249. 10.1001/jama.2023.26930 (2024).
15. Coalition for Health AI. *Blueprint for Trustworthy AI Implementation Guidance and Assurance for Healthcare* (Coalition for Health AI, 2023).
16. Qazi, I. A. et al. Automation bias in large language model–assisted diagnostic reasoning among physicians trained in AI literacy: a randomized clinical trial. *NEJM AI.* **3**, A10a2501001. 10.1056/A10a2501001 (2026).
17. Kalai, A. T. & Vempala, S. S. Calibrated language models must hallucinate. In Proceedings of the 56th Annual ACM Symposium on Theory of Computing (STOC) 160–171 (2024).
18. Banerjee, S., Agarwal, A. & Singla, S. LLMs will always hallucinate, and we need to live with this. Preprint at (2024). <https://arxiv.org/abs/2409.05746>
19. Xu, Z., Jain, S. & Kankanhalli, M. Hallucination is inevitable: an innate limitation of large language models. Preprint at (2024). <https://arxiv.org/abs/2401.11817>
20. Karpowicz, M. P. On the fundamental impossibility of hallucination control in large language models. In Proceedings of the International Conference on Learning Representations (ICLR) (2026).
21. Wind, S. et al. Safety and accuracy follow different scaling laws in clinical large language models. Preprint at (2026). <https://arxiv.org/abs/2605.04039>
22. Koh, P. W. et al. Concept bottleneck models. *Proc. Int. Conf. Mach. Learn.* 119, 5338–5348 (2020).
23. Yeh, C. K. et al. On completeness-aware concept-based explanations in deep neural networks. *Adv. Neural Inf. Process. Syst.* **33**, 20554–20565 (2020).
24. Riegel, R. et al. Logical Neural Networks. Preprint at (2020). <https://arxiv.org/abs/2006.13155>
25. Saab, K. et al. Advancing conversational diagnostic AI with multimodal reasoning. *Nat. Med.* **32**, 1726–1736. 10.1038/s41591-026-04371-0 (2026).
26. Chen, X. et al. All circuits lead to Rome: rethinking functional anisotropy in circuit and sheaf discovery for LLMs. *Proc. Int. Conf. Mach. Learn.* (2026).
27. Cruz Rivera, S. et al. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. *Nat. Med.* **26**, 1351–1363 (2020).

28. Liu, X. et al. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. *Nat. Med.* **26**, 1364–1374 (2020).
29. Gallifant, J. et al. The TRIPOD-LLM reporting guideline for studies using large language models. *Nat. Med.* **31**, 60–69 (2025).
30. Jin, D. et al. What disease does this patient have? A large-scale open-domain question-answering dataset from medical examinations. *Appl. Sci.* **11**, 6421 (2021).
31. Baum, A., Batniji, R., Ratcliffe, H., DeGosztonyi, M. & Basu, S. Supporting rising-risk Medicaid patients through early intervention. *NEJM Catal. Innov. Care Deliv.* **5** 10.1056/CAT.24.0060 (2024).
32. Brown, T. B. et al. Language models are few-shot learners. *Adv. Neural Inf. Process. Syst.* **33**, 1877–1901 (2020).
33. By the 2023 American Geriatrics Society Beers Criteria Update Expert Panel. American Geriatrics Society 2023 updated AGS Beers Criteria for potentially inappropriate medication use in older adults. *J. Am. Geriatr. Soc.* **71**, 2052–2081 (2023).
34. Ramaswamy, A. et al. ChatGPT Health performance in a structured test of triage recommendations. *Nat. Med.* **32**, 1671–1675. 10.1038/s41591-026-04297-7 (2026).
35. Bhattacharyya, M. et al. High rates of fabricated and inaccurate references in ChatGPT-generated medical content. *Cureus* **15**, e39238 (2023).
36. Stokel-Walker, C. Scientists invented a fake disease. AI told people it was real. *Nature* **652**, 559–561 (2026).
37. He, P., Gao, J. & Chen, W. DeBERTaV3: improving DeBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. In Proceedings of the International Conference on Learning Representations (ICLR) (2023).
38. Johnson, A. E. W. et al. MIMIC-IV, a freely accessible electronic health record dataset. *Sci. Data.* **10**, 1 (2023).
39. O'Mahony, D. et al. STOPP/START criteria for potentially inappropriate prescribing in older people: version 3. *Eur. Geriatr. Med.* **14**, 625–632 (2023).
40. Newcombe, R. G. Interval estimation for the difference between independent proportions: comparison of eleven methods. *Stat. Med.* **17**, 873–890 (1998).
41. Wilson, E. B. Probable inference, the law of succession, and statistical inference. *J. Am. Stat. Assoc.* **22**, 209–212 (1927).
42. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R Stat. Soc. B.* **57**, 289–300 (1995).

Figures

Verification framework architecture

A decoupled per-output certificate detects failure modes the upstream model cannot self-detect

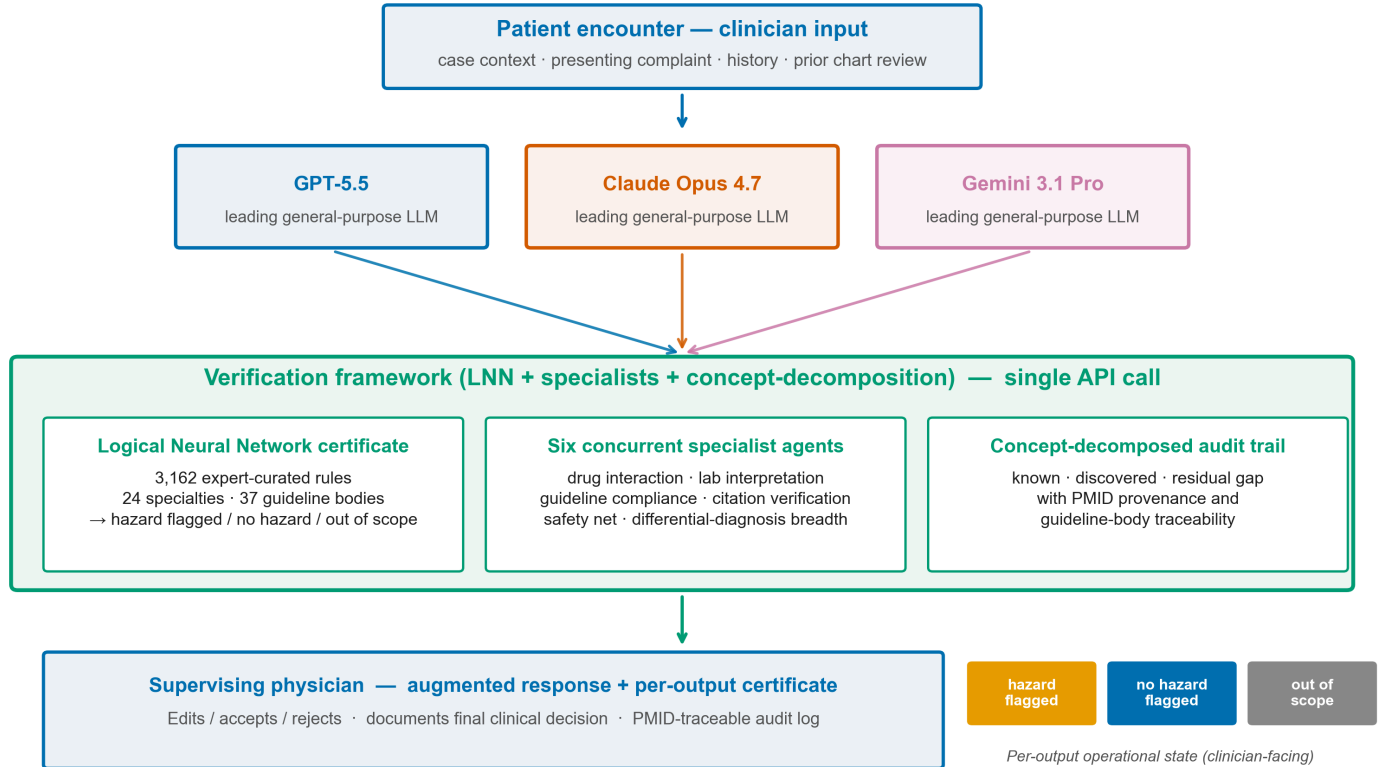


Figure 1

Verification Framework Architecture

Schematic of the three-component downstream verification framework layered on the model's standard-guardrail-stack output. The three components are a Logical Neural Network certificate over a 3,162-rule clinical-rule library, six concurrent specialist verification agents with a consensus gate requiring at least two agents to emit the same named hazard, and a concept-decomposition module that emits a deterministic residual-attribution-gap statement on every output. Arrows show the data flow from the model output through each component to the final verified recommendation. The colour palette is colourblind-safe and avoids red-green contrast.

Three-stage progression on safety-critical omission

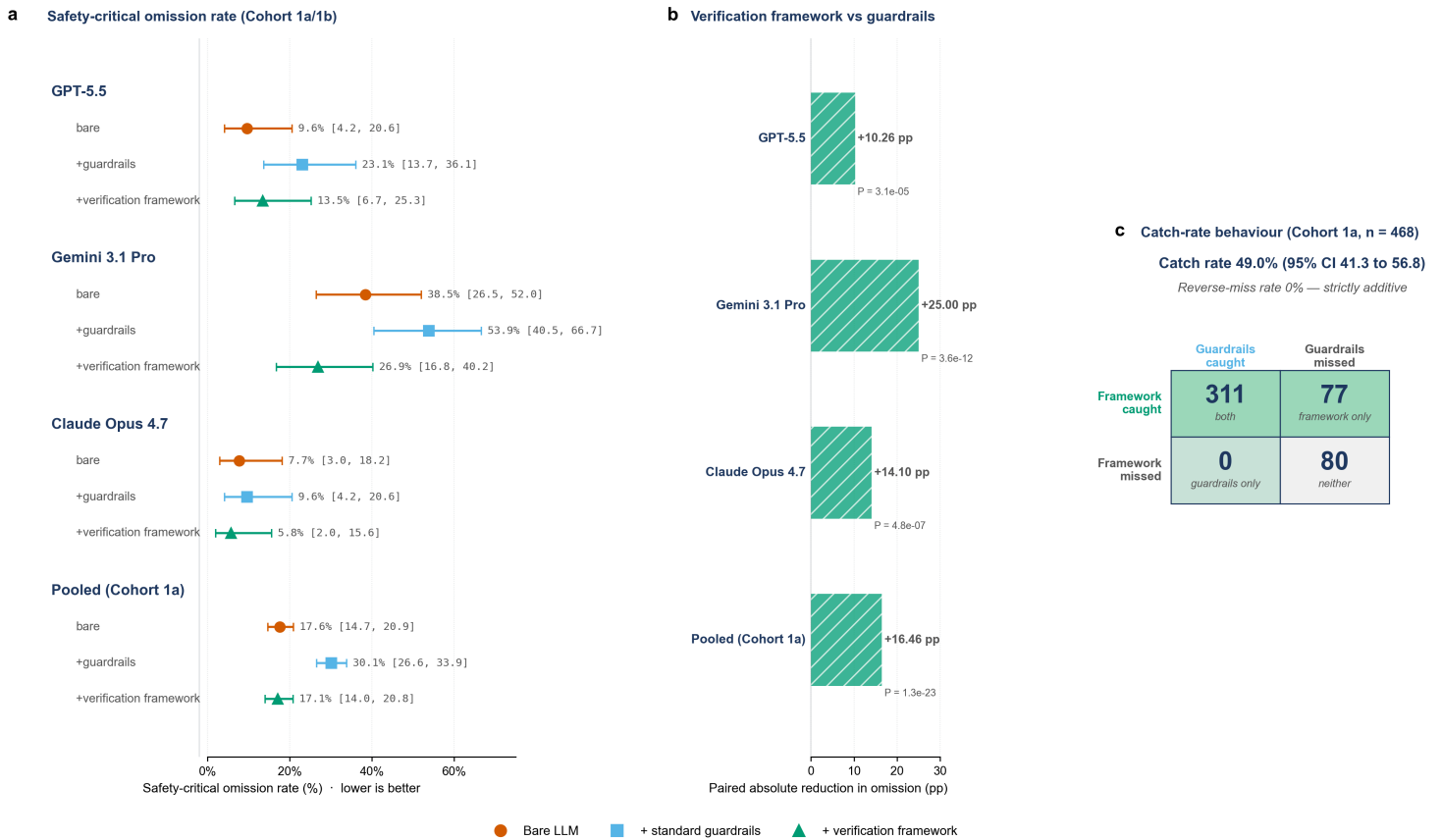


Figure 2

Three-Stage Progression on the Primary Safety-Critical Omission Endpoint

Panel a shows the per-model and pooled safety-critical omission rate on Cohort 1a and 1b across the three configurations (bare large language model 17.6 percent, standard guardrails 30.1 percent, standard guardrails plus verification framework 17.1 percent pooled), with Wilson 95 percent confidence intervals shown as error bars. Panel b shows the paired absolute reduction in safety-critical omission (verification framework versus standard guardrails) by model and pooled (plus 16.46 percentage points pooled; Newcombe 95 percent confidence interval 10.93 to 21.86; paired McNemar exact P equals 1.32 times ten to the minus twenty-third), with error bars showing the Newcombe interval. Panel c shows the catch-rate contingency table: of 157 safety-critical omissions missed by the standard guardrail stack, the verification framework caught 77 (49.0 percent; Wilson 95 percent confidence interval 41.3 to 56.8) and missed none of the omissions the guardrails caught, confirming strictly additive behaviour. The colour palette is colourblind-safe and avoids red-green contrast.

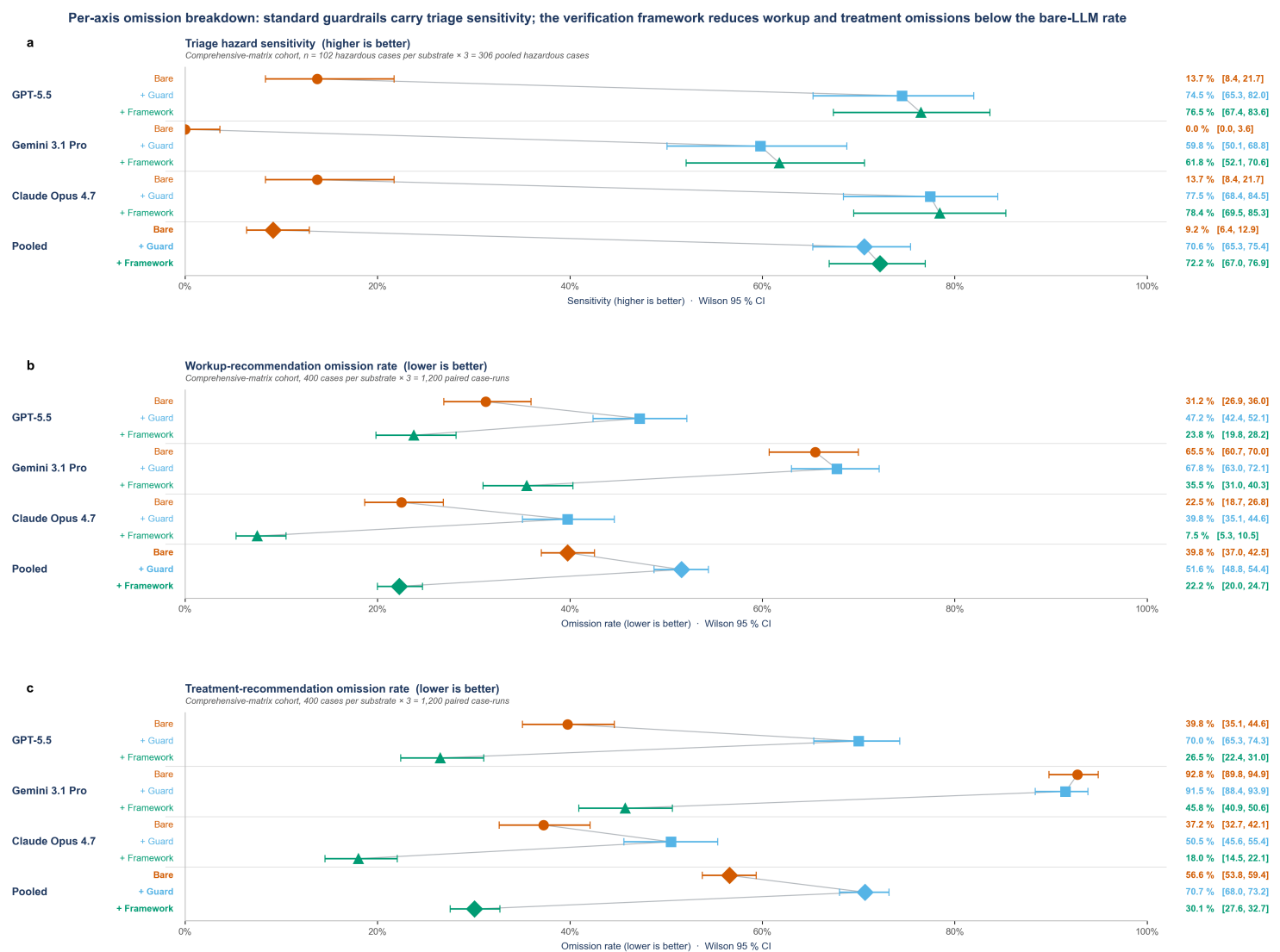


Figure 3

Per-Axis Safety-Critical Omission Breakdown Across the Three Configurations

Three forest plots show the bare large language model, the standard guardrail stack, and the verification framework progression on three output axes of Cohort 1b (comprehensive-matrix, n equals 400 cases per model times 3 models equals 1,200 paired case-runs). Panel a shows triage hazard sensitivity, where the standard guardrail stack carries the majority of the gain (9.2 percent under the bare large language model, 70.6 percent under standard guardrails, and 72.2 percent under the verification framework). Panel b shows the workup-recommendation safety-critical omission rate. Panel c shows the treatment-recommendation safety-critical omission rate. The pattern shows that the standard guardrail stack contributes most on the triage axis and the verification framework contributes most on the workup and treatment axes, complementing the comprehensiveness-axis effects in Figure 2. Error bars show Wilson 95 percent confidence intervals. The colour palette is colourblind-safe and avoids red-green contrast.