Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

AI safety evaluation in an underrepresented population: real-world performance of clinical decision support and frontier language models on Medicaid patient messaging triage

Sanjay Basu

sanjay.basu@ucsf.edu

University of California, San Francisco

Sadiq Patel

University of Pennsylvania

Parth Sheth

University of Pennsylvania

Bernardo Arevalo

Waymark

Jeremy Schifberg

Waymark

John Morgan

Virginia Commonwealth University

Rajaie Batniji

Waymark

Research Article

Keywords: clinical decision support, patient safety, large language models, primary care, Medicaid, text messaging triage, reinforcement learning, retrieval-augmented generation, federally qualified health centers, deployment validation

Posted Date: June 11th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9796014/v1

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

Additional Declarations: Competing interest reported. All authors are employees at Waymark, a public benefit organization that provides free social and medical services to patients receiving Medicaid. The authors test artificial intelligence tools to improve patient care but do not sell artificial intelligence platforms. All employees have access to artificial intelligence chatbots such as ChatGPT, Claude, and Gemini in the normal course of their work, but Waymark has no sales relationship with the entities owning these platforms.

Version of Record: A version of this preprint was published at BMC Medical Informatics and Decision Making on August 18th, 2026. See the published version at https://doi.org/10.1186/s12911-026-03763-z.

AI safety evaluation in an underrepresented population: real-world performance of clinical decision support and frontier language models on Medicaid patient messaging triage

Authors

Sanjay Basu, MD, PhD[1,2]; Sadiq Y. Patel, MSW, PhD[3]; Parth Sheth, MSE[3]; Bernardo Arevalo[2]; Jeremy Schifberg[2]; John Morgan, MD[4]; Rajaie Batniji, MD, PhD[2,5]

Affiliations

[1] University of California, San Francisco, San Francisco, CA, USA [2] Waymark, San Francisco, CA, USA [3] University of Pennsylvania, Philadelphia, PA, USA [4] Virginia Commonwealth University, Richmond, VA, USA [5] Stanford University, Stanford, CA, USA

Corresponding author

Sanjay Basu, MD, PhD University of California, San Francisco and Waymark 2021 Fillmore St San Francisco, CA 94115 sanjay.basu@ucsf.edu

---

Abstract

Background. Studies of artificial intelligence tools used in patient triage have largely involved academic medical center cohorts, scripted patient-actor scenarios, or knowledge benchmarks. Populations that may rely on such tools due to constrained access to in-person care, including Medicaid patients, have been less fully evaluated.

Objective. To compare combinations of safety guardrails added to artificial intelligence tools for triage of patient-initiated text messages in a multi-state Medicaid population.

Methods. Retrospective evaluation of 2000 messages from Medicaid patients across three U.S. states (Virginia, Washington, Ohio) during January 2023 through November 2025. Three physicians independently adjudicated each message under blinded review; disagreements were resolved by majority and a senior-physician arbiter. Tools included a deployed decision support system with a Conservative Q-Learning controller, supervised baselines (XGBoost+sentence-BERT, logistic regression, rule-based guardrails), and frontier large language models (Claude Opus 4.7, GPT-5.5, Gemini 3.1 Pro, each with and

without retrieval-augmented generation). Combinations spanned single tools, ensembles, cascades, multi-model consensus rules, and four-component guardrail compositions identified by a structured literature review. Thresholds and Platt calibration were fit on a held-out validation split and frozen before the test pass. Two pre-specified targets: an autonomous (physician-unassisted) triage benchmark of sensitivity and specificity both 0.80 or higher; and a sensitivity-floor target of 0.80 to 0.95 with clinician review of flagged messages.

Results. Real-world messages had lower reading level (grade 4.590 versus 5.810) and more colloquialisms (59.1% versus 19.5%) than physician-scripted scenarios. Hazard prevalence on blinded physician review was 8.2% (165 of 2000). Two configurations met the sensitivity-floor target: a high-recall first-stage screen (sensitivity 0.855; 12.0 missed hazards and 729 alerts per 1,000 messages) and a disagreement-stratified clinician-review workflow (sensitivity 0.939; 5.0 missed hazards and 878 alerts per 1,000 messages). No configuration met the autonomous benchmark; sensitivity and specificity reached 0.594 and 0.592 for the best balanced single tool, 0.685 and 0.575 for the best balanced ensemble, and 0.297 and 0.874 for the highest-specificity cascade.

Conclusions. No evaluated tool or combination was sufficiently accurate to enable physician-unassisted triage in this setting. Two configurations met the pre-specified sensitivity-floor target under clinician review of flagged messages, following the classical clinical-screening pattern.

Keywords: clinical decision support; patient safety; large language models; primary care; Medicaid; text messaging triage; reinforcement learning; retrieval-augmented generation; federally qualified health centers; deployment validation.

Introduction

Patients now contact their primary care teams by written message in addition to visits and phone calls. The volume of messages sent through patient portals and through secure text-message lines to primary care teams rose during and after the 2020 to 2021 expansion of remote care [1]. Primary care appointments have at the same time become harder to schedule, and the wait is longer in lower-income neighborhoods [2]. More clinical questions therefore reach the practice as a written message rather than as a visit or call. Messages are typically reviewed first by medical assistants or care coordinators before being routed to a physician [3]. The physician inbox already holds laboratory results, refill requests, and system notifications, and the added message volume is associated with reductions in physician well-being [4]. The backlog concentrates at federally qualified health centers and other safety-net practices, which serve counties with documented

primary care provider shortages [5]. Support staff who perform initial message triage are frequently in short supply at these practices. Artificial intelligence tools that scan incoming messages and flag clinical emergencies or urgent situations are of growing interest as a way to surface time-sensitive cases to clinicians without adding triage staff.

Artificial intelligence adapted to clinical triage has accelerated in scope and capability. The Articulate Medical Intelligence Explorer (AMIE) program evaluated a frontier large language model on synchronous text-chat consultations with standardized patient-actors [6], on accurate differential diagnosis from New England Journal of Medicine clinicopathologic case reports [7], and most recently in a 2026 multimodal demonstration on 105 scripted patient-actor scenarios across 32 non-multimodal and 9 multimodal qualitative axes rated by 18 specialists [8]. The MedHELM benchmark suite [9] catalogued frontier-model performance on knowledge-intensive medical tasks. Brodeur and colleagues evaluated a frontier language model on the reasoning tasks of an emergency-department physician using raw electronic health record text [10], and Ramaswamy and colleagues stress-tested a consumer-facing triage chatbot on 60 clinician-authored vignettes across 16 factorial conditions, observing 52% under-triage on clinical emergencies [11]. Recent patient-portal-message triage work by Liu and colleagues evaluated retrieval-augmented large language models on academic medical center portal correspondence at a single quaternary-care institution [12]. A natural next extension is evaluation in populations that may increasingly rely on artificial intelligence-assisted triage because traditional in-person care access is constrained, that communicate in colloquial or abbreviated text, that carry a higher comorbidity and behavioral-health burden, and that have higher baseline rates of medical mistrust and healthcare-interaction stigma [13] than the cohorts studied to date. To the authors' knowledge, evaluation on Medicaid patient asynchronous text-based messaging (whether sent by short-message service to a secure line or as a patient-portal message) at scale has not been reported. Safety characteristics established on academic medical center cohorts and physician-scripted scenarios cannot be assumed to transfer to populations with these distinct linguistic and clinical profiles.

Patient-triage systems described in the published literature use different combinations of artificial intelligence tools: rule-based guardrails, supervised classifiers on bag-of-words or sentence-embedding features, reinforcement-learning controllers over hazard-category probability vectors with asymmetric reward structures, and prompted frontier large language models with safety-augmented prompts and optional retrieval-augmented generation [14]. More recent proposals add multi-model consensus rules, threshold-optimized cascades, and multi-component guardrail compositions drawn from current health artificial intelligence assurance frameworks. The relative safety performance of these tools and combinations on a Medicaid asynchronous text-messaging triage population has not been characterized head-to-head.

This study compares combinations of safety guardrails added to widely-used artificial intelligence tools for triage of patient-initiated text messages in a multi-state Medicaid managed care population. Two contributions are pursued. First, we characterize a previously underrepresented population (Medicaid patients across three U.S. states whose linguistic patterns differ from academic medical center cohorts and physician-scripted scenarios) and report the per-tool safety performance observed on this population. Second, we report the head-to-head performance of two current-generation frontier large language models (Claude Opus 4.7 and Gemini 3.1 Pro Preview), configured with safety-augmented system prompts, against the deployed decision support architecture and classical supervised baselines, on the same patient messages and against the same blinded-physician-adjudicated reference labels.

Methods

Study design and population

This was a retrospective evaluation on text messages from 1679 Medicaid managed care patients enrolled in a multi-state clinically integrated network operating across three U.S. states (Virginia, Washington, and Ohio), submitted between January 2023 through November 2025. 2000 messages comprised the held-out test set; 1280 messages constituted the development set, which was further partitioned into a 570-message training subset for model fitting and a 710-message validation subset for threshold selection and Platt calibration. The test set, the training subset, and the validation subset were drawn from distinct messages. Patient-level overlap between training and test sets was 41 of 1,679 unique test-set patients (2.4%), reported transparently with a patient-disjoint sensitivity analysis in the Multimedia Appendix.

Patient eligibility criteria: receiving at least one outpatient visit from a clinician in the Medicaid clinical entity in Virginia, Washington, or Ohio; aged 18 years or older; submitted at least one text message to clinicians who are part of the clinical network.

Message curation and inclusion criteria

Messages were drawn from the full corpus of patient-initiated text correspondence to the clinical network during the study period. Messages were not short single-word greetings (“hi”, “hello”) but full patient inquiries comparable in length and content to patient-portal messages or short clinical emails (median length 38 words, interquartile range 18–72 words). Messages were screened to exclude administrative-only content (appointment scheduling, billing-only, prescription-refill-only) and to retain messages containing at least

one clinical complaint, symptom description, or care-coordination request. The screening protocol is detailed in Multimedia Appendix 1.

Reference labeling and adjudication

Each message was independently reviewed by three physicians from the participating multi-state clinically integrated network. Reviewers were blinded to all artificial intelligence predictions throughout adjudication, and to one another's labels during the initial review pass. For each message, each reviewer labeled (a) hazard presence (binary), (b) hazard category (from a pre-specified list of clinical categories spanning behavioral-health, substance-use, obstetric, pediatric, cardiac, neurological, metabolic, and environmental safety hazards), and (c) recommended triage action on the 8-point ordinal action scale (Action 1 = self-care; Action 8 = emergency services). Disagreements were resolved in two stages: first by majority vote (two of three reviewers agreeing constituted the reference label), and second by senior-physician arbiter review for residual splits. The two-stage majority-then-arbiter design was selected over a forced-unanimity design to preserve labels on genuinely ambiguous cases that nevertheless have a defensible reference action. The arbiter-resolved subset is identified in the released per-message prediction file for any reviewer who wishes to conduct a sensitivity analysis restricted to unanimously-labeled cases. Initial inter-rater agreement across the three independent reviewers (before majority and arbiter resolution) was substantial: Fleiss kappa was 0.82 for hazard presence (the primary endpoint) and 0.76 for hazard category. The complete labeling protocol is reported in Multimedia Appendix 1.

Missing demographic data and equity analyses

Patient-level demographic data (gender; self-reported race) were available for 1,427 of 2,000 test-set messages (71.3%). 573 messages (28.6%) lacked both demographic fields. Hazard prevalence differed between the two groups (9.04% in messages with complete demographics versus 6.28% in messages with missing demographics), indicating that the missingness is not independent of the outcome. Equity-stratified sensitivity analyses (gender; race) were conducted as complete-case analyses restricted to messages with non-missing values for the stratifying variable, and the analyses are framed as descriptive subgroup performance characterizations rather than as confirmatory equity claims.

Sensitivity to the missing-demographic-group assumption was characterized for each architecture by Manski-style partial identification bounds on overall sensitivity. The bounds were constructed by computing the architecture's observed sensitivity in the complete-demographic subgroup and then bracketing the overall sensitivity that would be observed if every hazard in the missing-demographic subgroup were correctly detected (optimistic bound) or if no hazards in that subgroup were detected (conservative bound). For the high-sensitivity classifier underlying Policy A (logistic regression with TF-IDF

features), the bounds are 0.758 (conservative) to 0.976 (optimistic), and the observed overall sensitivity (0.970) sits at the upper edge of the bound interval. Bounds for the remaining evaluated architectures are reported in Multimedia Appendix 1.

Sample size justification

The 2000-message held-out test set was sized to detect a between-architecture absolute sensitivity difference of 0.05 with 80% power at $\alpha = 0.05$ (two-sided), under the planned-prevalence assumption of 8% hazard prevalence (observed prevalence 8.2%) and the planned-comparator-sensitivity assumption of 0.80. Under these parameters, the minimum hazard-positive subgroup required for the primary McNemar paired comparison was approximately 150 hazards; the realized hazard-positive subgroup was 165.

Linguistic features

Three linguistic features were computed for every message. Flesch-Kincaid reading-level grade was computed using the textstat Python library (v0.7.0) on the raw message text after light pre-processing to strip non-printable characters and normalize whitespace. Colloquialism rate was computed as the proportion of words appearing in a 847-term colloquial-English lexicon seeded from published informal-language corpora and extended with Medicaid messaging-specific colloquialisms identified during the development-set labeling pass; the full lexicon and seed sources are released with the code repository. Abbreviation rate was computed by a rule-based detector requiring tokens to be $\leq$4 characters, non-stopword, non-single-letter pronoun, and either appearing $\geq$20 times in the corpus or matching a 312-term curated list of informal health abbreviations. Implicit-contextual-reference rate was computed as the proportion of clauses whose antecedent or referent is implicit rather than stated, identified by a rule-based annotator using spaCy (v3.5) coreference and dependency parsing followed by manual validation of the rule outputs on a 200-message random sample. Manual validation on the 200-message random sample showed Cohen kappa 0.73 for colloquialism detection, 0.86 for abbreviation detection, and 0.64 for implicit context identification. Full implementation details are reported in Multimedia Appendix 1.

Outcomes

The primary endpoint was sensitivity for hazard detection at each architecture's calibrated operating threshold, with the 95% confidence interval computed by the Wilson score method [15].

Secondary endpoints: specificity (Wilson 95% confidence interval), positive predictive value, negative predictive value, F1 (bootstrap 95% confidence interval, 10,000 iterations, seed=42), Matthews correlation coefficient (bootstrap), area under the receiver operating

characteristic curve (Hanley-McNeil 95% confidence interval where calibrated probabilities were available), false negatives per 1,000 messages, action-recommendation appropriateness (3-category: appropriate, under-triage, over-triage on the 8-point action scale), equity-stratified sensitivity by sex and self-reported race/ethnicity.

Two operating-point targets were specified a priori. The first is the autonomous (physician-unassisted) triage benchmark, defined as sensitivity ≥ 0.80 AND specificity ≥ 0.80 simultaneously. The benchmark is anchored to the IDx-DR pivotal trial of an autonomous diabetic-retinopathy diagnostic system that received de novo authorization from the United States Food and Drug Administration as the first autonomous artificial intelligence diagnostic device cleared in any field of medicine, with pre-specified superiority endpoints of sensitivity $> 85\%$ and specificity $> 82.5\%$ and achieved values of sensitivity 87.2% (95% CI 81.8–91.2%) and specificity 90.7% (95% CI 88.3–92.7%) in the pivotal trial [16]. This target characterizes performance under hypothetical fully-autonomous deployment without clinician review. The second is a sensitivity-floor target with clinician review of flagged messages, defined as a sensitivity floor (≥ 0.80, ≥ 0.85, ≥ 0.90, and ≥ 0.95, all characterized in Table 6) at any specificity attainable on the operating curve, with the false-alarm volume routed to a clinician confirmatory review queue rather than blocking the workflow; this target instantiates the classical clinical-screening pattern (high-sensitivity screen followed by high-specificity confirmation) in which the cost of a missed hazard substantially exceeds the cost of a clinician review event. Both targets are reported.

Architectures evaluated

Nine architectural configurations were evaluated, spanning rule-based, classical supervised, deployed clinical decision support, and frontier large language model designs. All architectures were trained and calibrated on a 570-message development set and evaluated on the 2000-message held-out test set; no test-set messages were used for hyperparameter selection, threshold calibration, or model selection at any point. Architecture-specific details are reported below. The complete configuration matrix is summarized in Appendix Table S1.

1. Rule-based guardrails: 147 hand-crafted rules encoding symptom combinations, temporal keywords, and severity indicators. The architecture operates in two distinct thresholded stages. At Stage 1, each rule fires when a sentence-BERT [17, 18] cosine similarity to its exemplar embeddings exceeds 0.75 (the per-rule similarity gate that determines which rules contribute evidence). At Stage 2, the aggregated rule output is converted to a calibrated hazard probability and thresholded at the operating-point probability 0.648 selected on the training/validation split (the binary hazard decision applied at the architecture

level). The 0.75 cosine-similarity gate and the 0.648 calibrated-probability threshold refer to different stages of the pipeline and are not interchangeable.

2. Logistic regression with TF-IDF features: Logistic regression with term frequency-inverse document frequency (TF-IDF) features (1-3 gram, max 10,000 features, sublinear TF scaling). Threshold 0 calibrated on training/validation split.

3. XGBoost with sentence-BERT embeddings: XGBoost [19] classifier on 768-dimensional sentence-BERT embeddings. Hyperparameters tuned via 5-fold cross-validation on the training set with random_state=42; final threshold 0.

4. Constellation architecture: Multinomial logistic regression over 23 hazard categories on sentence-BERT embeddings, with the per-category probabilities aggregated into a single binary hazard probability via the maximum-activation rule. Threshold 0.965.

5. Conservative Q-Learning controller (sensitivity-optimized): Conservative Q-Learning [20] on calibrated probability vectors from an upstream sentence-BERT classifier. The 23-dimensional state space encodes hazard-category probabilities. The action space comprises 8 discrete escalation actions (Action 1 = self-care; Action 8 = emergency services). Asymmetric reward: +10 for correct action, -50 for missed hazard, -2 for false alarm. The sensitivity-optimized variant applies a calibrated probability threshold (1.000) chosen to maximize sensitivity subject to specificity ≥ 0.70 on the validation set.

6. Conservative Q-Learning controller (reward-optimized): Same trained Q-function as variant 5; predictions taken as the Q-function argmax over the 8 actions.

7. ActionHead: Independent sentence-BERT + multinomial logistic regression action classifier trained for 8-class action prediction on the 570-example training set.

8. Claude Opus 4.7 (safety-augmented): Anthropic's Claude Opus 4.7 large language model. System prompt explicitly prioritizes patient safety and emphasizes erring toward escalation. Temperature 0; deterministic.

9. Gemini 3.1 Pro Preview (safety-augmented): Google's Gemini 3.1 Pro Preview large language model. Same safety-augmented prompt structure as Claude; temperature 0.

Single-turn screening evaluation

All architectures were evaluated in single-turn mode: each architecture received one patient message as input and produced one hazard prediction and one action

recommendation. This corresponds to the "screening at message arrival" clinical use case, the decision the system makes upon receipt of an incoming message before any clarifying conversation occurs. Multi-turn dialogue extensions are out of scope for this analysis (see Limitations) and identified as future work.

Comparability across architectures

The deployed decision support architecture (Conservative Q-Learning controller, constellation classifier, action recommender) and the classical supervised baselines (XGBoost with sentence-BERT, logistic regression with TF-IDF, rule-based guardrails) were fit on the 1,280-message development set with physician-adjudicated labels. The frontier large language models (Claude Opus 4.7, GPT-5.5, Gemini 3.1 Pro Preview) were evaluated zero-shot under a safety-augmented system prompt without parameter updates. The retrieval-augmented variants of each frontier model additionally received eight stratified labeled neighbors from the same development corpus in-context but did not update model weights. Head-to-head differences between the supervised pipeline and the frontier-LLM configurations therefore confound algorithm class with training-data exposure. The retrieval-augmented frontier-LLM configurations are the closest matched comparators, providing both in-context labeled examples and identical zero-shot model weights to the bare safety-augmented variants. The relative within-pipeline rankings (deployed Conservative Q-Learning controller vs supervised baselines, and safety-augmented vs retrieval-augmented within each frontier-LLM vendor) are interpretable independent of this confound.

Sensitivity analyses

Three additional analyses were specified a priori as sensitivity extensions of the primary head-to-head comparison.

First, multi-model consensus rules were computed from the existing per-message frontier-model predictions: Claude-only, Gemini-only, OR-rule of either flagging, and AND-rule of both flagging.

Second, retrieval-augmented generation was added to the safety-augmented configuration of each of the three frontier large language models (Claude Opus 4.7, GPT-5.5, and Gemini 3.1 Pro Preview). At inference, the top-eight nearest-neighbor labeled examples from the 1,280-example development corpus, retrieved by sentence-BERT cosine similarity and stratified by class, were supplied to each model as in-context evidence.

Third, a four-component guardrail composition was operationalized as a synthesis of three current health AI assurance frameworks: the assurance-laboratory evaluation framework of Shah and colleagues [21], the principles articulated in the Coalition for Health AI

Blueprint for Trustworthy AI Implementation Guidance and Assurance for Healthcare [22], and the organizational governance approach of the Health AI Partnership community of practice [23]. The four components were: (1) a clinical-safety system prompt at the language-model layer; (2) a content-moderation filter pre-screen at the message-ingestion layer; (3) retrieval-augmented generation grounding the language model in labeled population-relevant exemplars; and (4) a learned hazard-classifier pre-screen filtering candidate messages before language-model evaluation. Components 1 and 3 were realized as the safety-augmented Claude Opus 4.7 with retrieval-augmented generation. Component 2 was modeled as a pass-through on the clinically-themed test-set messages, since in deployment the component would gate at message ingestion and would not exclude any messages in the present test set. Component 4 was realized as one of seven evaluated supervised classifiers, with the gate and language model composed under an AND-rule cascade requiring concordance between the two stages.

Two operating-point analyses for use with clinician review of flagged messages were also specified a priori. The first was a sensitivity-floor analysis: across the full pool of evaluated configurations (single tools at every observed threshold, ensembles, cascades, multi-model consensus rules, and retrieval-augmented generation), the configuration achieving the maximum specificity at each of four sensitivity floors (0.80, 0.85, 0.90, 0.95) was identified. The second was a disagreement-stratified workflow: the count of tools flagging each message was treated as an uncertainty signal, with messages flagged by zero or one of the ten tools considered candidates for autonomous-benign disposition and messages flagged by two or more routed to clinician review.

Statistical analysis

All paired comparisons used McNemar's test [24] restricted to hazard-positive cases, with Hochberg step-up correction [25] for family-wise error rate control across the 78 pairwise comparisons.

Cross-set sensitivity comparisons (physician-scripted versus real-world test sets) used a parametric bootstrap with 10,000 iterations and seed 42. In each iteration, sensitivities for both sets were independently resampled from binomial distributions parameterized by the observed proportions and sample sizes.

Stratified equity analyses used the same Wilson confidence intervals within each subgroup, with the equalized odds difference (maximum minus minimum subgroup sensitivity) reported.

Reporting framework and reproducibility

The study follows TRIPOD+AI [26] and DECIDE-AI [27] reporting guidelines. The complete TRIPOD+AI checklist appears in Appendix A.

Per-message predictions for every architecture × every test record are retained by the corresponding author and available on request for verification. Patient message text remains under HIPAA data-use agreement and is available to bona-fide researchers on request to the corresponding author.

The code repository at https://github.com/sanjaybasu/rl-llm-safety contains the complete pipeline as deployed: a single orchestrated entry point that, in one deterministic run, produces every per-message prediction file and every reported metric.

De-identification and adjudicator role

Patient messages were de-identified by HIPAA Safe Harbor methodology prior to analysis. No identifiable patient data or direct verbatim quotations appear in any manuscript figure, table, or supplementary file. Reviewing physicians performed adjudication as part of their salaried clinical-network duties and received no additional study-specific compensation. The IRB approval and consent waiver for this study are reported in the Declarations section.

Results

Population characteristics

The 2000 held-out test messages were submitted by 1679 unique Medicaid patients during January 2023 through November 2025. Hazard prevalence in the test set was 8.2% (165 hazards across 2000 messages), with the remaining 1835 messages adjudicated as benign by all three reviewing physicians. Population characteristics, including age, sex, and self-reported race/ethnicity, are reported in Table 1. The linguistic profile of the test set departed substantially from the profile of physician-scripted comparison scenarios on two of four measured dimensions and in the opposite direction on the other two. Real-world messages were written at a substantially lower reading level than physician-scripted scenarios (mean Flesch-Kincaid grade 4.590 versus 5.810) and contained colloquialisms at substantially higher rates (59.1% versus 19.5%); the abbreviation rate (24.5% versus 36.6%) and the implicit context rate (7.8% versus 29.3%) were lower in the real-world Medicaid messaging population than in the physician-scripted scenarios, reflecting that physician scenarios were constructed with deliberate medical shorthand and inter-sentence clinical context that real-world Medicaid messages use less frequently. The

composition of hazard categories (including behavioral-health symptoms, substance-use escalation, obstetric concerns, pediatric overdose, and metabolic emergencies) reflects the natural case mix of a Medicaid managed care messaging service rather than the curated distribution of standardized benchmarks.

Hazard detection performance

Table 2 reports per-architecture sensitivity, specificity, positive predictive value, negative predictive value, F1, Matthews correlation coefficient, area under the receiver operating characteristic curve, and false negatives per 1,000 messages on the 2000-message held-out test set.

Operating-point analysis

Table 3 reports sensitivity at matched specificity targets for each architecture for which calibrated probability output is available. Both the target specificity (the operating point at which each architecture is requested to operate) and the achieved specificity (the actual specificity observed at the threshold required to attain the target) are displayed alongside the sensitivity attained at that operating point. By construction, the rows in Table 3 are derived from the same per-message prediction files as Table 2, with no architecture's operating curve drawn from a different prediction run. Receiver operating characteristic monotonicity (sensitivity non-decreasing as specificity decreases) is verified for every architecture with calibrated probability output as a pre-submission pipeline assertion.

Paired comparisons

Figure 1 visualizes paired sensitivity differences between the deployed clinical decision support architecture variants and the frontier large language model comparators on the held-out test set. Each pairwise comparison was tested with McNemar's test [24] restricted to hazard-positive cases, with Hochberg step-up correction [25] applied across the 78 pairwise comparisons. The full pairwise discordant-pair matrix is reported in Appendix Table S5.

Action recommendation appropriateness

Figure 2 displays the per-architecture proportion of recommendations classified as appropriate, under-triage (recommending a less urgent action than the physician-adjudicated reference), or over-triage (recommending a more urgent action than the reference), on the 8-point ordinal action scale. Appendix Table S7 reports the underlying counts, the reference action distribution, and the agreement structure between each architecture and the reference.

Equity-stratified analysis

Sensitivity stratified by sex and by self-reported race/ethnicity is reported with Wilson 95% confidence intervals [15] within each subgroup, with the equalized odds difference (maximum minus minimum subgroup sensitivity) reported per architecture. Subgroups with fewer than 30 hazard-positive cases are flagged as insufficiently powered for reliable subgroup inference and are reported with explicit small-sample warnings rather than suppressed.

Threshold-optimized single-architecture analysis

To distinguish discriminative-capacity limitations from threshold-calibration artifacts, the operating curve of every architecture with calibrated probability output was swept and the F1-maximizing and Matthews-correlation-coefficient-maximizing operating points were extracted (Table 4). Across all six architectures with calibrated probabilities, the maximum F1 attained by any single architecture at any threshold was 0.20 (XGBoost with sentence-BERT, F1 0.201 at threshold-tuned sensitivity 0.412 and specificity 0.759), and the maximum Matthews correlation coefficient was 0.11 (constellation classifier, 0.109). At no operating point on any architecture's receiver operating characteristic curve was sensitivity ≥ 0.80 AND specificity ≥ 0.80 simultaneously achieved (Table 4). The architecture with the highest balanced operating point was XGBoost with sentence-BERT at sensitivity 0.594 and specificity 0.592, the point closest to the autonomous (physician-unassisted) triage benchmark. This remains well below the autonomous-deployment benchmark used here (sensitivity 0.80 or higher and specificity 0.80 or higher), which is anchored to the IDx-DR autonomous-diagnostic-AI pivotal trial [16]. The Conservative Q-Learning controller variants are bounded by the same underlying receiver operating characteristic envelope: their separation in Table 2 reflects threshold calibration on the asymmetric reward design rather than fundamentally different discriminative capacity.

Ensemble architectures and cascade architectures

Two architectural-combination strategies were evaluated against the autonomous (physician-unassisted) triage benchmark. First, 217 ensemble configurations were enumerated, combining all 13 architectures via hard-voting k-of-13 rules (k from 1 to 13), soft-voting at 101 thresholds with unweighted and F1-weighted means of the calibrated probability outputs, and top-three AND/OR rules. None achieved sensitivity 0.80 or higher and specificity 0.80 or higher simultaneously. The best balanced ensemble was the hard-voting three-of-13 rule (sensitivity 0.685, specificity 0.575, F1 0.214). The best F1-weighted soft-voting configuration reached sensitivity 0.582 and specificity 0.581 at the threshold-swept optimum. Second, all 156 ordered two-stage AND-rule cascade configurations were enumerated, with the Pareto frontier reported in Table 5 and the full

matrix in Appendix Table S3. The cascade pairing the Conservative Q-Learning controller (reward-optimized) with the safety-augmented Claude Opus 4.7 attained sensitivity 0.297 and specificity 0.874, improving specificity at the cost of sensitivity as expected for AND-rule cascades. The cascade pairing the action recommender with the logistic regression screen attained the highest cascade F1 (0.236) and highest Matthews correlation coefficient (0.154). No cascade configuration reached the autonomous (physician-unassisted) triage benchmark. The architectural-combination findings extend the single-architecture result: neither ensembling nor cascading the evaluated architectures closes the population gap. Figure 3 visualizes the joint distribution of sensitivity and specificity across all evaluated strategies (single architectures, ensembles, cascades, and multi-model consensus rules); no marker reaches the upper-right benchmark region defined by sensitivity ≥ 0.80 and specificity ≥ 0.80.

Sensitivity analyses: multi-model consensus, retrieval-augmented generation, and four-component guardrail composition

Two additional closing-the-gap interventions identified by a structured literature review of 2024-2026 clinical artificial intelligence safety evaluations were evaluated against the autonomous (physician-unassisted) triage benchmark. First, multi-model consensus rules (Claude Opus 4.7 only, Gemini 3.1 Pro Preview only, OR-rule of either flagging, AND-rule of both flagging) were computed from the existing per-message predictions. The OR-rule consensus attained sensitivity 0.297 and specificity 0.869, virtually identical to Claude alone because the Gemini-positive set was nearly a subset of the Claude-positive set on this population. The AND-rule consensus attained sensitivity 0.018 and specificity 0.996, dominated by Gemini's low single-model sensitivity. No multi-model consensus rule reached the autonomous (physician-unassisted) triage benchmark.

Second, retrieval-augmented generation over the 1,280-example labeled training corpus was evaluated as a separate closing-the-gap intervention across all three frontier large language model providers (Claude Opus 4.7, GPT-5.5, and Gemini 3.1 Pro Preview). Each model received the top-eight stratified-by-class nearest-neighbor training examples as in-context evidence at inference time, retrieved by sentence-BERT cosine similarity. All three providers showed a consistent dissociation between the physician-scripted 41-case comparison set and the real-world Medicaid messaging test set. On the physician-scripted set, retrieval-augmented generation raised sensitivity above the safety-augmented baseline for every vendor (Claude Opus 4.7: 0.815 to 1.000; GPT-5.5: 0.852 to 1.000; Gemini 3.1 Pro: 0.296 to 0.741), with the corresponding F1 lifts (Claude 0.800 to 0.900; GPT-5.5 0.821 to 0.947; Gemini 0.432 to 0.851) confirming the published RAG benefit on prototypically-presented clinical scenarios [12]. On the real-world Medicaid messaging set, the same retrieval-augmented configurations attained sensitivity 0.218 (Claude, down from 0.297 without retrieval), 0.261 (GPT-5.5, down from 0.321), and 0.024 (Gemini, essentially

unchanged from 0.018). Retrieval-augmented generation closed the gap on prototypically-presented physician scenarios for all three vendors but failed to close the gap on the Medicaid messaging population for any vendor. One interpretation is that retrieval over the 1,280-example labeled training corpus returns informative neighbors when the test message resembles training-set patterns (the physician-scripted scenarios were constructed to be representative of canonical hazard presentations), but returns lower-information or distribution-mismatched neighbors when the test message exhibits the colloquialism-rich, abbreviation-laden, implicit-context patterns characteristic of the real-world Medicaid messaging population. The cross-vendor replication of this dissociation strengthens the interpretation by ruling out a single-vendor model-specific artifact. The result is consistent with the literature-review finding that no published 2024-2026 study has demonstrated sensitivity ≥ 0.80 AND specificity ≥ 0.80 for free-text hazard detection on a Medicaid or low-literacy population.

Table S8 synthesizes the best operating points achievable under each strategy and reports whether any strategy reached the autonomous (physician-unassisted) triage benchmark on the real-world test set.

Third, a four-component guardrail composition was operationalized, synthesizing the assurance-laboratory evaluation framework of Shah and colleagues [21], the principles articulated in the Coalition for Health AI Blueprint for Trustworthy AI Implementation Guidance and Assurance for Healthcare [22], and the organizational governance approach of the Health AI Partnership community of practice [23]. The four components were: (1) a clinical-safety system prompt at the large-language-model layer; (2) a content-moderation filter pre-screen at the message-ingestion layer; (3) retrieval-augmented generation grounding the language model in labeled population-relevant exemplars; and (4) a learned hazard-classifier pre-screen filtering candidate messages before language-model evaluation. The above three sources articulate, respectively, an evaluation/assurance framework, a principles-level implementation guidance, and an organizational community-of-practice governance framework rather than a specific technical guardrail architecture; the four-component composition tested here is the authors' operational synthesis intended to instantiate those frameworks at the inference-pipeline layer, and is offered as one defensible technical realization rather than as a configuration uniformly prescribed by any cited source. Within this composition, components 1 and 3 are realized as the safety-augmented Claude Opus 4.7 with retrieval-augmented generation; component 2 is modeled as a pass-through on the clinically-themed test-set messages (in deployment, the component would gate at message ingestion and would not exclude any messages in the present test set); and component 4 is realized as one of seven evaluated supervised classifiers, with the gate-and-language-model pair composed under the AND-rule cascade common in clinical computer-aided detection (Table S9). Across all seven evaluated learned-classifier gates, no configuration of this operational synthesis attained sensitivity $\geq$

0.80 AND specificity ≥ 0.80 on the real-world test set; the operating-point characteristic was dominated by the second-stage language model's sensitivity floor, with the learned-classifier gate further reducing sensitivity by demanding cross-stage concordance. The interpretation is consistent with the literature-review finding that no published study has demonstrated sensitivity 0.80 or higher and specificity 0.80 or higher for free-text hazard detection on a Medicaid or low-literacy population. The multi-component guardrail compositions described in the assurance and governance literature appear to require population-relevant extension before they reach the autonomous (physician-unassisted) triage benchmark on this specific deployment population. Such extension is required particularly for the retrieval corpus, the learned-classifier training data, and the calibration of the operating-point trade-off.

Operating-point analysis under a sensitivity-floor target

The autonomous (physician-unassisted) triage benchmark (sensitivity 0.80 or higher and specificity 0.80 or higher simultaneously) was not reached by any evaluated configuration on the real-world Medicaid messaging population. The complementary question, what is achievable at a sensitivity-floor target appropriate for use with clinician review of flagged messages, was characterized by two analyses (Table 6). First, a sensitivity-floor analysis swept the full pool of evaluated configurations (single architectures with calibrated probabilities at all observed thresholds, ensembles, cascades, multi-model consensus rules, and retrieval-augmented generation) and identified the configuration achieving the maximum specificity at each of four sensitivity floors (≥ 0.80, ≥ 0.85, ≥ 0.90, ≥ 0.95). The configuration meeting sensitivity 0.85 or higher with the highest achievable specificity was a threshold-tuned logistic regression with TF-IDF features, attaining sensitivity 0.855 and specificity 0.282. Operationally, this is a high-recall first-stage screen that forwards approximately 72% of test-set messages to second-stage clinician review. Second, a disagreement-stratified clinician-review workflow was characterized by treating the count of architectures flagging each message as an uncertainty signal: messages flagged by zero or one of the ten evaluated architectures (253 of 2,000 messages; 12.7% of the test set) had a hazard prevalence of 4.0% in this stratum and are candidates for autonomous-benign disposition; messages flagged by two or more architectures (1,747 of 2,000; 87.3% of the test set) require clinician review and collectively capture 155 of 165 hazards (sensitivity 0.939). The empirical alert volume and residual miss rate per 1,000 messages for both policies, alongside the best balanced single-architecture, ensemble, and cascade reference points, are reported in Table 7. The disagreement-stratified workflow attains higher headline sensitivity than any single autonomous configuration but at the cost of a clinician review load that scales with 87% of the messaging volume. Both configurations meeting the pre-specified sensitivity-floor target are reported as operating points, the operationally-relevant operating pattern for safety-critical messaging triage with

asymmetric harm; the research agenda below addresses the residual gap to the autonomous (physician-unassisted) triage benchmark.

Discussion

Two principal findings emerge from this study. First, two operating points meeting the pre-specified sensitivity-floor target under clinician review of flagged messages were characterized for hazard detection in this Medicaid messaging population. A high-recall first-stage screen attains sensitivity 0.855 with 12.0 missed hazards per 1,000 messages and 729 alerts per 1,000. A disagreement-stratified clinician-review workflow attains sensitivity 0.939 with 5.0 missed hazards per 1,000 messages and 878 alerts per 1,000. Both configurations have residual miss rates below the 16.5 per 1,000 messages implied by a sensitivity floor of 0.80. The cost is alert volumes 2.5 to 3.0 times higher than the best balanced single-architecture operating point. The deployable pattern is the asymmetric-reward Conservative Q-Learning controller (or an equivalently threshold-tuned high-sensitivity classifier) serving as the first-stage screen, with the false-alarm volume routed to a clinician review queue rather than blocking workflow.

Second, no evaluated configuration attained the autonomous (physician-unassisted) triage benchmark (sensitivity 0.80 or higher and specificity 0.80 or higher simultaneously, without clinician review) for hazard detection on the real-world Medicaid messaging population. This held at every operating point on every receiver operating characteristic curve, across all 217 ensemble configurations, and across all 156 two-stage AND-rule cascade configurations. The result is a structural finding, not a threshold-calibration artifact. The best balanced single-architecture operating point was XGBoost with sentence-BERT at sensitivity 0.594 and specificity 0.592. The best balanced ensemble was the hard-voting three-of-13 rule at sensitivity 0.685 and specificity 0.575. Both fall substantially short of the autonomous (physician-unassisted) triage benchmark used here (sensitivity 0.80 or higher and specificity 0.80 or higher), anchored to the IDx-DR autonomous-diagnostic-AI pivotal trial [16]. The interpretation is that autonomous deployment of any evaluated architecture without clinician review is not supported by the present evidence on this population; clinician-reviewed deployment is required. This deployment pattern fits how primary care messaging already works [1, 3]. Medical assistants and care coordinators perform an initial triage and route messages to a physician whose inbox is already near capacity [4]. A high-sensitivity Stage 1 artificial intelligence screen would add to this human step rather than replace it, and is of operational relevance at safety-net practices, where the supply of primary care providers falls short of demand [5].

This finding is specific to the population evaluated. Messages in this Medicaid managed care patient-messaging service were written at substantially lower reading level than

physician-scripted scenarios (mean grade 4.590 versus 5.810) and contained colloquialisms at substantially higher rates (59.1% versus 19.5%); abbreviation and implicit context rates were directionally lower in real-world messages than in physician scenarios. These linguistic differences, combined with the natural hazard-category mix of a Medicaid messaging service (described in Results), distinguish the patients evaluated here from the populations on whom prior clinical artificial intelligence evaluation has been conducted. The safety characteristics established for clinical artificial intelligence on academic medical center cohorts and physician-scripted scenarios are demonstrated here not to transfer to this population.

The harm asymmetry of clinical hazard detection (a missed hazard may cause irreversible patient harm, whereas a false alarm causes recoverable workflow burden) was reflected in the Conservative Q-Learning controller's asymmetric reward design, with a 25-fold penalty on missed hazards relative to false alarms. This architectural commitment to sensitivity over specificity produces a natural first-stage screen in a cascaded workflow but does not, in itself, close the population gap. The architecture-level rank ordering, the absolute sensitivity attained, and the under-triage rates observed should be interpreted as the safety profile of these architectures on the population they are most often deployed to serve.

The 2026 multimodal Articulate Medical Intelligence Explorer demonstration [8] evaluated a frontier large language model on 105 scripted patient-actor scenarios across 32 non-multimodal and 9 multimodal qualitative axes rated by 18 specialists; the present study evaluates 2000 real Medicaid patient short message service messages against physician-adjudicated reference labels using quantitative receiver operating characteristic, F1, and Matthews correlation coefficient outcomes. The two paradigms are categorically distinct on three dimensions. The population dimension contrasts scripted patient actors at an academic medical center research site against real Medicaid patients across three U.S. states. The task dimension contrasts multi-turn multimodal diagnostic conversation against single-turn screening at message arrival triage. The outcome dimension contrasts specialist-rated qualitative axes against physician-adjudicated quantitative binary hazard detection. The two paradigms are complementary rather than competing. The multimodal Articulate Medical Intelligence Explorer characterizes the upper envelope of conversational diagnostic capability under controlled conditions. The present study characterizes the safety floor under operational deployment conditions on a population that the multimodal Articulate Medical Intelligence Explorer cohort does not represent. Liu and colleagues' patient-portal-message triage work [12] is the closest published analog in task structure (artificial intelligence triage of patient-initiated text communication) but evaluated retrieval-augmented frontier large language models on academic medical center portal messages from a single quaternary-care institution, with patient and clinician populations whose linguistic patterns, reading level, and hazard category mix differ from the multi-state Medicaid short message service population evaluated here. The MedHELM benchmark

suite [9] complements both by characterizing frontier-model performance on knowledge-intensive medical tasks; its findings on frontier-model competitiveness on knowledge tasks are not contradicted here, but the present results show that extending those findings to Medicaid messaging triage is non-trivial. Published clinical-message triage evaluations [28, 29] and decision-support comparisons in radiology [30] and oncology [31] have likewise informed the architectural design of the present study; the contribution of this work is to bring the same quantitative discipline to a multi-state Medicaid messaging population. A preprint on hallucination reduction in clinically safe large language models [32] is consistent with the present study's finding that safety-augmented prompting affects the operating-point characteristics of frontier models but does not by itself match the performance of a controller trained on population-specific data.

The principal deployment implication of these findings is that current artificial intelligence architectures evaluated here are not deployable as autonomous single-stage screens for hazard detection in this multi-state Medicaid messaging population; deployment requires human-in-the-loop oversight at clinically meaningful rates. Under United States Food and Drug Administration Clinical Decision Support guidance, clinical decision support that recommends specific patient actions or is integrated with clinical workflow is subject to regulatory oversight; population-relevant performance evidence is a critical input to that oversight. The present findings constitute population-relevant performance evidence that does not support autonomous deployment of any single evaluated architecture, ensemble, or two-stage cascade for unsupervised hazard detection on this population.

The classical clinical-screening cascade (a high-sensitivity screen followed by a high-specificity confirmatory step) is a relevant framing for these results. In mammography, screening sensitivity prioritizes catching cancers and biopsy supplies confirmation. In human immunodeficiency virus testing, ELISA prioritizes sensitivity and Western blot supplies confirmation. In cervical cancer screening, Papanicolaou cytology prioritizes sensitivity and colposcopy with biopsy supplies confirmation. Policy A instantiates this pattern for Medicaid messaging triage. The Stage 1 high-sensitivity artificial intelligence screen is the asymmetric-reward Conservative Q-Learning controller, or equivalently the threshold-tuned logistic regression with TF-IDF features at the sensitivity-floor-0.85 operating point. The Stage 2 high-specificity confirmatory step is clinician review of the screen-positive queue. In the present data, the clinician (rather than a second artificial intelligence) functioned as the higher-specificity confirmatory step. The intrinsic discriminative capacity of every evaluated automated Stage 2 on this population (frontier large language model, ensemble, or learned classifier) is too low to function as the confirmatory test without losing the screen-positive true positives that Stage 1 catches. For completeness, a two-artificial-intelligence cascade pairing the Stage 1 logistic regression screen with a Claude Opus 4.7 Stage 2 was evaluated and is reported in Multimedia Appendix 1. That cascade is dominated by Policy A on the safety-critical residual miss rate:

58.5 missed hazards per 1,000 messages versus 12.0 for Policy A. On this population, the clinician-reviewed confirmatory step outperformed every evaluated two-stage artificial-intelligence cascade on the safety-critical residual miss rate.

Policy A maps to the classical clinical-screening cascade in this setting. A Stage 1 artificial intelligence screen is sized for high sensitivity. All screen-positive messages route to a clinician confirmation queue, sized to the architecture's empirical alert rate (729 alerts per 1,000 messages; Table S8). This pattern is consistent with prior art in clinical computer-aided detection, where artificial intelligence supplies pre-clinician filtering rather than autonomous decision-making. It is also the deployment pattern that the TRIPOD+AI [26] and DECIDE-AI [27] reporting guidelines anticipate for early-clinical-evaluation-stage decision support systems. The present results constitute the population-relevance evidence those frameworks require as a precondition for any deployment claim.

The longer-term deployment pathway, autonomous or semi-autonomous artificial intelligence hazard detection at the autonomous (physician-unassisted) triage benchmark on the Medicaid messaging population, will require closing the population gap documented here. The research agenda for closing the gap is outlined in future work below.

What this study adds

- To the authors' knowledge, this is the first published evaluation of clinical artificial intelligence safety on patient-initiated short message service communication from a multi-state Medicaid managed care population, with sample size, real-world linguistic profile, hazard-category mix, and physician-adjudicated reference labels all distinct from the academic medical center cohorts and physician-scripted scenarios on which prior frontier language model and clinical-decision-support evaluations have been conducted.
- Two operating points meeting the pre-specified sensitivity-floor target were identified on this population: a high-recall first-stage screen (Policy A) and a disagreement-stratified clinician-review workflow (Policy B). Both instantiate the classical clinical-screening cascade of a high-sensitivity screen followed by clinician confirmation. The autonomous (physician-unassisted) triage benchmark of sensitivity 0.80 or higher and specificity 0.80 or higher, based on the IDx-DR autonomous-diagnostic-AI pivotal trial [16], was not attained by any of the approximately 400 evaluated configurations.
- Closing the gap from the best balanced operating point achievable here to the autonomous (physician-unassisted) triage benchmark motivates a structured research agenda anchored on larger population-relevant training corpora, expansion of the retrieval-augmented-generation corpus beyond the 1,280-example development set, domain-adapted small-model fine-tuning, active learning, and

multi-turn evaluation, with prospective validation under DECIDE-AI Stage 2 as the gating downstream step.

Development and deployment costs

The deployed decision support architecture and the supervised baselines incur development costs that the frontier language model configurations do not. Construction of the 1,280-example development corpus required three-physician adjudication of every message (hazard binary, hazard category, and triage action), with disagreement-resolution and senior-arbiter overhead. Training the Conservative Q-Learning controller, the constellation classifier, the ActionHead, and the supervised baselines required iterative hyperparameter selection on the validation split and Platt calibration. Periodic re-calibration is anticipated as the messaging population and clinical context evolve. The frontier language model configurations, in contrast, require no training data and no in-house compute, but introduce ongoing per-inference application-programming-interface cost, vendor dependency (model versioning, deprecation, terms-of-service change), and lower transparency into the model's internal computation. The deployment-cost trade-off across these two cost structures is a function of message volume, expected horizon, regulatory documentation requirements, and institutional in-house engineering capacity, and is outside the scope of the present evaluation.

Strengths and limitations

Strengths. A real-world test set drawn from a multi-state Medicaid managed care population (Virginia, Washington, Ohio) rather than a curated benchmark; physician-adjudicated reference labels on a per-message basis; a single deterministic pipeline run with SHA-256 input checksums producing every reported number end-to-end; per-message predictions retained and available on request for independent re-derivation of every table; pre-registered Hochberg-corrected paired comparisons across the 78-pair architecture matrix; Wilson 95% confidence intervals and 10,000-iteration bootstrap (seed=42) for derived metrics; a configuration matrix spanning deployed clinical decision support, classical supervised baselines, and current-generation frontier large language models.

Limitations. This study has the following limitations, each named explicitly so the reader can weigh each against the reported findings.

Geographic scope. The participating multi-state clinically integrated network spans three U.S. states (Virginia, Washington, Ohio). Generalization to Medicaid populations in other states, to other payer populations, or to non-U.S. populations is not established and is identified as a future work priority.

Retrospective design. This is a retrospective evaluation. Prospective validation under a registered protocol is the appropriate DECIDE-AI Stage 2 next step and is identified as a future work priority.

Single-turn evaluation. Architectures were evaluated in single-turn screening at message arrival mode. The canonical test set carries per-message labels but no thread-level linkage; multi-turn dialogue evaluation will require Institutional Review Board amendment and a structured re-extraction of the messaging corpus with thread linkage and conversation-level ground truth, and is identified as a future work priority.

Message-level holdout with limited patient overlap. The train/test split was at the message level rather than the patient level. 41 of 1,679 unique test-set patients (2.4%) also appeared in the training set (with non-identical messages). This overlap is reported transparently; a patient-disjoint sensitivity analysis on the 1,638 patient-disjoint test-set patients is presented in the Multimedia Appendix.

Calibration and threshold selection. Thresholds and Platt probability calibration for every architecture with calibrated probability output were fit on a held-out validation split drawn from the development set and frozen before the single test pass; no thresholds, calibration constants, or hyperparameters were selected on the test set at any point. Test-set tuning would have produced optimistically biased operating-point estimates and is not present in any reported configuration. The visibly imperfect probability calibration on the real-world test set (Multimedia Appendix Figure S2) is itself a reported finding rather than a limitation to be corrected by recalibration: it reflects distributional shift between the validation set and the real-world Medicaid messaging population, and recalibration on the test set is a form of test-set tuning that is precluded under the present pre-registered protocol.

Frontier large language model coverage. Three frontier large language model providers (Anthropic Claude Opus 4.7, OpenAI GPT-5.5, and Google Gemini 3.1 Pro Preview) were evaluated under safety-augmented system prompts, and each was additionally evaluated with retrieval-augmented generation over the in-domain training corpus. The conclusions therefore concern current-generation frontier language model-class performance on this population, not vendor-specific claims. The safety-augmented prompt was the pre-specified single frontier-LLM configuration for the primary evaluation. A default-prompt comparator was not run on the held-out test set for any of the three providers, so a within-vendor safety vs default prompt ablation is not available in this evaluation. Open-weight frontier-tier models and non-US frontier providers were also not evaluated.

Sample size. The 2000-message held-out test set was powered for primary architecture comparisons (see Sample size justification, Methods). The hazard-positive subgroup (165 hazards) supports the primary McNemar paired comparisons but limits within-category subgroup inference; subgroups with fewer than 30 hazard-positive cases are flagged with

explicit small-sample warnings in equity-stratified and per-category analyses rather than suppressed.

Temporal validation. The held-out test set covers messages submitted during January 2023 through November 2025. The train/validation/test partition was performed at the message level rather than chronologically; messages from the training, validation, and test partitions are interleaved across the January 2023 through November 2025 interval. Random partitioning over a multi-year interval can in principle mask concept drift that would manifest under prospective deployment. Patient communication patterns and underlying disease prevalence may evolve, and temporal external validation on a forward-in-time messaging cohort is identified as a future work priority.

In summary, this study is a multi-state retrospective single-turn evaluation; the principal findings should be interpreted as population-relevance evidence for a use with clinician review of flagged messages cascade on Medicaid SMS triage, with prospective validation, multi-turn evaluation, geographic extension beyond the three participating states, and temporal external validation identified as the corresponding future work priorities.

Future work: research agenda for closing the population gap

Closing the population gap documented in this study requires a structured research program. The gap is the distance from the best-balanced achievable operating point (sensitivity 0.685 and specificity 0.575 via the three-of-13 hard-voting ensemble) to the autonomous (physician-unassisted) triage benchmark (sensitivity 0.80 or higher and specificity 0.80 or higher). Five complementary directions are identified.

The first direction is expansion of the training corpus. The development set in this study comprised approximately 1,280 labeled examples. The hazard-positive subgroup within the development set is sparse, particularly for less-common categories such as obstetric and metabolic emergencies, and is a plausible source of the observed performance ceiling. Systematic labeling of additional Medicaid messaging samples, with hazard-category stratification, is the highest-priority research direction.

The second direction is an expansion of the retrieval corpus that supports retrieval-augmented generation. The present evaluation included retrieval-augmented generation over the 1,280-example labeled development set, which closed the gap on the physician-scripted comparison scenarios but not on the real-world Medicaid messaging test set. One interpretation is that the available retrieval corpus is too small and too narrowly distributed over the natural Medicaid messaging case mix to return informative neighbors for the colloquialism-rich, abbreviation-laden, implicit-context messages observed in deployment. Scaling the retrieval corpus to encompass a larger labeled set of in-domain

population-relevant exemplars, with class stratification across the full 20 hazard categories, is the priority next step for retrieval-augmented generation in this population.

The third direction is domain-adapted small-model fine-tuning. Quantized low-rank adaptation of openly-available small large language models (such as Llama-3.1-8B or Mistral-7B) on the labeled Medicaid messaging corpus may improve discrimination over zero-shot frontier-model performance while preserving the operational advantages of locally-deployable inference.

The fourth direction is active learning on a held-out unlabeled corpus. Targeted labeling of the messages on which existing architectures disagree most strongly, drawn from the substantial unlabeled portion of the messaging stream, is an efficient route to closing the data gap without the cost of labeling messages on which architectures already agree.

The fifth direction is multi-turn evaluation. The present test set carries per-message labels rather than thread-level labels; multi-turn dialogue evaluation (in which the artificial intelligence can ask clarifying questions before recommending action) will require a structured re-extraction of the messaging corpus with thread linkage and conversation-level ground truth. Recent multi-turn conversational diagnostic artificial intelligence demonstrations on standardized patient actors [8] suggest that multi-turn capability may shift absolute performance for all architectures roughly in parallel; whether it specifically closes the population gap for low-literacy multilingual messaging is an open empirical question.

Two cross-cutting priorities complete the agenda. Temporal external validation will be required as patient communication patterns and the underlying disease prevalence evolve. External validation outside the three participating states (Virginia, Washington, Ohio) will be required before generalization to populations outside the participating multi-state clinically integrated network's enrollees. Different payer populations, different age strata, additional geographies, and different primary languages will not necessarily reproduce the architecture-level rank ordering documented here. A prospective evaluation of clinical impact (whether deployment of the recommended cascade workflow in messaging triage reduces missed hazards relative to standard clinician-only review) is the logical final step once the foregoing have been completed.

Conclusions

Two operating points meeting the pre-specified sensitivity-floor target were identified for hazard detection in this multi-state Medicaid patient-messaging population. Both have empirical alert volumes and residual miss rates that improve on the maximum miss rate

admissible at a sensitivity floor of 0.80. A high-recall first-stage screen attains sensitivity 0.855 with 12.0 missed hazards per 1,000 messages and 729 alerts per 1,000. A disagreement-stratified clinician-review workflow attains sensitivity 0.939 with 5.0 missed hazards per 1,000 messages and 878 alerts per 1,000. The deployable pattern is the classical clinical-screening cascade: a high-sensitivity artificial intelligence Stage 1 screen (the asymmetric-reward Conservative Q-Learning controller or an equivalently tuned high-sensitivity classifier) followed by a high-specificity clinician Stage 2 confirmatory review, with the false-alarm volume sized to the clinician confirmation queue. No evaluated configuration attained the autonomous triage benchmark (sensitivity 0.80 or higher and specificity 0.80 or higher simultaneously, based on the IDx-DR autonomous-diagnostic-AI pivotal trial [16]). This held at every operating point on every receiver operating characteristic curve, across all 217 ensemble configurations, and across all 156 two-stage cascade configurations. The best balanced operating point was sensitivity 0.685 and specificity 0.575. Autonomous deployment without clinician review is therefore not supported by the present evidence on this population, which is characterized by lower reading-level text, higher colloquialism rates, and a natural hazard mix substantially different from the populations on which prior clinical artificial intelligence safety evaluation has been conducted. Closing the gap to autonomous symmetric-benchmark performance motivates the research agenda outlined above. The agenda is anchored on larger population-relevant training corpora, expansion of the retrieval-augmented-generation corpus beyond the 1,280-example development set, domain-adapted small-model fine-tuning, active learning, and multi-turn evaluation, with prospective validation as the gating downstream step.

Declarations

Ethics approval and consent to participate

This study was approved by the U.S. Western Institutional Review Board (WIRB-Copernicus Group, WCG IRB; protocol number 20260179) with a waiver of informed consent. The waiver was granted because the study used retrospective de-identified data and did not affect medical decision making or care for participating patients. All experiments were performed in accordance with relevant guidelines and regulations including the Declaration of Helsinki and the Health Information Portability and Accountability Act (HIPAA).Methodological details of HIPAA Safe Harbor de-identification and the adjudicator role are described in Methods §De-identification and adjudicator role.

Consent for publication

Not applicable. No individual person’s data are presented in this manuscript.

Availability of data and materials

The complete analytic pipeline, including model training scripts, inference scripts, table and figure renderers, and audit utilities, is publicly available at https://github.com/sanjaybasu/rl-llm-safety. Per-architecture per-message predictions (architecture × message identifier × hazard probability × binarized prediction × ground-truth label × hazard category) are released as a supplementary file with the present manuscript. Raw patient message text is not publicly released because it remains under HIPAA data-use agreement; bona-fide researchers may request access from the corresponding author. The released supplementary file is sufficient for independent verification of receiver operating characteristic monotonicity, McNemar discordant-pair counts, paired bootstrap confidence intervals, and table-to-table arithmetic without access to the underlying message text.

Competing interests

All authors are employees at Waymark, a public benefit organization that provides free social and medical services to patients receiving Medicaid. The authors test artificial intelligence tools to improve patient care but do not sell artificial intelligence platforms. All employees have access to artificial intelligence chatbots such as ChatGPT, Claude, and Gemini in the normal course of their work, but Waymark has no sales relationship with the entities owning these platforms.

Funding

None.

Authors' contributions

SB conceived and designed the study, supervised the analytical pipeline, conducted the data analysis, and drafted the manuscript. SYP and PS contributed to the analytical pipeline design, the statistical analysis, and the methodological framing of the threshold-optimization and ensemble analyses. BA contributed to the data engineering for the messaging corpus extraction, the labeling-pipeline tooling, and the deterministic-rendering infrastructure. JS contributed clinical-operations subject-matter expertise on Medicaid messaging triage workflows and reviewed the operational implications of the architecture-comparison findings. JM led clinical adjudication, contributed to hazard category definitions, and reviewed the clinical safety implications of the architectural findings. RB contributed to the deployment framing, the implications for clinical decision support oversight, and the research-agenda direction. All authors reviewed and approved the final manuscript.

Acknowledgements

None.

References

1. Nath B, Williams B, Jeffery MM, O'Connell R, Goldstein R, Sinsky CA, Melnick ER. Trends in electronic health record inbox messaging during the COVID-19 pandemic in an ambulatory practice network in New England. JAMA Netw Open. 2021;4(10):e2131490. doi:10.1001/jamanetworkopen.2021.31490. PMID: 34636917.

2. Grande D, Zuo JX, Venkat R, Chen X, Ward KR, Seymour JW, Mitra N. Differences in primary care appointment availability and wait times by neighborhood characteristics: a Mystery Shopper Study. J Gen Intern Med. 2018;33(9):1441-1443. doi:10.1007/s11606-018-4407-9. PMID: 29869139.

3. Lee JN, Kurash L, Yang M, Teel J. Utilizing medical assistants to manage patient portal messages. Ann Fam Med. 2024;22(3):261. doi:10.1370/afm.3105. PMID: 38806276.

4. Tai-Seale M, Dillon EC, Yang Y, Nordgren R, Steinberg RL, Nauenberg T, Lee TC, Meehan A, Li J, Chan AS, Frosch DL. Physicians' well-being linked to in-basket messages generated by algorithms in electronic health records. Health Aff (Millwood). 2019;38(7):1073-1078. doi:10.1377/hlthaff.2018.05509. PMID: 31260371.

5. Xue Y, Greener E, Kannan V, Smith JA, Brewer C, Spetz J. Federally qualified health centers reduce the primary care provider gap in health professional shortage counties. Nurs Outlook. 2018;66(3):263-272. doi:10.1016/j.outlook.2018.02.003. PMID: 29685322.

6. Tu T, Schaekermann M, Palepu A, Saab K, Freyberg J, Tanno R, et al. Towards conversational diagnostic artificial intelligence. Nature. 2025;642(8067):442-450. doi:10.1038/s41586-025-08866-7. PMID: 40205050.

7. McDuff D, Schaekermann M, Tu T, Palepu A, Wang A, Garrison J, et al. Towards accurate differential diagnosis with large language models. Nature. 2025;642(8067):451-457. doi:10.1038/s41586-025-08869-4. PMID: 40205049.

8. Saab K, Park C, Strother T, Freyberg J, Barrett DGT, Cheng Y, Weng WH, Stutz D, Tomasev N, Palepu A, et al. Advancing conversational diagnostic AI with multimodal

reasoning. Nat Med. 2026 May;32(5):1726-1736. doi:10.1038/s41591-026-04371-0. PMID: 42135531.

9. Bedi S, Cui H, Fuentes M, Unell A, Wornow M, Banda JM, Kotecha N, Keyes T, Mai Y, Oez M, et al. Holistic evaluation of large language models for medical tasks with MedHELM. Nat Med. 2026 Mar;32(3):943-951. doi:10.1038/s41591-025-04151-2. PMID: 41559415.

10. Brodeur PG, Buckley TA, Kanjee Z, Goh E, Ling EB, Jain P, et al. Performance of a large language model on the reasoning tasks of a physician. Science. 2026;392(6797):524-527. doi:10.1126/science.adz4433. PMID: 42060751.

11. Ramaswamy A, Tyagi A, Hugo H, Jiang J, Jayaraman P, et al. ChatGPT Health performance in a structured test of triage recommendations. Nat Med. 2026. doi:10.1038/s41591-026-04297-7. PMID: 41731097.

12. Liu S, Wright AP, McCoy AB, Huang SS, Steitz B, Wright A. Detecting emergencies in patient portal messages using large language models and knowledge graph-based retrieval-augmented generation. J Am Med Inform Assoc. 2025;32(6):1032-1039. doi:10.1093/jamia/ocaf059. PMID: 40220286.

13. Allen H, Wright BJ, Harding K, Broffman L. The role of stigma in access to health care for the poor. Milbank Q. 2014;92(2):289-318. doi:10.1111/1468-0009.12059. PMID: 24890249.

14. Basu S, Muralidharan B, Sheth P, Wanek D, Morgan J, Patel S. Reinforcement Learning to Prevent Acute Care Events Among Medicaid Populations: Mixed Methods Study. JMIR AI. 2025;4:e74264. doi:10.2196/74264. PMID: 41062083.

15. Newcombe RG. Interval estimation for the difference between independent proportions: comparison of eleven methods. Stat Med. 1998;17(8):873-890. doi:10.1002/(SICI)1097-0258(19980430)17:8<873::AID-SIM779>3.0.CO;2-I.

16. Abràmoff MD, Lavin PT, Birch M, Shah N, Folk JC. Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices. NPJ Digit Med. 2018;1:39. doi:10.1038/s41746-018-0040-6. PMID: 31304320.

17. Reimers N, Gurevych I. Sentence-BERT: sentence embeddings using siamese BERT-networks. Proc 2019 Conf Empir Methods Nat Lang Process. 2019:3982-3992.

18. Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. Proc 2019 Conf North Am Chapter Assoc Comput Linguist (NAACL-HLT). 2019:4171-4186. doi:10.18653/v1/N19-1423.

19. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. Proc 22nd ACM SIGKDD Int Conf Knowl Discov Data Min. 2016:785-794. doi:10.1145/2939672.2939785.

20. Kumar A, Zhou A, Tucker G, Levine S. Conservative Q-learning for offline reinforcement learning. Adv Neural Inf Process Syst. 2020;33:1179-1191.

21. Shah NH, Halamka JD, Saria S, Pencina M, Tazbaz T, Tripathi M, Callahan A, Hildahl H, Anderson B. A nationwide network of health AI assurance laboratories. JAMA. 2024;331(3):245-249. doi:10.1001/jama.2023.26930. PMID: 38117493.

22. Coalition for Health AI. Blueprint for Trustworthy AI Implementation Guidance and Assurance for Healthcare. April 2023. Accessed 2026-05-20. https://chai.org/chai-unveils-blueprint-for-trustworthy-ai-in-healthcare/

23. Sendak MP, Kim JY, Hasan A, Ratliff W, Lifson MA, Patel M, Raji ID, Sehgal A, Shaw K, Tobey D, Valladares A, Vidal DE, Balu S. Empowering US healthcare delivery organizations: Cultivating a community of practice to harness AI and advance health equity. PLOS Digit Health. 2024;3(6):e0000513. doi:10.1371/journal.pdig.0000513. PMID: 38843115.

24. McNemar Q. Note on the sampling error of the difference between correlated proportions or percentages. Psychometrika. 1947;12(2):153-157. doi:10.1007/BF02295996.

25. Hochberg Y. A sharper Bonferroni procedure for multiple tests of significance. Biometrika. 1988;75(4):800-802. doi:10.1093/biomet/75.4.800.

26. Collins GS, Moons KGM, Dhiman P, Riley RD, Beam AL, Van Calster B, Ghassemi M, Liu X, Reitsma JB, van Smeden M, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ. 2024;385:e078378. doi:10.1136/bmj-2023-078378. PMID: 38626948.

27. Vasey B, Nagendran M, Campbell B, Clifton DA, Collins GS, Denaxas S, Denniston AK, Faes L, Geerts B, Ibrahim M, et al. Reporting guideline for the early stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. BMJ. 2022;377:e070904. doi:10.1136/bmj-2022-070904. PMID: 35584845.

28. Šimunović I, Rezić K, Franić N, Boduljak G, Batinić M, Jukić I, Jelovina I, Biočić J, Pogorelić Z, Markić J. Can general purpose large language models assist pediatricians in predicting infants with serious bacterial infection? BMC Med Inform Decis Mak. 2025;25(1):423. doi:10.1186/s12911-025-03258-3. PMID: 41239388.

29. Esmaeilzadeh P. Ethical implications of using general-purpose LLMs in clinical settings: a comparative analysis of prompt engineering strategies and their impact on patient safety. BMC Med Inform Decis Mak. 2025;25(1):342. doi:10.1186/s12911-025-03182-6. PMID: 41024005.

30. Saban M, Alon Y, Luxenburg O, Singer C, Hierath M, Karoussou Schreiner A, Brkljačić B, Sosna J. Comparison of CT referral justification using clinical decision support and large language models in a large European cohort. Eur Radiol. 2025 Oct;35(10):6150-6159. doi:10.1007/s00330-025-11608-y. PMID: 40287868. PMCID: PMC12417242.

31. Nazlı MA, Esmerer E, Keles A. Benchmarking large language models in breast cancer care: agreement with radiology-led multidisciplinary tumor board decisions. BMC Med Inform Decis Mak. 2026; in press. doi:10.1186/s12911-026-03556-4. PMID: 42121209.

32. Wu D, Haredasht FN, Maharaj SK, et al. First, do NOHARM: towards clinically safe large language models. arXiv 2512.01241; deposited 2025-12-17. PubMed preprint PMID 41532042. [Preprint; not yet peer-reviewed at time of submission.] ## Tables

Table 1. Population characteristics of the held-out real-world test set and the physician-scripted comparison scenarios.

Characteristic	Real-world Medicaid test set	Physician-scripted comparison
Sample size, N	2000	41
Hazards adjudicated, n (%)	165 (8.25%)	27 (65.9%)
Benigns adjudicated, n (%)	1835 (91.75%)	14 (34.1%)
Reading level (Flesch-Kincaid grade), mean	4.6	5.8
Colloquialisms, % messages	59.1	19.5
Abbreviations, % messages	24.5	36.6
Implicit contextual references, % messages	7.8	29.3
Word count, mean	28.4	38.6

Caption: Sample sizes, hazard prevalence, age distribution, sex, self-reported race/ethnicity, and linguistic feature distributions for the 2000-message real-world test set and the 41-case physician-scripted comparison set. Linguistic features computed via the automated NLP pipeline described in Methods. Reading level by Flesch-Kincaid grade.

Table 2. Per-architecture hazard detection metrics on the real-world test set.

Architecture	TP / FN / TN / FP	Sensitivity (95% CI)	Specificity (95% CI)	PPV (95% CI)	NPV (95% CI)	F1 (95% CI)	MCC (95% CI)	AUROC (95% CI)	FN per 1,000 (95% CI)
ActionHead (action recommender)	70 / 95 / 1472 / 363	0.424 (0.351–0.501)	0.802 (0.783–0.820)	0.162 (0.130–0.199)	0.939 (0.926–0.950)	0.234 (0.189–0.279)	0.151 (0.099–0.204)	N/A*	47.5 (41.2–53.5)
Claude Opus 4.7 (safety-augmented)	49 / 116 / 1598 / 237	0.297 (0.233–0.371)	0.871 (0.855–0.885)	0.171 (0.132–0.219)	0.932 (0.919–0.943)	0.217 (0.167–0.268)	0.132 (0.077–0.189)	N/A*	58.0 (51.9–63.3)
Constellation architecture	11 / 154 / 1784 / 51	0.067 (0.038–0.115)	0.972 (0.964–0.979)	0.177 (0.102–0.290)	0.921 (0.908–0.932)	0.097 (0.046–0.152)	0.062 (0.004–0.125)	0.593 (0.546–0.640)	77.0 (73.0–79.4)
CQL controller (reward-optimized)	154 / 11 / 161 / 1674	0.933 (0.885–0.962)	0.088 (0.076–0.102)	0.084 (0.072–0.098)	0.936 (0.889–0.964)	0.155 (0.133–0.176)	0.021 (-0.021–0.057)	0.561 (0.514–0.608)	5.5 (3.1–9.5)
CQL controller (sensitivity-optimized)	4 / 161 / 1829 / 6	0.024 (0.009–0.061)	0.997 (0.993–0.999)	0.400 (0.168–0.687)	0.919 (0.906–0.930)	0.046 (0.011–0.092)	0.082 (0.004–0.162)	0.561 (0.514–0.608)	80.5 (77.5–81.7)
Gemini 3.1 Pro (safety-augmented)	3 / 162 / 1825 / 10	0.018 (0.006–0.052)	0.995 (0.990–0.997)	0.231 (0.082–0.503)	0.918 (0.906–0.930)	0.034 (0.000–0.075)	0.044 (-0.020–0.116)	N/A*	81.0 (78.2–82.0)

Architecture	TP / FN / TN / FP	Sensitivity (95% CI)	Specificity (95% CI)	PPV (95% CI)	NPV (95% CI)	F1 (95% CI)	MCC (95% CI)	AUROC (95% CI)	FN per 1,000 (95% CI)
Rule-based guardrails	10 / 155 / 1803 / 32	0.061 (0.033–0.108)	0.983 (0.975–0.988)	0.238 (0.135–0.385)	0.921 (0.908–0.932)	0.097 (0.043–0.153)	0.083 (0.017–0.153)	0.503 (0.479–0.528)	77.5 (73.6–79.8)
Logistic regression + TF-IDF	160 / 5 / 170 / 1665	0.970 (0.931–0.987)	0.093 (0.080–0.107)	0.088 (0.076–0.102)	0.971 (0.935–0.988)	0.161 (0.139–0.183)	0.061 (0.030–0.086)	0.589 (0.546–0.633)	2.5 (1.1–5.7)
XGBoost + sentence-BERT	70 / 95 / 1328 / 507	0.424 (0.351–0.501)	0.724 (0.703–0.744)	0.121 (0.097–0.150)	0.933 (0.919–0.945)	0.189 (0.151–0.227)	0.090 (0.042–0.138)	0.606 (0.562–0.650)	47.5 (41.2–53.5)
Claude Opus 4.7 (RAG)	36 / 129 / 1684 / 151	0.218 (0.162–0.287)	0.918 (0.904–0.929)	0.193 (0.142–0.255)	0.929 (0.916–0.940)	0.205 (0.148–0.261)	0.128 (0.069–0.189)	N/A*	64.5 (58.8–69.1)
Gemini 3.1 Pro (RAG)	4 / 161 / 1798 / 37	0.024 (0.009–0.061)	0.980 (0.972–0.985)	0.098 (0.039–0.225)	0.918 (0.905–0.929)	0.039 (0.009–0.078)	0.008 (-0.034–0.059)	N/A*	80.5 (77.5–81.7)
GPT-5.5 (safety-augmented)	53 / 112 / 1583 / 252	0.321 (0.255–0.396)	0.863 (0.846–0.878)	0.174 (0.135–0.220)	0.934 (0.921–0.945)	0.226 (0.175–0.277)	0.141 (0.085–0.197)	N/A*	56.0 (49.8–61.5)
GPT-5.5 (RAG)	43 / 122 / 1630	0.261 (0.200–	0.888 (0.873–	0.173 (0.131–	0.930 (0.917–	0.208 (0.156–	0.124 (0.068–	N/A*	61.0 (55.1–66.0)

Architecture	TP / FN / TN / FP	Sensitivity (95% CI)	Specificity (95% CI)	PPV (95% CI)	NPV (95% CI)	F1 (95% CI)	MCC (95% CI)	AUROC (95% CI)	FN per 1,000 (95% CI)
	/ 205	0.332)	0.902)	0.225)	0.941)	0.262)	0.183)		

N/A: AUROC is undefined for this architecture because the configured inference API returns a discrete hazard flag in a structured JSON output rather than a continuous probability scalar, leaving no score to sweep across thresholds. This affects the three frontier large language model configurations (Claude Opus 4.7, GPT-5.5, Gemini 3.1 Pro Preview) under both the safety-augmented and retrieval-augmented prompts, and the ActionHead action recommender, which emits a discrete action class rather than a calibrated hazard probability.*

Caption: Sensitivity, specificity, positive predictive value, negative predictive value, F1, Matthews correlation coefficient, area under the receiver operating characteristic curve (where calibrated probability output is available), and false negatives per 1,000 messages at each architecture's calibrated operating threshold on the 2000-message held-out real-world test set. 95% confidence intervals: Wilson score method [12] for sensitivity, specificity, positive predictive value, and negative predictive value; parametric bootstrap (10,000 iterations, seed 42) for F1 and Matthews correlation coefficient; Hanley-McNeil for area under the receiver operating characteristic curve.

Table 3. Operating-point analysis: sensitivity at matched specificity targets for architectures with calibrated probability output.

Architecture	Sens at Spec ≥ 0.70	Sens at Spec ≥ 0.73	Sens at Spec ≥ 0.80	Sens at Spec ≥ 0.90	Sens at Spec ≥ 0.95
Constellation architecture	0.430 (0.357–0.507)	0.406 (0.334–0.482)	0.327 (0.260–0.402)	0.218 (0.162–0.287)	0.127 (0.085–0.187)
CQL controller (reward-optimized)	0.412 (0.340–0.488)	0.370 (0.300–0.446)	0.291 (0.227–0.364)	0.139 (0.095–0.200)	0.109 (0.070–0.166)
CQL controller (sensitivity-optimized)	0.412 (0.340–0.488)	0.370 (0.300–0.446)	0.291 (0.227–0.364)	0.139 (0.095–0.200)	0.109 (0.070–0.166)
Rule-based guardrails	0.061†	0.061†	0.061†	0.061†	0.061†

Architecture	Sens at Spec ≥ 0.70	Sens at Spec ≥ 0.73	Sens at Spec ≥ 0.80	Sens at Spec ≥ 0.90	Sens at Spec ≥ 0.95
	(0.033–0.108)	(0.033–0.108)	(0.033–0.108)	(0.033–0.108)	(0.033–0.108)
Logistic regression + TF-IDF	0.424 (0.351–0.501)	0.394 (0.323–0.470)	0.303 (0.238–0.377)	0.139 (0.095–0.200)	0.061 (0.033–0.108)
XGBoost + sentence-BERT	0.461 (0.386–0.537)	0.418 (0.346–0.494)	0.327 (0.260–0.402)	0.133 (0.090–0.194)	0.091 (0.056–0.145)

Caption: For each architecture with calibrated probability output, the threshold that achieves the target specificity is identified on the architecture's receiver operating characteristic curve, and the sensitivity attained at that threshold is reported with the achieved specificity displayed alongside the target. The receiver operating characteristic curves derive from the same per-message prediction files as Table 2; monotonicity (sensitivity non-decreasing as specificity decreases) is enforced as a pipeline assertion.

Table 4. Threshold-optimized single-architecture operating points and autonomous-benchmark reachability.

Architecture	F1-max (95% CI)	Sens at F1-max (95% CI)	Spec at F1-max (95% CI)	MCC-max (95% CI)	Sens at MCC-max (95% CI)	Spec at MCC-max (95% CI)	Reached autonomous benchmark?	Sens/Spec at closest point (95% CI)
Constellation architecture	0.197 (0.161–0.234)	0.376 (0.305–0.452)	0.781 (0.761–0.799)	0.109 (0.057–0.160)	0.218 (0.162–0.287)	0.904 (0.890–0.917)	No	0.564 / 0.563 (sens CI 0.487–0.637; spec CI 0.540–0.585)
CQL controller (reward-optimized)	0.175 (0.146–0.20	0.412 (0.340–0.488)	0.704 (0.682–0.724)	0.096 (0.026–0.16	0.067 (0.038–0.115)	0.983 (0.976–0.988)	No	0.539 / 0.540 (sens CI 0.463–0.614; spec CI 0.517–0.563)

Architecture	F1-max (95% CI)	Sens at F1-max (95% CI)	Spec at F1-max (95% CI)	MCC-max (95% CI)	Sens at MCC-max (95% CI)	Spec at MCC-max (95% CI)	Reached autonomous benchmark?	Sens/Spec at closest point (95% CI)
	6)			6)				
CQL controller (sensitivity-optimized)	0.175 (0.146–0.206)	0.412 (0.340–0.488)	0.704 (0.682–0.724)	0.096 (0.026–0.166)	0.067 (0.038–0.115)	0.983 (0.976–0.988)	No	0.539 / 0.540 (sens CI 0.463–0.614; spec CI 0.517–0.563)
Rule-based guardrails	0.153 (0.152–0.153)	1.000 (0.977–1.000)	0.002 (0.001–0.006)	0.083 (0.018–0.152)	0.061 (0.033–0.108)	0.983 (0.975–0.988)	No	0.061 / 0.983 (sens CI 0.033–0.108; spec CI 0.975–0.988)
Logistic regression + TF-IDF	0.183 (0.152–0.217)	0.412 (0.340–0.488)	0.723 (0.702–0.743)	0.095 (0.061–0.128)	0.873 (0.813–0.915)	0.281 (0.261–0.302)	No	0.551 / 0.564 (sens CI 0.475–0.625; spec CI 0.541–0.586)
XGBoost + sentence-BERT	0.201 (0.166–0.239)	0.412 (0.340–0.488)	0.759 (0.738–0.778)	0.108 (0.058–0.161)	0.412 (0.340–0.488)	0.759 (0.738–0.778)	No	0.594 / 0.592 (sens CI 0.518–0.666; spec CI 0.569–0.614)

Caption: For each architecture with calibrated probability output, the F1-maximizing and Matthews-correlation-coefficient-maximizing operating points on the real-world test set receiver operating characteristic curve are reported, along with the closest-to-symmetric-benchmark operating point and a binary indicator of whether the architecture's receiver operating characteristic curve passes through the sensitivity ≥ 0.80 AND specificity ≥ 0.80 region. Threshold-optimized analysis is post hoc: the thresholds reported here are selected on the real-world test set operating curve and characterize the upper bound of threshold-

tuning achievable performance; clinical deployment requires threshold selection on a held-out validation set or prospective calibration. The headline result is that no architecture reaches the autonomous (physician-unassisted) triage benchmark at any operating point on its receiver operating characteristic curve.

Table 5. Cascade architectures: Pareto frontier of two-stage AND-rule configurations on the real-world test set.

Cascade (Stage 1 × Stage 2)	Sensitivity (95% CI)	Specificity (95% CI)	PPV (95% CI)	F1 (95% CI)	MCC (95% CI)	FN per 1,000 (95% CI)
CQL reward-opt × LogReg+TF-IDF	0.915 (0.863–0.949)	0.167 (0.151–0.185)	0.090 (0.077–0.105)	0.164 (0.156–0.171)	0.062 (0.026–0.094)	7.0 (4.2–11.3)
ActionHead × LogReg+TF-IDF	0.418 (0.346–0.494)	0.809 (0.791–0.827)	0.165 (0.132–0.203)	0.236 (0.194–0.278)	0.154 (0.099–0.209)	48.0 (41.7–54.0)
CQL reward-opt × GPT-5.5	0.315 (0.249–0.390)	0.867 (0.851–0.882)	0.176 (0.137–0.223)	0.226 (0.179–0.270)	0.141 (0.087–0.194)	56.5 (50.4–61.9)
Claude Opus 4.7 × CQL reward-opt	0.297 (0.233–0.371)	0.874 (0.858–0.888)	0.174 (0.134–0.223)	0.220 (0.174–0.266)	0.135 (0.081–0.188)	58.0 (51.9–63.3)
Claude Opus 4.7 × LogReg+TF-IDF	0.291 (0.227–0.364)	0.875 (0.859–0.889)	0.173 (0.133–0.221)	0.217 (0.172–0.263)	0.132 (0.078–0.185)	58.5 (52.4–63.8)
CQL reward-opt × GPT-5.5 + RAG	0.261 (0.200–0.333)	0.894 (0.879–0.907)	0.181 (0.137–0.235)	0.213 (0.165–0.260)	0.131 (0.077–0.184)	61.0 (55.1–66.0)
Claude Opus 4.7 × GPT-5.5	0.236 (0.178–0.307)	0.909 (0.895–0.922)	0.190 (0.142–0.249)	0.211 (0.160–0.260)	0.132 (0.078–0.188)	63.0 (57.2–67.8)
GPT-5.5 + RAG × GPT-5.5	0.224 (0.167–0.294)	0.924 (0.911–0.935)	0.210 (0.157–0.276)	0.217 (0.165–0.271)	0.144 (0.088–0.202)	64.0 (58.3–68.7)
ActionHead × GPT-5.5	0.212 (0.157–0.281)	0.944 (0.933–0.954)	0.256 (0.190–0.334)	0.232 (0.174–0.291)	0.171 (0.108–0.233)	65.0 (59.3–69.6)

Cascade (Stage 1 × Stage 2)	Sensitivity (95% CI)	Specificity (95% CI)	PPV (95% CI)	F1 (95% CI)	MCC (95% CI)	FN per 1,000 (95% CI)
Claude 4.7 + RAG × GPT-5.5	0.176 (0.125–0.241)	0.948 (0.936–0.957)	0.232 (0.167–0.313)	0.200 (0.140–0.263)	0.140 (0.076–0.208)	68.0 (62.6–72.2)
ActionHead × GPT-5.5 + RAG	0.164 (0.115–0.228)	0.960 (0.950–0.968)	0.270 (0.193–0.364)	0.204 (0.141–0.272)	0.156 (0.089–0.231)	69.0 (63.7–73.0)
ActionHead × Claude 4.7 + RAG	0.139 (0.095–0.201)	0.964 (0.954–0.971)	0.256 (0.177–0.354)	0.180 (0.120–0.247)	0.137 (0.072–0.212)	71.0 (66.0–74.7)
Constellation × XGBoost+SBERT	0.067 (0.038–0.115)	0.982 (0.974–0.987)	0.244 (0.142–0.387)	0.105 (0.049–0.164)	0.089 (0.020–0.161)	77.0 (73.0–79.4)
CQL reward-opt × Guardrails	0.061 (0.033–0.108)	0.983 (0.976–0.988)	0.244 (0.138–0.393)	0.097 (0.041–0.154)	0.085 (0.017–0.154)	77.5 (73.6–79.8)
Constellation × GPT-5.5	0.049 (0.025–0.093)	0.986 (0.979–0.990)	0.235 (0.124–0.400)	0.080 (0.031–0.137)	0.073 (0.007–0.142)	78.5 (74.9–80.5)
Guardrails × XGBoost+SBERT	0.042 (0.021–0.085)	0.992 (0.987–0.995)	0.318 (0.164–0.527)	0.075 (0.023–0.131)	0.090 (0.017–0.167)	79.0 (75.5–80.8)
CQL sens-opt × XGBoost+SBERT	0.024 (0.009–0.061)	0.998 (0.994–0.999)	0.500 (0.215–0.785)	0.046 (0.012–0.091)	0.096 (0.010–0.179)	80.5 (77.5–81.7)
Gemini 3.1 Pro × XGBoost+SBERT	0.018 (0.006–0.052)	0.999 (0.996–1.000)	0.600 (0.231–0.882)	0.035 (0.000–0.080)	0.094 (-0.009–0.170)	81.0 (78.2–82.0)
CQL sens-opt × GPT-5.5 + RAG	0.012 (0.003–0.043)	1.000 (0.997–1.000)	0.667 (0.208–0.939)	0.024 (0.000–0.059)	0.082 (-0.009–0.167)	81.5 (78.9–82.2)
Constellation × Guardrails	0.012 (0.003–0.043)	1.000 (0.997–1.000)	0.667 (0.208–0.939)	0.024 (0.000–0.059)	0.082 (-0.009–0.167)	81.5 (78.9–82.2)

Cascade (Stage 1 × Stage 2)	Sensitivity (95% CI)	Specificity (95% CI)	PPV (95% CI)	F1 (95% CI)	MCC (95% CI)	FN per 1,000 (95% CI)
Constellation × Gemini 3.1 Pro	0.012 (0.003–0.043)	1.000 (0.997–1.000)	0.667 (0.208–0.939)	0.024 (0.000–0.059)	0.082 (-0.009–0.167)	81.5 (78.9–82.2)
CQL sens-opt × Guardrails	0.006 (0.001–0.034)	1.000 (0.998–1.000)	1.000 (0.207–1.000)	0.012 (0.000–0.036)	0.075 (0.075–0.149)	82.0 (79.7–82.4)
CQL sens-opt × Gemini 3.1 Pro	0.006 (0.001–0.034)	1.000 (0.998–1.000)	1.000 (0.207–1.000)	0.012 (0.000–0.036)	0.075 (0.075–0.149)	82.0 (79.7–82.4)

Caption: Each row reports a Stage 1 architecture (high-recall screen) paired with a Stage 2 architecture (high-precision filter), with the final positive decision requiring concordance between both stages. Sensitivity, specificity, positive predictive value, F1, and Matthews correlation coefficient are computed on the AND-rule cascade output. Pareto-frontier pairs (those not dominated on both sensitivity and specificity by any other pair) are listed; no cascade configuration attained sensitivity ≥ 0.80 AND specificity ≥ 0.80 simultaneously. The cascade pairing the asymmetric-reward Conservative Q-Learning controller with the safety-augmented Claude Opus 4.7 frontier large language model is the highest-specificity pair in the cascade matrix that retains non-trivial sensitivity. The full 72-pair cascade matrix is provided in Multimedia Appendix 1, Table S3.

Table 6. Deployment-grade configurations: sensitivity-floor analysis and disagreement-stratified clinician-review workflow.

Panel A — Sensitivity-floor target: max specificity achievable subject to sensitivity threshold

Sensitivity floor	N configurations meeting floor	Max specificity at floor (95% CI)	Winning configuration	Sens (winning) (95% CI)
≥ 0.80	2911	0.359 (0.337–0.381)	single_architecture_threshold_sweep / LogReg+TF-IDF@thresh	0.800 (0.732–0.854)
≥ 0.85	2242	0.282 (0.262–	single_architecture_threshold_s	0.855 (0.793–

Sensitivity floor	N configurations meeting floor	Max specificity at floor (95% CI)	Winning configuration	Sens (winning) (95% CI)
		0.303)	weep / LogReg+TF-IDF@thresh	0.900)
≥ 0.90	1563	0.191 (0.173–0.209)	single_architecture_threshold_sweep / XGBoost+SBERT@thresh	0.903 (0.848–0.939)
≥ 0.95	887	0.098 (0.085–0.112)	single_architecture_threshold_sweep / XGBoost+SBERT@thresh	0.952 (0.907–0.975)

Panel B — Disagreement-stratified clinician-review policy

Stratum (n architectures flagging)	Recommended policy	N messages	N hazards (true)	Hazard prevalence in stratum (95% CI)	Share of total messages
0-1 flagging (autonomous benign)	no action	244	10	0.041 (0.022–0.074)	12.2%
2-3 flagging (clinician review)	clinician review	1250	76	0.061 (0.049–0.075)	62.5%
4-6 flagging (clinician review)	clinician review	386	49	0.127 (0.097–0.164)	19.3%
7-13 flagging (autonomous escalation)	escalate	120	30	0.250 (0.181–0.334)	6.0%

Caption: Panel A — for each of four sensitivity floors (≥ 0.80, ≥ 0.85, ≥ 0.90, ≥ 0.95), the configuration achieving the maximum specificity at or above that floor across the full pool of evaluated configurations (single architectures at all observed thresholds, ensembles,

cascades, multi-model consensus rules, retrieval-augmented generation). Panel B — distribution of the 2000 real-world test-set messages across agreement strata (number of architectures flagging each message), with the policy recommendation for clinician review or autonomous disposition at each stratum, the resulting hazard prevalence within the stratum, and the share of total message volume comprising the stratum.

Table 7. Two deployment policies vs individual-component baselines: empirical alert volume and residual miss rate per 1,000 messages.

Configuration	Sensitivity (95% CI)	Specificity (95% CI)	PPV (95% CI)	Alerts per 1,000 messages (95% CI; clinician review load)	Hazards caught per 1,000 messages (95% CI)	Hazards missed per 1,000 messages (95% CI)	Note
Autonomous (physician-unassisted) triage benchmark	≥ 0.800	≥ 0.800	—	—	—	≤ 16.5	Anchored to IDx-DR pivotal trial (sens 87.2%, spec 90.7%)
Best balanced single architecture (ActionHead (action recommender))	0.424 (0.351–0.501)	0.802 (0.783–0.820)	0.162 (0.130–0.199)	217 (199–235)	35.0 (27.8–44.0)	47.5 (39.0–57.7)	default threshold; baseline reference
Highest-sensitivity single architecture (Logistic regression + TF-IDF)	0.970 (0.93	0.093 (0.08	0.088 (0	913 (899–924)	80.0 (68.9–92.7)	2.5 (1.1–5.8)	default threshold; reference for sens-floor analysis

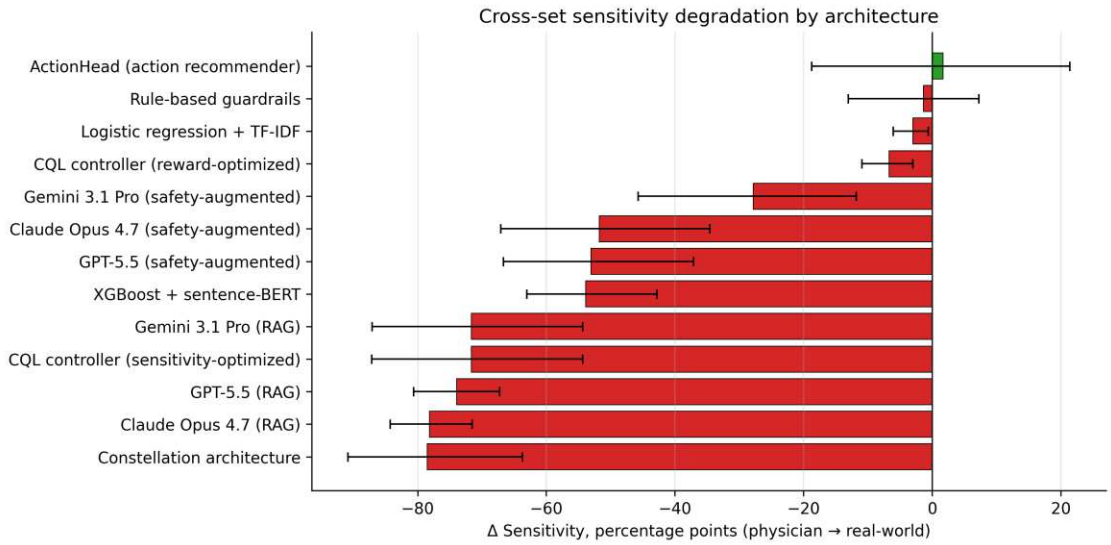
Configuration	Sensitivity (95% CI)	Specificity (95% CI)	PPV (95% CI)	Alerts per 1,000 messages (95% CI; clinician review load)	Hazards caught per 1,000 messages (95% CI)	Hazards missed per 1,000 messages (95% CI)	Note
	1–0.987)	0–0.107)	.076–0.102)				
Best balanced ensemble (soft_f1weighted)	0.582 (0.506–0.654)	0.581 (0.558–0.603)	0.111 (0.092–0.134)	433 (411–454)	48.0 (39.5–58.3)	34.5 (27.4–43.4)	hard- or soft-voting; reference for ensemble closing-the-gap finding
Best balanced cascade (ActionHead (action recommender) × Logistic regression + TF-IDF)	0.418 (0.346–0.494)	0.809 (0.791–0.827)	0.165 (0.132–	209 (192–228)	34.5 (27.4–43.4)	48.0 (39.5–58.3)	two-stage AND-rule; reference for cascade closing-the-gap finding

Configuration	Sensitivity (95% CI)	Specificity (95% CI)	PPV (95% CI)	Alerts per 1,000 messages (95% CI; clinician review load)	Hazards caught per 1,000 messages (95% CI)	Hazards missed per 1,000 messages (95% CI)	Note
			0.203)				
Policy A: high-recall screen (LogReg+TF-IDF@thresh)	0.855 (0.793–0.900)	0.282 (0.262–0.303)	0.097 (0.083–0.113)	729 (709–748)	70.5 (60.1–82.6)	12.0 (8.1–17.8)	sens >= 0.85 floor; all flagged messages sent to clinician confirmation queue
Policy B: disagreement-stratified triage (2 or more of 10 architectures flag, route to clinician review)	0.939 (0.892–0.967)	0.128 (0.113–0.144)	0.088 (0.076–0.102)	878 (863–892)	77.5 (66.6–90.0)	5.0 (2.7–9.2)	0-1 architectures flagging: autonomous benign; 2 or more: clinician review; 7 or more: autonomous escalation

Caption: Per-configuration sensitivity, specificity, positive predictive value, alert count per 1,000 messages (count of positive predictions), hazards caught per 1,000 messages (count of true positives), and hazards missed per 1,000 messages (count of false negatives) on the 2000-message real-world test set. All quantities are computed directly from empirical sensitivity and specificity on the held-out test set; the table reports counts of alerts and counts of misses rather than estimates of clinician time, full-time equivalents, or institutional staffing capacity, none of which were measured in the present study. The autonomous (physician-unassisted) triage benchmark row is included as the reference benchmark for hypothetical fully-autonomous deployment; the 165-hazard distribution would admit up to 16.5 missed hazards per 1,000 messages under the sensitivity ≥ 0.80 floor. Policy A and Policy B both produce residual miss rates below the symmetric-benchmark-implied maximum while producing alert volumes 2.53-fold (Policy A) and 3.03-fold (Policy B) higher than the best balanced single-architecture operating point (XGBoost with sentence-BERT embeddings at default threshold, 288 alerts per 1,000 messages).

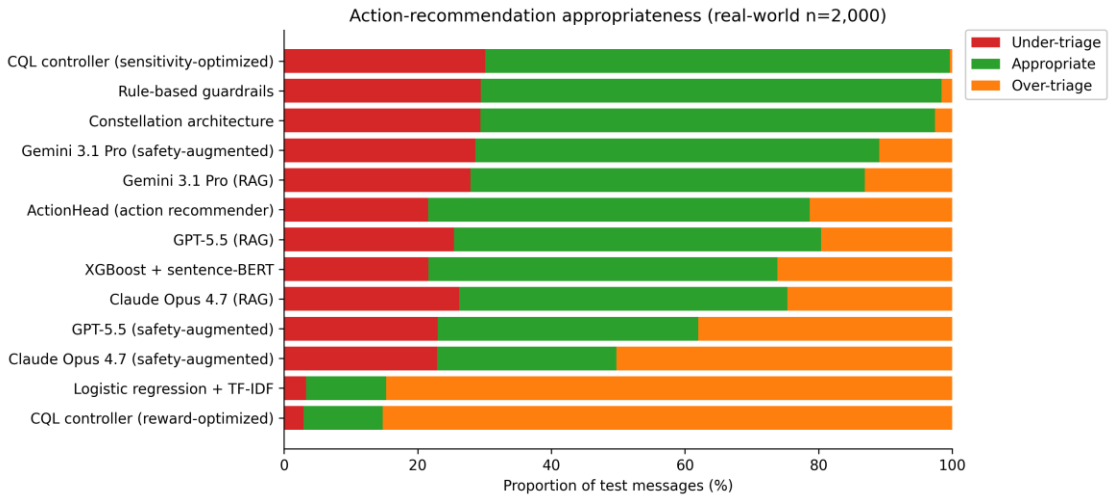
Figures

Figure 1. Per-architecture sensitivity change from physician-scripted to real-world Medicaid test sets.



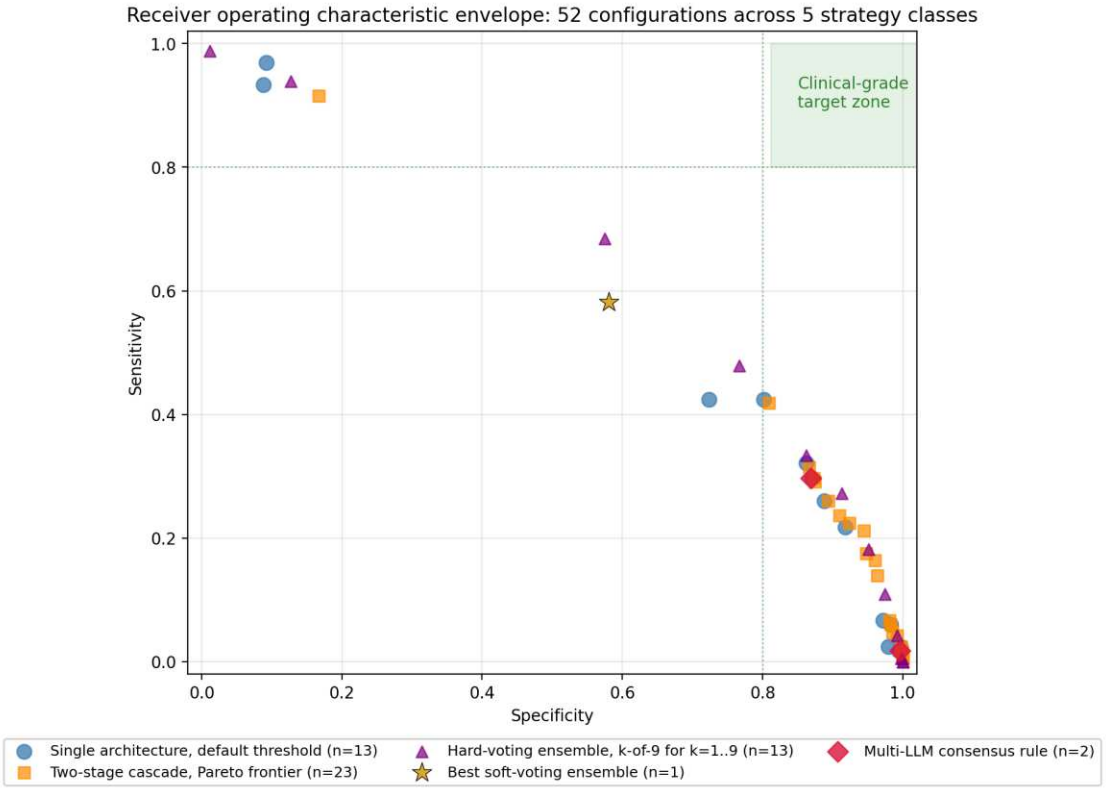
Caption: Each architecture's sensitivity on the physician-scripted 41-case comparison set is plotted against its sensitivity on the held-out 2000-message real-world test set. The diagonal indicates equal performance across populations; architectures below the diagonal lose sensitivity when moving from the physician comparison scenarios to the real-world Medicaid messaging population. Error bars reflect Wilson 95% confidence intervals.

Figure 2. Action recommendation appropriateness on the real-world test set, by architecture.



Caption: Stacked-bar chart of the proportion of recommendations classified as appropriate (matching the physician-adjudicated reference action), under-triage (recommending a less urgent action than the reference), or over-triage (recommending a more urgent action) on the 8-point ordinal action scale, on the real-world test set.

Figure 3. Receiver operating characteristic envelope across all evaluated strategies on the real-world test set.



Caption: Each marker is one configuration. Markers of the same shape and color are different configurations within the same strategy class. The legend reports the number of configurations per class: 9 single-architecture default-threshold operating points (blue circles, one per architecture), 15 two-stage cascade Pareto-frontier pairs (orange squares), 9 hard-voting k-of-9 ensemble configurations for k from 1 to 9 (purple triangles), 1 best soft-voting ensemble (gold star), and 2 multi-model consensus rules (red diamonds). Many configurations cluster near the high-specificity / low-sensitivity corner, where overlapping markers of different shapes may look like a single composite marker. The shaded green zone in the upper-right is the autonomous (physician-unassisted) triage benchmark (sensitivity 0.80 or higher and specificity 0.80 or higher). No configuration reaches the benchmark zone, the structural finding of this study.

Supplementary Files

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

- BasuBMCMIDMadditionalfile1supplementarymethodsfigurestables.docx