

Inconsistent and divergent large language model safety guidance during intraoperative crises: a guideline- anchored benchmark

Brian H. Park

bhpark@health.ucsd.edu

University of California San Diego

Claire S. Soria

University of California San Diego

Preetham Suresh

University of California San Diego

Ryan Broderick

University of California San Diego

Bryan Sandler

University of California San Diego

Research Article

Keywords: Large language models, Patient safety, Intraoperative crises, Clinical decision support, Benchmarking, Anesthesiology, Surgery

Posted Date: September 7th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10438275/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

Large language models (LLMs) are increasingly proposed for clinical decision support, yet no benchmark tests whether they perform guideline-anchored safety actions or give consistent guidance across repeated runs during intraoperative crises, where a single omission or contraindicated recommendation may cause immediate harm. We developed a phase-based benchmark of ten intraoperative crisis archetypes, each unfolding over five to seven deterioration phases and scored against 150 binary sentinel actions anchored to clinical practice guidelines and published crisis checklists. Three frontier LLMs (ChatGPT 5.2, Claude Opus 4.6, Gemini 3.1 Pro) were evaluated through consumer web interfaces, three runs per scenario, and independently double-scored by blinded clinicians. The proportion of required safety actions a model completed varied across repeated runs of the same scenario in 29 of 30 model-scenario combinations. Between two runs of the same model, completion differed by a mean of 14.7 to 18.1 percentage points, with overlapping confidence intervals that do not rank the models by stability. Omissions fell most on communication and escalation actions, where omission varied nearly threefold across models under an identical prompt (24.6 to 72.6 percent). In a secondary comparison, aggregate completion also differed (Claude 83.7 percent, ChatGPT 69.0 percent, Gemini 61.2 percent), but unsafe actions were rare with overlapping intervals. Frontier LLMs gave inconsistent safety guidance across repeated runs and diverged sharply on communication and escalation under an identical prompt. Models proposed for perioperative crisis roles should be judged on how reliably and reproducibly they act, not on aggregate accuracy alone.

Introduction

Despite rapid expansion of large language model (LLM) research in healthcare, existing evaluation frameworks focus on outpatient counseling, diagnostic reasoning, or general clinical safety [1–3]. These benchmarks do not examine LLM performance during intraoperative crises. In these settings, deteriorating physiology, time pressure, and the interdependence of surgical and anesthetic management drive patient outcomes [4, 5]. Recent benchmarks have begun to score the overall quality of model responses to perioperative crises, but none tests whether models perform the specific guideline-anchored safety actions these crises require, avoid unsafe maneuvers, or remain stable across repeated runs.

Intraoperative emergencies differ fundamentally from outpatient or diagnostic tasks. Decision pathways are tightly coupled to ongoing surgical maneuvers, anesthetic interventions, and rapidly changing physiology. A single omission or contraindicated recommendation may cause immediate harm. Examples include failure to activate massive transfusion, failure to decompress tension physiology, and continued carbon dioxide insufflation despite suspected gas embolism. Structured crisis checklists and emergency manuals are established components of perioperative safety practice and improve team performance during simulated and actual crises [6–8]. The operating room is therefore a natural stress test for LLM safety rather than a downstream deployment setting.

We developed a phase-based benchmark for evaluating LLM behavior across ten intraoperative emergencies. The benchmark measures whether models perform the required safety actions, avoid the unsafe ones, escalate appropriately, and stay consistent across repeated runs. By locating where models fail rather than only how often, it provides a reusable instrument for judging whether and how LLMs might support perioperative crisis care.

Methods

Study design

Ten crisis scenarios were constructed. Model outputs were scored against predefined positive and unsafe sentinel actions drawn from perioperative crisis guidelines, resuscitation standards, and emergency manuals.

Scenario development

Scenarios were drafted by an attending anesthesiologist using published perioperative bleeding, airway, resuscitation, sepsis, and crisis-management guidance as reference standards [9-19]. Each scenario was designed to reflect the clinical progression of an evolving intraoperative emergency, from early instability through decompensation, strategic inflection points, and immediate postoperative disposition.

Each sentinel named a single discrete action rather than a broad goal or a fine sub-step. Examples of the intended level are activating massive transfusion or declaring an emergency, as opposed to a general instruction to resuscitate or a specific drug dose. Actions that depend on local resources or practice patterns were excluded. These included calling a named institutional team, requiring a specific device model, or following a unit-specific protocol variant. Resource-dependent interventions were placed in later phases when they represented escalation rather than immediate rescue.

Each scenario stem gave the patient's age, comorbidities, procedure, and current physiologic trajectory. This was enough information for a clinician to respond to the crisis without the diagnosis being named. Diagnostic recognition and differential reasoning were treated as latent abilities, rather than scored as sentinels. Each was captured indirectly through whether the model selected the guideline-concordant actions as the scenario evolved.

Sentinel framework

Sentinel actions were categorized as positive or unsafe. Positive sentinels were high-impact actions expected of competent crisis management. Unsafe sentinels were contraindicated maneuvers likely to worsen the patient's trajectory.

Each sentinel was mapped to one of six cross-crisis safety axes: stopping ongoing harm, activating the correct algorithm, definitive life-saving maneuvers, global strategy and disposition, unsafe red-flag actions, and communication and escalation. Stability across repeated runs was treated as an analysis-

only dimension rather than a sentinel category. These six axes constitute a study-developed framework, informed by crisis resource management principles and operating-room crisis-checklist literature [6-8]. The first four axes capture the therapeutic and strategic actions that change physiology. The fifth captures the contraindicated actions to avoid. The sixth captures the team behaviors that crisis checklists identify as central to performance [6-8].

Sentinel anchoring was tiered. The full mapping is reported in Supplementary Table S1. Of the 150 sentinels, 113 mapped to a named clinical practice guideline recommendation or a directly entailed guideline principle. These drew on perioperative bleeding, pulmonary embolism, sepsis, anaphylaxis, malignant hyperthermia, difficult-airway, and advanced cardiac life support guidelines [9-19]. A further 26 covered the communication and escalation actions and the decisions to stop or continue surgery during a crisis. These were anchored to published operating-room crisis checklists and emergency manuals [6-8]. The remaining 11 sentinels reflected established surgical principle and expert consensus rather than a dedicated perioperative crisis guideline. These were limited to capnothorax-specific maneuvers and hepatobiliary damage-control steps. Team-based sentinels were limited to explicit crisis declaration and calling for help, rather than broader teamwork and leadership behaviors.

Expert review

Draft scenarios and sentinel lists underwent iterative expert review by surgeons and anesthesiologists to assess clinical accuracy, physiologic coherence, phase alignment, and appropriateness of positive versus unsafe classification. This process validated safety-critical content and cross-scenario consistency rather than performing formal consensus development.

Models and prompting

Three frontier general-purpose LLMs were evaluated through consumer web interfaces: ChatGPT 5.2 (OpenAI), Claude Opus 4.6 (Anthropic), and Gemini 3.1 Pro (Google). Consumer interfaces were chosen to capture the sampling variability end users encounter rather than the fixed-parameter conditions of application programming interface (API) access. Each model completed three independent runs of each scenario between February and March 2026. The prompt text, run count, and output format were standardized across runs. Each run was conducted in a fresh session with no custom instructions or cross-session memory enabled. The models were evaluated as publicly released, without fine-tuning or retrieval augmentation, and the consumer interfaces did not disclose exact build identifiers or training-data cutoff dates.

All primary runs used a standardized prompting protocol. Each scenario began with a short operative briefing containing patient demographics and operative context, followed by sequential phase cues. After each phase, the model was asked, “What should the OR team do next?” Models were not shown the sentinel lists. To enable reproducible binary scoring, the system prompt constrained each response to four to eight single-action imperative bullets. It also directed the model to “prioritize decisive management actions over monitoring, confirmation, or evaluation steps” and to “avoid simulation micro-

steps (monitor-by-monitor troubleshooting, exhaustive checklists).” This action-oriented format is central to interpreting the axis-level results, because it discouraged the checklist-style output to which several communication and escalation sentinels are anchored. A prompt-sensitivity analysis comparing three prompt variants on one scenario was specified in advance, although these variants altered persona and reasoning style rather than the output-format constraints. The full system briefing, the per-phase query, and all scenario phase cues are provided verbatim in Supplementary Methods S1.

Scoring

Primary scoring used binary sentinel assessment, with positive sentinels scored present or absent and unsafe sentinels scored triggered or not triggered. Raters scored clinical equivalence rather than exact wording, and judged whether each action was present rather than when it occurred. If a rater left a sentinel unscored, that sentinel was excluded from the run’s denominator for that rater. Phase assignments were retained as clinical context for scenario construction and interpretation.

Scoring used a two-tier rater design with a pool of four clinicians (three anesthesiologists and one surgeon), all of them study investigators and all blinded to model identity. In the first tier, the four raters independently scored a calibration subset spanning protocol-heavy, judgment-heavy, and ambiguous scenarios, which refined the operational definitions before full scoring. In the second tier, every model-scenario combination (one model on one scenario, 30 in all) was independently double-scored, with reliability-fill batches added where initial coverage was unbalanced so that each combination carried at least two doubly scored runs. Because the two raters worked without an adjudicator, each sentinel was scored as the mean of the two raters’ binary judgments, so a split between raters counted as one half. Headline completion and unsafe-action estimates were computed from doubly scored runs only, whereas run-to-run instability used all three runs of every combination as a measure of model behavior. Supplementary prompt-sensitivity runs (variants in Supplementary Methods S1.4) followed a single-investigator protocol and were excluded from inter-rater reliability analyses.

Outcomes

The following outcomes were specified a priori: positive-sentinel completion rate, critical omission rate, unsafe-action rate, and run-to-run instability. Run-to-run instability was defined as the standard deviation of per-run completion across the three runs of each model-scenario combination. Secondary analyses, also specified a priori, included axis-level completion, completion by sentinel anchoring tier, scenario-level risk profiles, the distribution of failures across deterioration phases, and prompt sensitivity. Ranking stability across disagreement-resolution rules and against the full single-rater cohort was an exploratory robustness check.

Statistical analysis

Each model’s completion rate was estimated from the doubly scored runs as the mean of scenario-level means rather than by pooling across runs. This estimator accounts for the nesting of repeated runs

within scenarios and multiple sentinels within runs [20, 21]. Ninety-five percent confidence intervals were obtained by a nonparametric cluster bootstrap that resampled scenarios as the independent unit [22], treating the ten scenarios rather than the nested runs as the source of sampling variation. Pairwise differences in completion between models were tested with the Wilcoxon signed-rank test across the ten scenarios, using the normal approximation because tied and zero differences were present. Because three comparisons were made, the three raw *P* values were ranked and interpreted against Holm-Bonferroni thresholds to control the family-wise error rate. Inter-rater reliability was summarized with Cohen's kappa, with percent agreement reported alongside kappa for the rare unsafe events. The benchmark used no human-clinician reference standard, so between-model differences are interpreted relative to one another rather than against an external gold standard, and probability calibration was not assessed because every sentinel was scored as a binary present-or-absent outcome. Analyses were conducted in Python. This study is reported against the TRIPOD-LLM and MI-CLEAR-LLM guidelines [23, 24]. The completed MI-CLEAR-LLM checklist is provided in Supplementary Appendix S1, and a model card and benchmark datasheet documenting intended use, evaluation factors, performance across settings, and recommended monitoring are provided in Supplementary Appendix S2.

Ethics and oversight

This study evaluated the outputs of large language models and did not involve human participants, patient data, or protected health information. Institutional review board review was therefore not required. Because no human participants were studied, sex- and gender-based reporting under the SAGER guidelines does not apply.

Results

Scoring completeness

Scoring covered all 90 runs across the 30 model-scenario combinations, of which 62 were independently double-scored so that every combination carried at least two doubly scored runs (Claude *n* = 20, ChatGPT *n* = 21, Gemini *n* = 21).

Positive-sentinel completion and critical omissions

Positive-sentinel completion was highest for Claude, at 83.7 percent (95% CI, 77.4 to 89.4), followed by ChatGPT at 69.0 percent (95% CI, 62.6 to 77.3) and Gemini at 61.2 percent (95% CI, 50.2 to 71.8) (Table 1, Fig. 1). In paired comparisons across the ten scenarios, Claude completed more safety actions than both competitors, exceeding ChatGPT by 14.7 percentage points (95% CI, 7.4 to 21.3; Wilcoxon signed-rank *P* = .015) and Gemini by 22.5 percentage points (95% CI, 14.4 to 30.8; *P* = .008). Both of Claude's separations remained significant under Holm-Bonferroni correction for the three pairwise comparisons. ChatGPT and Gemini were not statistically distinguishable in this ten-scenario sample, differing by 7.8 percentage points (95% CI, -0.5 to 15.9; *P* = .086). The complementary critical-omission rate followed the same ordering, rising from 16.3 percent for Claude to 31.0 percent for ChatGPT and

38.8 percent for Gemini (Table 1). Completion also varied across the ten scenarios, from 56 percent in septic shock to 90 percent in the can't-intubate can't-oxygenate airway emergency (Fig. 2). These between-model completion comparisons are secondary to the reliability findings that follow, because the benchmark includes no human-clinician reference standard to anchor an absolute ranking.

Table 1

Model safety outcomes across ten intraoperative crisis scenarios. Estimates are the mean of scenario-level means from double-scored runs, with 95% cluster-bootstrap confidence intervals resampling the ten scenarios. Critical omission is the complement of positive-sentinel completion.

Model	Positive-sentinel completion, % (95% CI)	Critical omission, %	Unsafe-action, % (95% CI)
Claude Opus 4.6	83.7 (77.4 to 89.4)	16.3	3.8 (1.0 to 7.5)
ChatGPT 5.2	69.0 (62.6 to 77.3)	31.0	8.3 (3.7 to 13.2)
Gemini 3.1 Pro	61.2 (50.2 to 71.8)	38.8	10.3 (3.7 to 17.8)

Unsafe actions

Unsafe sentinels were triggered in a minority of runs for every model, in 3.8 percent for Claude (95% CI, 1.0 to 7.5), 8.3 percent for ChatGPT (95% CI, 3.7 to 13.2), and 10.3 percent for Gemini (95% CI, 3.7 to 17.8) (Table 1). These rates ordered in the same direction as the omissions, though the overlapping confidence intervals do not support a ranking among the three models.

Run-to-run instability

Completion changed between repeated runs in 29 of the 30 model-scenario combinations. Between two runs of the same model, completion differed by a mean of 14.7 percentage points for Claude (95% CI, 10.8 to 18.9), 15.9 for Gemini (95% CI, 9.8 to 22.2), and 18.1 for ChatGPT (95% CI, 12.8 to 24.8) (Fig. 3). The mean within-model standard deviation of per-run completion was 0.11 (95% CI, 0.09 to 0.15), 0.13 (95% CI, 0.08 to 0.18), and 0.15 (95% CI, 0.10 to 0.21), respectively. These confidence intervals overlap, so the models cannot be ranked by stability. A given model therefore offered materially different safety guidance on identical input, with no reliable between-model difference in how much.

Failure taxonomy by safety axis

Omissions were not distributed evenly across the six safety axes (Fig. 4, Panel A). For every model, the communication and escalation axis carried the highest omission rate, reaching 72.6 percent for Gemini against 36.1 percent for ChatGPT and 24.6 percent for Claude. The definitive life-saving-maneuver axis separated the models in the opposite direction, where Claude omitted 4.6 percent of these actions whereas ChatGPT and Gemini omitted 31.9 and 25.5 percent. A similar gradient appeared for global strategy and disposition, in which Claude omitted no required action across 34 observations, compared with 7.8 percent for ChatGPT and 19.0 percent for Gemini. Stop-harm omissions were more uniform,

between 22 and 39 percent across the three models. The activate-algorithm axis held six to seven observations per model and is directional, with omission ranging from 16.7 to 33.3 percent.

Failures across deterioration phases

Failures also clustered by deterioration phase, concentrating in the first and third phases (Fig. 4, Panel B). First-phase omission, at the point of crisis recognition, reached 57.6 percent for Gemini against 40.0 percent for ChatGPT and 26.2 percent for Claude. The third phase, where escalation decisions fall due, showed Gemini and ChatGPT near half at 51.0 and 46.5 percent, with Claude lower at 23.4 percent. Omissions were rarest in the final disposition phase, where Claude and ChatGPT missed no required action and Gemini omitted 16.7 percent. Completion by deterioration phase for each model is reported in full in Supplementary Table S4.

Completion by sentinel anchoring tier

Completion varied with the evidence base of each sentinel (Supplementary Table S2). Across all models, completion was highest for the small set of consensus-anchored actions, at 92.9 percent, and for clinical-practice-guideline actions, at 81.2 percent. It then fell to 68.5 percent for guideline-concordant actions and to 57.2 percent for the 26 actions anchored to crisis checklists and crew-resource-management standards. This checklist-anchored tier accounted for most of the between-model spread, where Claude completed 77.1 percent and Gemini 35.3 percent. Completion broken out by individual guideline and checklist source is reported in Supplementary Table S3. Only 27 of the 100 positive sentinels were completed by all three models on every double-scored run, and only two were missed by all three. The remaining 71 separated the models, so completion gaps reflect model-specific behavior rather than uniform item difficulty.

Inter-rater reliability

Among double-scored runs, agreement between the two primary anesthesiologist raters corresponded to a Cohen's kappa of 0.74 across 520 sentinel observations. The second-pass anesthesiologist agreed with each at kappa 0.73 and 0.74 (Table 2). Cross-specialty agreement between the surgeon rater and the two anesthesiologists was comparable, at kappa 0.77 and 0.71. Reliability was lower for the positive-sentinel subset alone, where kappa was 0.60 across 345 observations. For unsafe sentinels, raters agreed on more than 91 percent of observations, yet the corresponding kappa was 0.25. Unsafe events were rare, so agreement expected by chance was high.

Table 2

Inter-rater reliability among the four blinded raters. Cohen's kappa and percent agreement are reported for all sentinels and, where noted, for the positive and unsafe subsets. N is the number of paired sentinel observations.

Rater comparison	Cohen's kappa	Agreement, %	N
Anesthesiologist 1 vs 2	0.74	86.9	520
Anesthesiologist 1 vs 3	0.73	86.5	347
Anesthesiologist 2 vs 3	0.74	87.2	274
Surgeon vs Anesthesiologist 1	0.77	88.6	132
Surgeon vs Anesthesiologist 2	0.71	85.5	131
Positive sentinels (Anesthesiologist 1 vs 2)	0.60	84.6	345
Unsafe sentinels (Anesthesiologist 1 vs 2)	0.25	91.4	175

Sensitivity analyses

The completion ranking was preserved across all three disagreement-resolution rules and when the full single-rater cohort was substituted for the doubly scored primary. In the single-scenario prompt-sensitivity analysis, ChatGPT completion ranged from 70.0 to 76.7 percent across three prompt variants, a span of 6.7 percentage points.

Discussion

Principal findings

Across ten guideline- and checklist-anchored intraoperative crisis scenarios, the two most consistent findings concerned reliability rather than rank. Every model gave different safety guidance when the same crisis was presented again, changing completion in 29 of 30 model-scenario combinations. The models also diverged sharply on the communication and escalation axis, where omission varied nearly threefold under an identical prompt and identical scoring. Against this backdrop, aggregate completion differed between models, highest for Claude, which separated from both competitors, whereas ChatGPT and Gemini did not differ significantly. All three triggered unsafe actions in a minority of runs, with overlapping intervals leaving that ranking uncertain.

From prediction to safety

Most prior work on large language models in perioperative care has asked whether a model can recognize a diagnosis or predict an outcome. No widely accepted evaluation standard exists, and benchmarks rarely measure concordance with time-critical, guideline-anchored crisis actions [4]. During a crisis the question is different. It is not whether a model names the problem but whether it recommends the actions that prevent harm and avoids those that cause it. Recent evidence that models

give accurate yet unsafe answers makes this distinction consequential [25, 26]. The present benchmark measures guideline- and checklist-anchored safety actions under physiologic stress rather than diagnostic accuracy.

This benchmark builds on emerging work evaluating LLMs in perioperative crisis care. Hurley and colleagues benchmarked twenty-two models on seven perioperative crisis scenarios, scoring each on a quality rubric at temperature zero [27]. Al-Azzam and colleagues evaluated five models across twenty intraoperative emergencies on a structured rubric [28]. Related work has examined model-generated anesthesia plans [29] and decision-making for a single procedure [30]. Each graded a single response per case on overall quality rather than specific guideline-anchored safety actions, and none repeated a scenario, so their designs cannot observe run-to-run variation or ask whether guidance is reproducible. By repeating every scenario under consumer sampling conditions and scoring each required and unsafe action individually, the present benchmark exposes the instability and between-model divergence these single-pass designs cannot detect.

What the sentinel framework shows

Each score is a binary, phase-anchored sentinel tied to a stated safety rationale, and the cross-crisis axes locate failures rather than merely counting them. Models generally performed the named definitive maneuvers that rescue each crisis. Their omissions fell disproportionately on the crisis-checklist and crew-resource-management layer, whose actions (crisis declaration, structured help-seeking, explicit escalation) are drawn from validated emergency manuals rather than soft teamwork preferences [6–8]. The models also differed sharply on this axis. Under an identical prompt and identical scoring, omission ranged from 24.6 percent for Claude to 72.6 percent for Gemini. Holding prompt and scoring constant rules out a scoring artifact, but not whether the spread reflects differing safety behavior or differing compliance with the action-first instruction. These omissions peaked at the recognition and escalation phases, a structure a single aggregate score would have hidden.

Implications for evaluation and selection

The practical safety question is whether a model reliably avoids critical omissions and unsafe recommendations, not whether it is usually correct. Two of these findings speak directly to how such models should be compared and selected. First, the models diverged most on the communication and escalation layer, exactly the scaffolding that crisis checklists already supply, so a model's value depends on how it is paired with those tools, not on its aggregate score. Second, run-to-run instability is itself a selection criterion. A tool that gives different safety guidance on the same crisis cannot serve as consistent decision support, a limitation documented for other clinical tasks [31–33]. This instability is not a universal property of current models. On tightly constrained classification tasks, repeated runs agree almost perfectly [34], and the variation seen here emerges under open-ended crisis guidance delivered through consumer interfaces. Reproducibility should therefore be measured and reported alongside accuracy whenever such a tool is evaluated for a perioperative role.

Limitations

The benchmark rests on ten synthetic scenarios, which limits power to separate models of similar performance. Each scenario nonetheless unfolds over multiple deterioration phases with several sentinels apiece, so the analysis draws on more than two thousand individual safety judgments. The 150 sentinel actions were selected and refined through iterative expert review rather than formal consensus development. Sentinels were scored as present or absent, crediting neither partially correct nor appropriately timed actions. Some anchoring tiers and safety axes rest on only a handful of sentinels or scenarios, so those completion rates are directional.

Phase cues were brief and objective rather than the full context of a live operating room, and model outputs were constrained to short action lists. This format standardizes scoring but may shift model behavior relative to richer clinical input. The system prompt directed models toward decisive management and away from exhaustive checklists, likely depressing completion of the checklist-anchored communication and escalation sentinels. The prompt-sensitivity analysis did not vary this output-format constraint, so it cannot isolate its effect. The absolute omission rate on that axis should therefore be read as conditional on an action-first protocol rather than as an intrinsic model property, though the large between-model differences under an identical prompt are not explained by the constraint alone.

In four runs one rater left a single positive sentinel unscored, too few to affect coverage or ordering. The ranking was stable across three disagreement-resolution rules and the full single-rater cohort, with substantial inter-rater reliability. The raters were study investigators rather than independent evaluators, which could introduce expectation bias despite blinding to model identity. Excluding any single rater nonetheless left both the completion ranking and the separation of the leading model unchanged.

Because models were accessed through consumer web interfaces, the evaluation reflects real-world conditions but allowed no control over decoding parameters such as temperature and could not guarantee exact reproducibility. The observed instability therefore reflects the deployed consumer system, which may apply its own sampling, rather than necessarily an intrinsic property of the underlying model. The scenarios are written vignettes, not live simulations, and probe recommended actions rather than execution. They also include no human-clinician comparator, leaving the intrinsic difficulty of individual sentinels externally uncalibrated. Finally, frontier models change frequently, and specific versions may be updated or withdrawn without notice. These rankings are therefore a snapshot rather than a durable property of any product, a reproducibility limit that also constrains API-based evaluations.

Conclusions

Across ten guideline- and checklist-anchored intraoperative crisis scenarios, frontier large language models gave inconsistent safety guidance when the same crisis was presented again and diverged sharply from one another on communication and escalation actions under an identical prompt.

Aggregate completion differed between models, but the run-to-run instability and the between-model divergence bear most directly on deployment. A model that returns different safety guidance on repeated presentations cannot serve as consistent decision support. Reproducibility should therefore be measured and reported alongside accuracy whenever such a tool is evaluated or selected for a perioperative role.

The principal contribution is the benchmark instrument rather than the current model scores. The scenario definitions, sentinel lists, safety axes, scoring rules, and prompt templates are designed for controlled release to independent evaluators. Other groups can then ask the same guideline-anchored safety question of each new model release rather than rebuilding the evaluation each time.

Declarations

Competing interests: The authors declare no competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Funding: No funding was received for conducting this study.

Ethics approval: This study evaluated the outputs of large language models and did not involve human participants, patient data, or protected health information. Institutional review board approval was not required.

Consent to participate and consent to publish: Not applicable. No human participants were involved.

Registration: This study was not registered.

Clinical trial number: Not applicable.

Data availability: The benchmark assets that support the findings of this study (scenario definitions, sentinel lists, safety axes, scoring rules, and prompt templates) are available to qualified researchers from the corresponding author on reasonable request. They are not posted in a fully public repository to limit their incorporation into model training data, which would compromise the benchmark's validity for evaluating future models.

Code availability: The analysis code is available from the corresponding author on reasonable request.

Author contributions: B.H. Park: Conceptualization, Methodology, Software, Formal analysis, Investigation, Data curation, Writing – original draft, Writing – review and editing, Visualization, Project administration. C.S. Soria: Methodology, Investigation, Validation, Writing – review and editing. P. Suresh: Methodology, Investigation, Validation, Writing – review and editing. R. Broderick: Investigation, Validation, Writing – review and editing. B. Sandler: Investigation, Validation, Writing – review and editing.

Patient and public involvement: No patients or members of the public were involved in the design, conduct, reporting, or dissemination of this research.

Use of generative artificial intelligence: The authors did not use generative artificial intelligence tools in the preparation of this manuscript. Large language models were the subject of study, and their evaluation is described in the Methods.

References

1. Singhal K, Tu T, Gottweis J et al (2025) Toward expert-level medical question answering with large language models. *Nat Med* 31:943-950. <https://doi.org/10.1038/s41591-024-03423-7>
2. Rao AS, Esmail KP, Lee RS et al (2026) Large language model performance and clinical reasoning tasks. *JAMA Netw Open* 9:e264003. <https://doi.org/10.1001/jamanetworkopen.2026.4003>
3. Waldock WJ, Zhang J, Guni A, Nabeel A, Darzi A, Ashrafian H (2024) The accuracy and capability of artificial intelligence solutions in health care examinations and certificates: systematic review and meta-analysis. *J Med Internet Res* 26:e56532. <https://doi.org/10.2196/56532>
4. Shi T, Ma J, Yu Z et al (2025) Large language models in critical care medicine: scoping review. *JMIR Med Inform* 13:e76326. <https://doi.org/10.2196/76326>
5. Daccache N, Zako J, Morisson L, Laferrière-Langlois P (2025) The applications of ChatGPT and other large language models in anesthesiology and critical care: a systematic review. *Can J Anaesth* 72:904-922. <https://doi.org/10.1007/s12630-025-02973-9>
6. Arriaga AF, Bader AM, Wong JM et al (2013) Simulation-based trial of surgical-crisis checklists. *N Engl J Med* 368:246-253. <https://doi.org/10.1056/NEJMsa1204720>
7. Goldhaber-Fiebert SN, Bereknyei Merrell S, Agarwala AV et al (2020) Clinical uses and impacts of emergency manuals during perioperative crises. *Anesth Analg* 131:1815-1826. <https://doi.org/10.1213/ANE.0000000000005012>
8. Hepner DL, Arriaga AF, Cooper JB et al (2017) Operating room crisis checklists and emergency manuals. *Anesthesiology* 127:384-392. <https://doi.org/10.1097/ALN.0000000000001731>
9. Kietaihl S, Ahmed A, Afshari A et al (2023) Management of severe peri-operative bleeding: guidelines from the European Society of Anaesthesiology and Intensive Care: second update 2022. *Eur J Anaesthesiol* 40:226-304. <https://doi.org/10.1097/EJA.0000000000001803>
10. American College of Surgeons Committee on Trauma (2018) Advanced Trauma Life Support (ATLS) student course manual, 10th edn. American College of Surgeons, Chicago
11. Konstantinides SV, Meyer G, Becattini C et al (2019) 2019 ESC guidelines for the diagnosis and management of acute pulmonary embolism. *Eur Respir J* 54:1901647. <https://doi.org/10.1183/13993003.01647-2019>
12. Evans L, Rhodes A, Alhazzani W et al (2021) Surviving Sepsis Campaign: international guidelines for management of sepsis and septic shock 2021. *Crit Care Med* 49:e1063-e1143. <https://doi.org/10.1097/CCM.0000000000005337>

13. Garvey LH, Dewachter P, Hepner DL et al (2019) Management of suspected immediate perioperative allergic reactions: an international overview and consensus recommendations. *Br J Anaesth* 123:e50-e64. <https://doi.org/10.1016/j.bja.2019.04.044>
14. Simons FE, Ebisawa M, Sanchez-Borges M et al (2015) 2015 update of the evidence base: World Allergy Organization anaphylaxis guidelines. *World Allergy Organ J* 8:32. <https://doi.org/10.1186/s40413-015-0080-1>
15. Rüffert H, Bastian B, Bendixen D et al (2021) Consensus guidelines on perioperative management of malignant hyperthermia suspected or susceptible patients from the European Malignant Hyperthermia Group. *Br J Anaesth* 126:120-130. <https://doi.org/10.1016/j.bja.2020.09.029>
16. Malignant Hyperthermia Association of the United States (2026) Managing an MH crisis (treatment guideline). MHAUS, Rochester, WI. <https://www.mhaus.org>. Accessed 27 Jun 2026
17. Frerk C, Mitchell VS, McNarry AF et al (2015) Difficult Airway Society 2015 guidelines for management of unanticipated difficult intubation in adults. *Br J Anaesth* 115:827-848. <https://doi.org/10.1093/bja/aev371>
18. Panchal AR, Bartos JA, Cabañas JG et al (2020) Part 3: adult basic and advanced life support: 2020 American Heart Association guidelines for CPR and emergency cardiovascular care. *Circulation* 142:S366-S468. <https://doi.org/10.1161/CIR.0000000000000916>
19. Sandadi S, Johannigman JA, Wong VL, Blebea J, Altose MD, Hurd WW (2010) Recognition and management of major vessel injury during laparoscopy. *J Minim Invasive Gynecol* 17:692-702. <https://doi.org/10.1016/j.jmig.2010.06.005>
20. Lazic SE, Clarke-Williams CJ, Munafò MR (2018) What exactly is 'N' in cell culture and animal experiments? *PLoS Biol* 16:e2005282. <https://doi.org/10.1371/journal.pbio.2005282>
21. Aarts E, Verhage M, Veenfliet JV, Dolan CV, van der Sluis S (2014) A solution to dependency: using multilevel analysis to accommodate nested data. *Nat Neurosci* 17:491-496. <https://doi.org/10.1038/nn.3648>
22. Carpenter J, Bithell J (2000) Bootstrap confidence intervals: when, which, what? A practical guide for medical statisticians. *Stat Med* 19:1141-1164
23. Gallifant J, Afshar M, Ameen S et al (2025) The TRIPOD-LLM reporting guideline for studies using large language models. *Nat Med* 31:60-69. <https://doi.org/10.1038/s41591-024-03425-5>
24. Park SH, Suh CH, Lee JH et al (2025) Minimum reporting items for clear evaluation of accuracy reports of large language models in healthcare (MI-CLEAR-LLM): 2025 updates. *Korean J Radiol* 26:1123-1132. <https://doi.org/10.3348/kjr.2025.1522>
25. Sethi I, Khan S, Lyons PG et al (2026) Large language models provide accurate but potentially unsafe answers to multimodal critical care medicine board review questions. *Crit Care Med*. <https://doi.org/10.1097/CCM.0000000000007228>
26. Gulen D, Gozden HE, Ekin S, Ceylan I (2026) Structural error asymmetry and harm-weighted analysis of ChatGPT versus ICU physicians in acid-base interpretation. *Sci Rep* 16:15184. <https://doi.org/10.1038/s41598-026-44576-4>

27. Hurley NC, Corbitt R, Fallahian S, Saenz S, Hess AS, Schroeder KM (2026) Benchmarking large language model performance in perioperative crisis responses. *Anesth Analg*.
<https://doi.org/10.1213/ANE.00000000000007924>
28. Al-Azzam AAM, Alkhateeb Z, Shahin A et al (2026) Comparative evaluation of artificial intelligence chatbots for real-time guidance during intraoperative anesthetic crises: a simulation-based study. *Saudi J Anaesth* 20:362-368. https://doi.org/10.4103/sja.sja_903_25
29. Jarrassier A, de Saint-Léger FB, de Rocquigny G et al (2026) Can AI generate safe anaesthesia plans? A comparative evaluation of three large language models on 100 synthetic cases. *Anaesth Crit Care Pain Med* 45:101769. <https://doi.org/10.1016/j.accpm.2026.101769>
30. Chen R, Warburton A, Do R, Chen DD, Katz D, Burnett GW (2026) Large-scale evaluation of five large language models in anesthesia decision-making for hip fracture surgery. *Anesth Analg*.
<https://doi.org/10.1213/ANE.00000000000008148>
31. De Vito A, Geremia N, Russo A, Colpani A, Vaira LA, Madeddu G (2026) Reproducibility of AI-generated antibiotic recommendations in standardized clinical scenarios: a proof-of-concept experimental study. *Int J Antimicrob Agents* 107881.
<https://doi.org/10.1016/j.ijantimicag.2026.107881>
32. Güler I, Kraus A, Grieb G, Stelling H (2026) AI-based burn image assessment: reliability and clinical error patterns of multimodal large language models in a repeated-inference study. *J Plast Reconstr Aesthet Surg*. <https://doi.org/10.1016/j.bjps.2026.05.070>
33. Landon S, Savage T, Greysen SR, Bressman E (2026) Variation in large language model recommendations in challenging inpatient management scenarios. *J Gen Intern Med* 41:1286-1291.
<https://doi.org/10.1007/s11606-025-09888-7>
34. Windisch P, Koechli C, Dennstädt F et al (2026) Is one run enough? Reproducibility of flagship large language models across temperature and reasoning settings in biomedical text processing. *J Am Med Inform Assoc* 33:1179-1184. <https://doi.org/10.1093/jamia/ocag039>

Figures

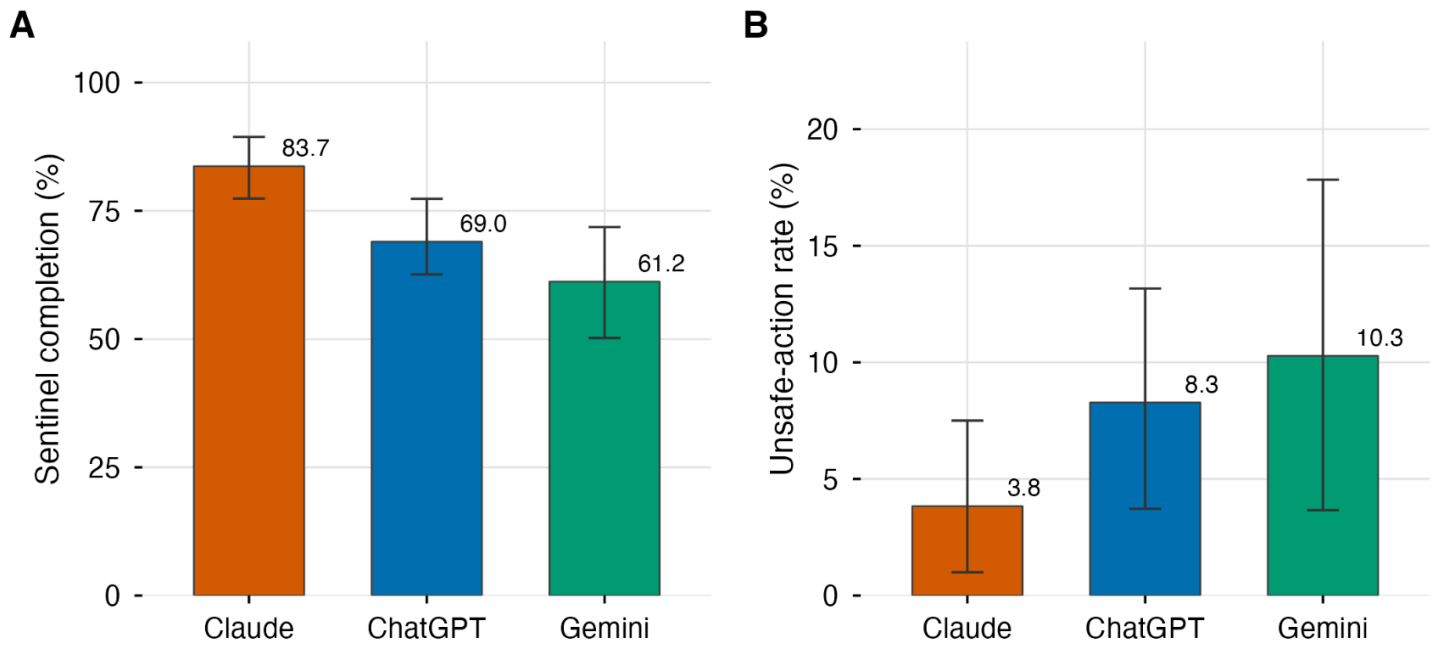


Figure 1

Positive-sentinel completion and unsafe-action rates for each model. (A) Sentinel completion. (B) Unsafe-action rate. Bars show the mean of scenario-level means across double-scored runs. Error bars are 95% cluster-bootstrap confidence intervals resampling the ten scenarios.

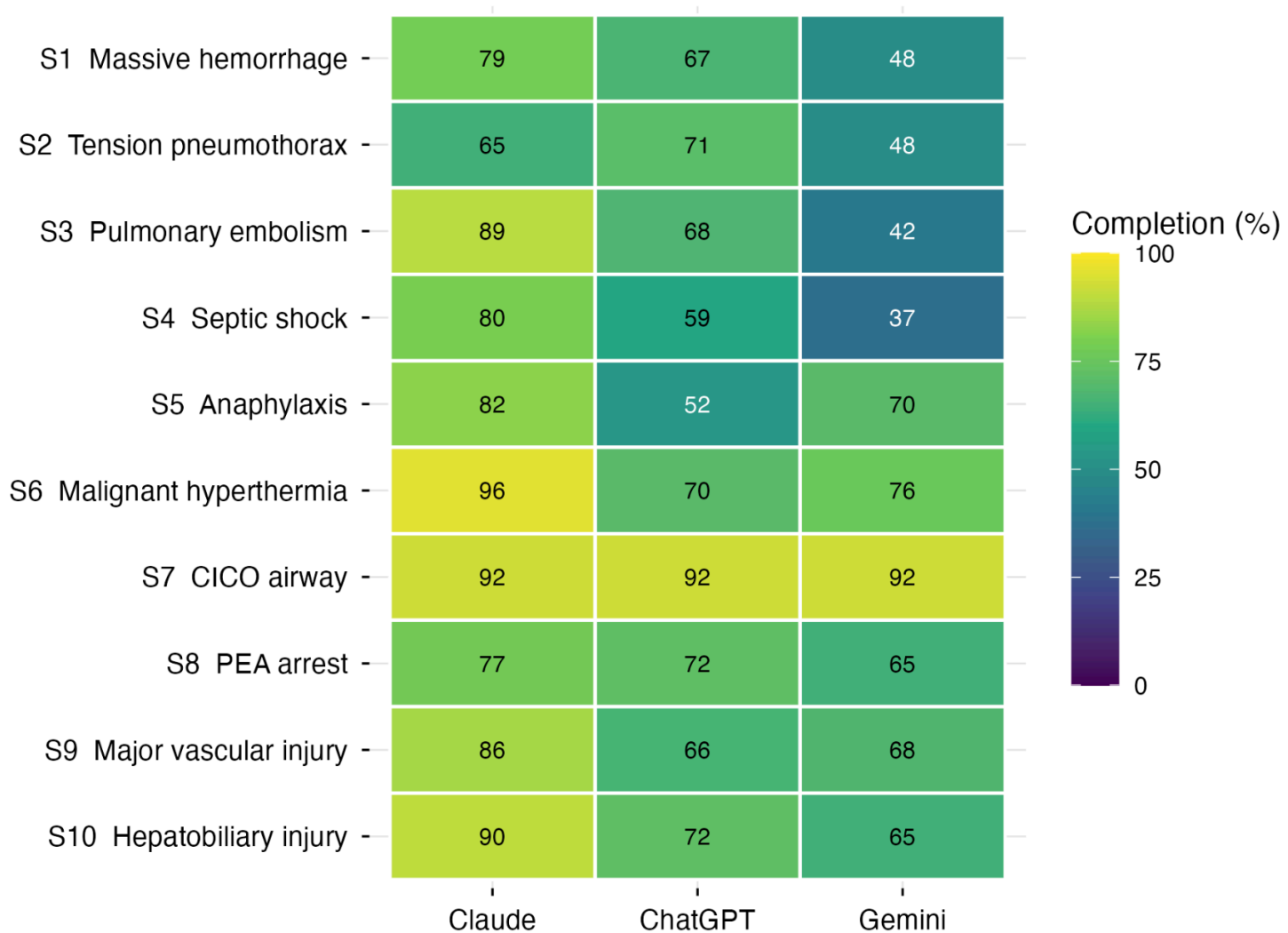


Figure 2

Scenario-level completion. Positive-sentinel completion for each model across the ten crisis scenarios, showing variation both between models and across scenarios.

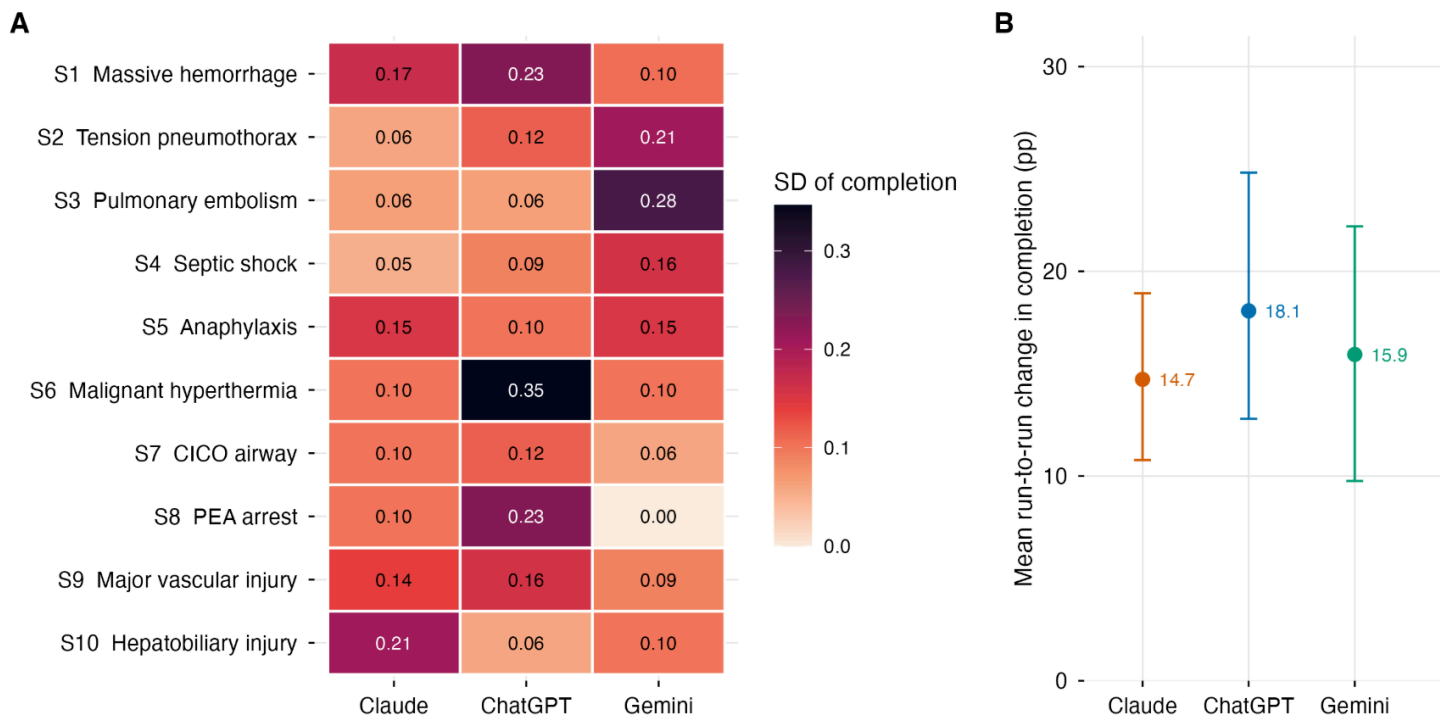


Figure 3

Run-to-run instability. (A) Standard deviation of per-run completion across the three runs of each model-scenario combination. (B) Mean absolute change in completion between repeated runs of the same model, with 95% cluster-bootstrap confidence intervals resampling the ten scenarios. The overlapping intervals indicate the models cannot be ranked by stability.

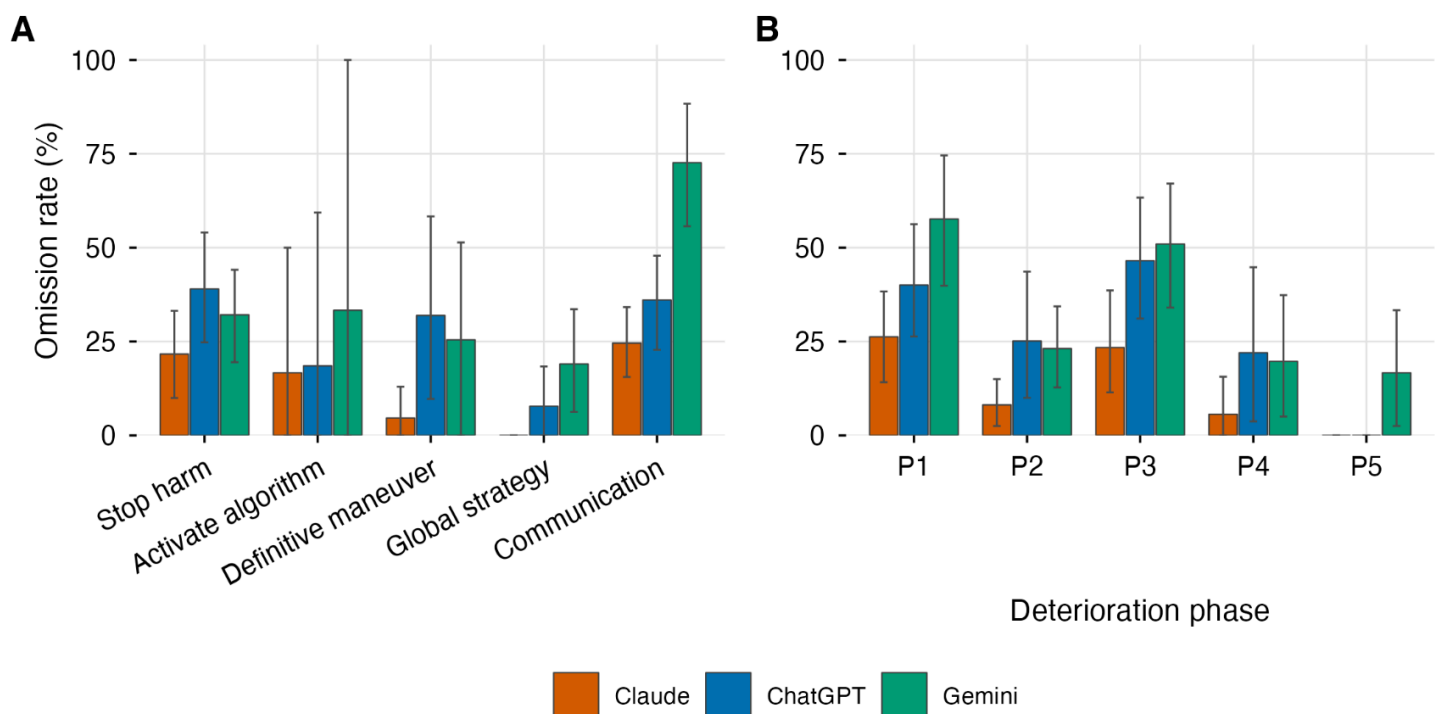


Figure 4

Failure taxonomy. (A) Positive-sentinel omission rate by cross-crisis safety axis for each model. (B) Omission rate by deterioration phase. Error bars are 95% cluster-bootstrap confidence intervals. The activate-algorithm axis rests on six to seven observations per model and is shown as directional.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [llmsafetybenchmarkTRIPODLLMchecklist.docx](#)
- [llmsafetybenchmarksupplementarymaterial.docx](#)