GUIDELINE-AS-ORACLE: RULE-DERIVED LABELS TEACH A TRIAGE AGENT WHEN TO ESCALATE

Anonymous authors
Paper under double-blind review

ABSTRACT

Can a published clinical guideline replace per-dialogue expert labels when training an interactive triage agent? We study ophthalmic telephone triage: a guideline, compiled by hand into a 70-row rule table, labels every synthetic training dialogue with the most severe tier among the rows its caller profile cites, so no dialogue is labeled by a person. On a 201-case development reference adjudicated by the authors, fine-tuning a 9B agent raises emergent recall in simulated calls from 9.5% to 69.0% and halves under-triage, and this escalation gain holds in every evaluated setting (not fully crossed): five training runs, two software environments, two caller simulators, and five reference-robustness views, including a clinician's review of the flagged cases. The rule-derived labels carry the escalation gain: controls trained on the same dialogues with the final tier words, or entire final messages, permuted raise the same topics but call 0 and 2 of 42 emergent cases emergent, while the first control's next-step recommendations, whose training text was not permuted, still track the rule-derived tier; given the complete case, the untuned model still calls only 7 of 42 emergent cases emergent. Training shifts the balance between under- and over-triage, by amounts that vary across runs of one configuration. Agreement rises from 61.7% to 74.1% in the pre-specified comparison, but over-triage ranges from 26 to 54 calls across runs, so the agreement difference ranges from $+12.4$ to -5.5 points. Every run costs less than its base model once an under-triaged call counts more than 1.41 over-triaged ones, and significantly less in all five once it counts three. On 40 external vignettes the trained model escalates more but does not agree more. The main results measure agreement with a development reference, not clinical validity.

1 INTRODUCTION

In ophthalmic telephone triage, a nurse decides from a conversation alone whether a caller needs emergent, urgent, or routine care, based on a few decisive facts such as an eye procedure last week, contact-lens wear, or a shadow over part of the visual field. An LLM agent for this task must learn both what to ask and how to map what it hears to a disposition, and clinician-labeled calls that could supervise either are costly and rarely shareable.

Specialty societies already publish guidance that maps presentations to urgency. We ask whether that guidance can replace per-dialogue annotation and, if it can, what it teaches. We compile American Academy of Ophthalmology (AAO) materials (American Academy of Ophthalmology, 2026) into a 70-row table in which each row pairs a trigger with a tier. Synthetic caller profiles cite rows of the table, and each dialogue generated from a profile inherits the most severe tier among its cited rows. We call this construction **Guideline-as-Oracle (GAO)** and the fine-tuned 9B model **GAO-Triage** (Figure 1). "Oracle" names a construction role: the table is an author-compiled operationalization, not a clinically validated standard.

Agreement alone cannot separate learning to ask from learning to decide, or under-triage from over-triage. A synthetic dialogue supervises both the questions and the final tier, and triage errors are asymmetric: missing an emergency can cost sight, whereas an unnecessary escalation costs time. We therefore evaluate GAO-Triage in simulated calls on 201 cases, separating questions from decisions and pricing errors by direction, across five training runs, two software environments, two caller simulators, and five reference-robustness views; these checks are not fully crossed.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

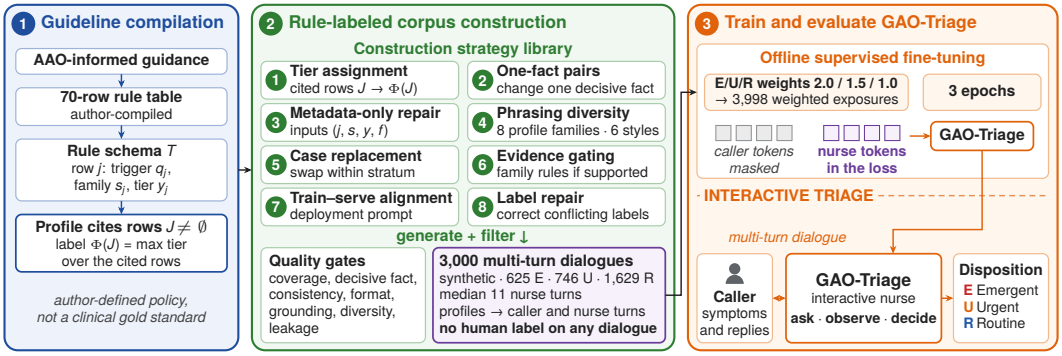


Figure 1: Overview of Guideline-as-Oracle (GAO). **(1)** Guideline compilation: the authors compile AAO-informed guidance into a 70-row rule table; a synthetic caller profile cites a nonempty row set J and inherits $\Phi(J)$, the most severe tier among the cited rows (Eq. (3)). **(2)** Rule-labeled corpus construction: role-separated LLM generators turn profiles into 3,000 multi-turn dialogues; numbered boxes summarize strategies used across corpus versions (Appendix C.14). **(3)** Training and evaluation: class-weighted fine-tuning on nurse tokens yields GAO-Triage, an interactive nurse. The rule table is written by people, but no training dialogue receives a human-assigned label.

Three findings emerge. (i) *Rule-derived labels reliably teach escalation.* In the pre-specified comparison, emergent recall rises from 4/42 to 29/42 and under-triage falls from 51 to 25 calls; the emergent gain holds in every training run, environment, caller simulator, and reference-robustness view (Section 4.1). (ii) *The labels teach when to escalate, not what to ask.* Controls trained on the same dialogues with permuted tier words or permuted final messages raise the same topics but call 0 and 2 of 42 emergent cases emergent, while the first control's unpermuted next-step recommendations still track the rule-derived tier; given the complete case, the untuned model identifies only 7 of 42 (Section 4.2). (iii) *Training moves the operating point, by an amount that varies between runs.* Agreement rises from 61.7% to 74.1% in the primary test (p=0.0046), but over-triage varies from 26 to 54 calls across runs of one configuration, and two such runs differ in agreement by 9.5 points (p=0.018). Every run costs less than its base model once an under-triaged call is priced above 1.41 over-triaged ones, significantly so in all five from a ratio of three (Section 4.3). The main results measure agreement with an author-adjudicated development reference, and on 40 external vignettes the trained model escalates more but does not agree more (Section 4.5).

Our contributions are:

- **Supervision.** A pipeline in which every training label of an interactive clinical agent traces to cited rows of a compiled guideline.
- **Mechanism.** Diagnostic evidence that rule-derived labels teach the decision rather than the questioning, from two label permutations (tier word only; entire final message), a full-information condition, and the positions of decisive facts in transcripts; permuting the entire final message removes the escalation gain.
- **Evaluation.** A replicated, direction-aware protocol for triage agents (runs, environments, robustness views, cost curves, run-level prediction intervals) that shows how single-run agreement misreads an operating-point shift.

2 Method: Guideline-as-Oracle

2.1 Problem Formulation

Triage as an interactive decision. A case is a caller profile x: a chief complaint plus hidden facts such as onset, laterality, pain, redness, visual symptoms, contact-lens wear, recent eye surgery, and injury or chemical exposure. A nurse policy π_θ talks with a caller who reveals hidden facts only when asked. At turn t, given the history h_t, the nurse either asks one question or ends the call with a disposition $\hat{y} \in \mathcal{Y} = \{\text{Routine}, \text{Urgent}, \text{Emergent}\}$, ordered Routine $\prec$ Urgent $\prec$ Emergent (routine appointment, care within 24–48 hours, emergency care now); the final turn also gives a reason that

may cite only facts the caller stated. Let $y^\star(x)$ be the reference tier. Under-triage ($\hat{y} \prec y^\star$) can delay emergency care, whereas over-triage ($\hat{y} \succ y^\star$) costs a visit. With $k > 0$ the cost ratio of under- to over-triage, the expected cost of a policy is

$$C_k(\pi_\theta) = \mathbb{E}_x \big[\, k\, \mathbb{I}\{\hat{y} \prec y^\star(x)\} + \mathbb{I}\{\hat{y} \succ y^\star(x)\} \,\big]. \tag{1}$$

Agreement weighs both errors alike (k=1, up to calls without a disposition). If π_θ under-triages less than a comparator π_0 ($U_\theta < U_0$), it costs less exactly when k exceeds the break-even ratio

$$k^* = \frac{O_\theta - O_0}{U_0 - U_\theta}, \tag{2}$$

where U and O denote under- and over-triage rates; $k^* \leq 0$ means that π_θ costs less at every k.

Supervision from a rule table. Instead of labeling dialogues, GAO labels profiles. A rule table $\mathcal{T} = \{(q_j, s_j, y_j)\}_{j=1}^m$ maps each textual trigger q_j and symptom family s_j to a tier y_j. A synthetic profile cites a nonempty set $J \subseteq \{1, \ldots, m\}$ of instantiated rows and receives

$$\Phi(J) = \max_{j \in J} y_j, \tag{3}$$

so every training label is a deterministic function of named rows. A generator turns the profile into a dialogue d whose final nurse turn states $\Phi(J)$, and the nurse is trained on its own turns only:

$$\mathcal{L}(\theta) = -\,\mathbb{E}_{d \sim \mathcal{D}_w} \sum_{t \in N(d)} \log \pi_\theta(a_t \mid h_t), \tag{4}$$

where $N(d)$ indexes the nurse turns a_t of d (each log-probability sums over the turn's tokens; caller tokens are masked from the loss) and $\mathcal{D}_w$ resamples dialogues with class weights $w_{\Phi(J)}$.

2.2 Compiling the Guideline

The authors compiled AAO Preferred Practice Pattern materials (American Academy of Ophthalmology, 2026) into an initial 64-row table, extended to m=70 rows (38 emergent, 22 urgent, 10 routine) after error analysis on the development cases of Section 3; no case text was copied into the table. The symptom families s_j are vision loss or change, eye pain, flashes and floaters, redness or discharge, photophobia, eyelid, pupil, and other. Rows are operational triggers: *"Redness or discharge in a contact lens wearer"* is emergent; *"Mild ocular pain if accompanied by redness"* is urgent. The table, released with the supplementary material, has no map from rows to source passages and has not been validated by an independent clinician panel.

By Eq. (3), a label is fixed when its profile is written. *One-fact boundary pairs* are two profiles that differ in one decision-critical fact and in the row that fact triggers, with no other cited row at or above the higher tier, so their labels differ by construction.

2.3 From Rows to Training Dialogues

Figure 1 shows the pipeline, and Appendix Figure 5 traces one training example through it.

Profiles. For a target family and tier, a teacher LLM writes a structured profile containing an opening line from the chief complaint, hidden facts, cited rows J, and caller attributes such as emotion, health literacy, and willingness to volunteer information. Profiles of the same family and tier provide schema examples in the prompt. Some seeds were abstracted from public patient posts; generation rules excluded evaluation posts, although later audits traced paraphrase clusters to development errors (Section 7). A validator checks that the cited rows exist and that the label equals $\Phi(J)$; whether hidden facts instantiate the cited triggers receives only limited consistency checks (Appendix B).

Dialogues. Separate LLM roles generate caller and nurse turns. The caller sees the profile but is instructed to reveal hidden facts only when asked. Family-specific instructions give the nurse a priority order: follow up on volunteered concerns, resolve the active family's highest-risk undecided trigger, cover five safety items when relevant (laterality, onset, contact-lens use, recent eye surgery or procedure, and injury or chemical exposure), check the other families, and close with a final sweep. Calls target 7 to 12 nurse questions but escalate early on a decisive emergent finding. Five LLM families, templates, and scripted sources generated the dialogues (Appendix A).

Quality gates. A label validator requires the final tier to equal $\Phi(J)$ and to be supported by the most severe cited row. The seven intended gates (Table 5) check role and verdict format, caller-answer consistency with the profile, caller disclosure of decisive facts before the verdict, whether every finding in the reason appears in a caller turn, quotas over the 24 family-by-tier cells, template repetition, and 8-gram overlap with evaluation text. The intended repair regenerates a rejected dialogue from its row identifier, family, tier, and failure tag alone; gates and repair evolved across corpus versions, and audits found historical batches that violated them (Appendix D).

Corpus. The final corpus has 3,000 dialogues (625 emergent, 746 urgent, 1,629 routine) with a median of 11 nurse turns, including 160 urgent one-fact boundary dialogues. Class-weighted resampling (w_{Emergent}=2.0, w_{Urgent}=1.5, w_{Routine}=1.0) yields 3,998 training exposures. Every closing nurse turn states the tier $\Phi(J)$, a reason citing caller-stated facts, a next step with tier-fixed wording, and a safety net. Class mix, gates, and generators changed across corpus versions (Appendix C.14); only the boundary pairs were ablated.

2.4 Training

We minimize Eq. (4) over all Qwen3.5-9B-Base parameters for three epochs on the resampled corpus (cosine-decayed learning rate 5×10^{-6}, effective batch size 32, maximum length 4,096, DeepSpeed ZeRO-3). Training prompts include a family-specific block; evaluation uses a generic nurse prompt (Appendix B). No training dialogue receives a human label.

3 Experimental Setup

Reference. The 201 evaluation cases are public, patient-authored Reddit posts about eye symptoms, each converted into an interactive profile with account handles and URLs removed; each carried an original tier label. We flagged 25 labels for review against the table. Two LLM auditors (A: DeepSeek-V3; B: Qwen3-235B) re-derived every tier with the table but without the labels, and the authors adjudicated the flagged cases, changing 12 labels (9 downgrades, 3 upgrades; Appendix D). The resulting *reference* has 42 emergent, 47 urgent, and 112 routine cases. A clinician also reviewed the 25 flagged cases and agreed with 9 of the 12 revisions; the clinician-adjudicated labels differ from the reference in 3 cases (Appendix H.1). We keep our adjudication as the reference because the primary test was pre-specified on it. These cases informed the table's extension from 64 to 70 rows, corpus development, and checkpoint selection, so agreement measures consistency with a policy-shaped development reference, not held-out generalization or clinical accuracy.

Reference-robustness views. We rescore the main comparisons under five views (Section 4.4): the clinician-adjudicated labels, the original labels, each LLM auditor's tiers, and the 124-case subset on which all labelings agree. None is an independent clinical reference: the clinician worked within our review workflow, and the auditors read the rule table.

Interactive protocol and environments. GPT-5.5 plays the caller at temperature 0. It sees the full profile and is instructed to answer only the question asked and to reveal hidden facts gradually. The nurse opens with a fixed greeting, then decodes greedily for up to 20 nurse turns; the tier is read from the final nurse turn, and a call without a parseable tier is an error. DeepSeek-V4-Pro serves as a second caller. The main comparisons use *overwrite-free* runs, without the post-hoc heuristic that had overwritten verdicts in earlier runs; historical results in the text and appendices are marked as such (Appendix B.6). We used two software environments: the first for every arm of Table 1; the second, on a separate cluster with a different inference stack and caller gateway, for three further training runs, two label-permutation controls, and a re-evaluated base model (Appendix I). The same base weights change about a quarter of their verdicts across environments, so we compare arms only within one environment.

Compared systems. We retrain GAO-Triage from its archived configuration and SHA-256-verified corpus, because the original checkpoint was not retained. Five training runs share that configuration: runs 1 and 2 in the first environment with the default seed 42, and runs 3, 4, and 5 in the second with seeds 42, 43, and 44, each using its final weights, with no checkpoint selection. Unless otherwise specified, GAO-Triage refers to run 1. The untuned Qwen3.5-9B-Base is

Table 1: Overwrite-free results on 201 development-reference cases (42 emergent; first environment, GPT-5.5 caller); GAO-Triage is run 1. Agr.: agreement (%). Δ: arm minus GAO-Triage in points, with paired 95% interval. p: exact McNemar test against GAO-Triage; only the comparison with the base model is primary, and other p-values are uncorrected. E rec.: emergent recall (%, hits of 42). No v.: calls without a disposition (errors). Msgs.: mean messages per call, nurse and caller. $\hookrightarrow$: variant of the row above. (†) GRPO (Section 6) starts from the original SFT weights, not the retrained GAO-Triage. Confusion matrices: Table 14.

Nurse	Agr.	Δ	95% CI	p	E rec.	No v.	Msgs.
GAO-Triage (ours)	74.1	—	—	—	69.0 (29)	1	20.2
$\hookrightarrow$ second run	73.1	-1.0	$[-7.3, 5.3]$	0.878	78.6 (33)	0	20.6
Base 9B (untuned)	61.7	-12.4	$[-20.6, -4.3]$	0.0046	9.5 (4)	0	20.2
Always routine (floor)	55.7	-18.4	$[-26.7, -10.1]$	3.8×10^{-5}	0.0 (0)	0	—
Outcome GRPO (†)	45.3	-28.9	$[-37.5, -20.2]$	1.9×10^{-9}	42.9 (18)	80	13.6
GPT-5.5	75.1	$+1.0$	$[-6.7, 8.7]$	0.899	64.3 (27)	0	12.3
$\hookrightarrow$ given the rule table	67.7	-6.5	$[-14.2, 1.2]$	0.130	66.7 (28)	0	16.0
GPT-5.4	70.6	-3.5	$[-11.4, 4.5]$	0.464	64.3 (27)	0	7.9
DeepSeek-V4-Pro	71.6	-2.5	$[-9.7, 4.7]$	0.590	40.5 (17)	0	8.7
GLM-5.2	69.2	-5.0	$[-13.3, 3.4]$	0.295	45.2 (19)	0	8.6
Qwen3.5-397B	68.2	-6.0	$[-13.2, 1.3]$	0.141	73.8 (31)	0	4.7
Kimi-K2.6	71.1	-3.0	$[-10.8, 4.8]$	0.532	76.2 (32)	2	7.9
Qwen3-235B	62.2	-11.9	$[-19.7, -4.2]$	0.0043	35.7 (15)	0	3.8

the main baseline, and an always-routine rule gives the majority-class floor. Eight general-purpose nurse arms cover seven models: GPT-5.5 with and without the 70-row table in its prompt, GPT-5.4, DeepSeek-V4-Pro, GLM-5.2, Qwen3.5-397B, Kimi-K2.6, and Qwen3-235B.

Metrics. We report agreement with the reference, emergent-reference recall, under- and over-triaged calls, calls without a disposition, and messages per call. Following cost-sensitive learning (Elkan, 2001), we price errors with Eq. (1), by convention counting a missing disposition as routine because the caller receives no instruction. Cost curves (Drummond & Holte, 2006) plot cost differences over k, and we report each run's break-even ratio k^* (Eq. (2)) against its base model.

Statistics. We pair all comparisons over the same cases, using exact McNemar tests, 95% Wald intervals for paired differences, and paired bootstrap intervals for cost differences (10,000 case resamples). Inference is conditional on the evaluated checkpoints; Table 2 and a post hoc run-level summary (Section 4.1) add run-to-run variation. The sole primary test compares GAO-Triage and base-model agreement in the first environment. Its endpoint and decision rule were fixed in a written plan before scoring, but the plan has no external timestamp, so we call it pre-specified, not registered. Other p-values are secondary and uncorrected unless stated. At n=201, paired 95% agreement intervals extend 6 to 9 points each way; non-significance does not establish equivalence.

4 Results

4.1 Rule-Derived Labels Teach Escalation

GAO-Triage agrees with the reference on 149 of 201 cases, against 124 for its untuned base model (Table 1; paired difference 12.4 points, 95% CI 4.3–20.6, exact McNemar p=0.0046). The gain is concentrated in emergent cases: emergent recall rises from 4/42 to 29/42 (26 versus 1 discordant pairs, $p=4.2\times10^{-7}$, significant under Holm correction across the three class-level tests), while urgent recall (29/47 for GAO-Triage versus 32/47 for the base model) and routine recall (91/112 versus 88/112) do not change significantly. The base model says "emergent" in only 6 of its 201 calls. Under-triage falls from 51 to 25 calls (paired difference -12.9 points, 95% CI -19.0 to -6.9, $p=6.9\times10^{-5}$), over-triage stays at 26 calls, and emergent cases called routine fall from 11 to 5.

Across the five further settings of Table 2, every run calls 23 to 33 of the 42 emergent cases emergent, against at most 4 for its base model, and every run with full counts cuts under-triage by 26 to 34 calls

(all $p<10^{-4}$). The agreement gain holds in the first environment, including with the DeepSeek-V4-Pro caller, but not in the second, whose three runs differ from their base model by -5.5 to $+4.0$ points, none significantly. Over-triage separates the runs, ranging from 26 to 54 calls (Section 4.3). For the five runs with the GPT-5.5 caller, a post hoc i.i.d. normal approximation (t, df$=4$) gives nominal 95% prediction intervals on this reference of 14.6 to 37.4 additional emergent hits and 18.5 to 40.7 fewer under-triaged calls, but -17.5 to $+27.3$ points of agreement (mean $+4.9$, 95% CI -4.3 to 14.0). These intervals are descriptive: runs share bases, cases, and seeds across two environments, so coverage for a new run is unverified.

Takeaway 1. Rule-derived labels reliably teach escalation: in all five runs, emergent recall rises from $\leq 4/42$ to 23–33/42 and under-triage falls by 26–34 calls, but agreement rises significantly only in environment 1.

4.2 The Labels Teach When to Escalate, Not What to Ask

A dialogue supervises both questions and tier; three analyses separate the two (Figure 2).

Questions come from the dialogues. GAO-Triage raises eye surgery or procedures in 186 of 201 calls against 96 for its base model (96 versus 6 discordant calls, $p=5.7\times10^{-22}$) and injury or chemical exposure in 179 against 134, with the same mean number of messages (20.2; Appendix Table 28). These topics define 15 of the table's 38 emergent rows and were prioritized during generation (Section 2.3). To test whether asking them carries the gain, a control trained alongside the second-environment runs on the same 3,998 exposures permutes only the tier words of final nurse messages, preserving class counts and all other text (Appendix I). It raises these topics about as often as the three runs or more (eye procedures in 190 calls against 180 to 195, all $p\geq0.06$; injury or chemical exposure in 187 against 174 to 183) but answers routine in 183 of 201 calls and never urgent: agreement is 100/201, below even the base model (-14.9 points, $p=1.0\times10^{-4}$), and emergent recall is 0/42.

The decision comes from the labels. The permutation left each tier's fixed next-step sentence (Section 2.3) in place; read post hoc, that sentence agrees with the reference in 126 of 201 calls, including 25 of 42 emergent cases, not significantly fewer than the third run's tier word (138; $p=0.21$), while the control's own tier word and next-step sentence agree in only 97. A second control permutes the entire final message and keeps every earlier turn (Appendix I): it still raises eye procedures in 190 calls but answers routine in 188 and calls 2 of 42 emergent cases emergent (2.7 expected by chance), and 1 and 0 in two runs given the complete post. Keeping the questions and earlier turns did not rescue escalation once the final message was permuted.

The base model's deficit is in deciding. Given the complete post as the caller's first message, so that no fact has to be elicited, the base model still calls only 7 of the 42 emergent cases emergent, against 28 to 33 for the three second-environment runs (all $p<10^{-5}$). The same deficit appears in interactive calls: in 16 of the 22 hand-verified procedure or injury cases with an emergent reference (Appendix H), the caller states the fact in the first message in every arm, and given that opening the base model escalates 3 of the 16 calls, GAO-Triage 12, its second run 14, and GPT-5.5 10. Elicitation

Table 2: GAO-Triage (G) versus base (B) across runs. Each row pairs a run with its same-environment (1 or 2), same-caller base; repeated B entries reuse one evaluation. Agreement, emergent hits (of 42), Under, and Over are counts; run 1 with the GPT-5.5 caller has one missing disposition. Δ is the paired agreement difference in points (G minus B), with its exact McNemar p; k^* is the cost ratio above which the run costs less than its base. The DeepSeek-V4-Pro caller evaluations retain per-class and discordant-pair counts, but not per-case verdicts.

Env., caller	Run	Agr. G / B	Δ	p	E G / B	Under G / B	Over G / B	k^*
1, GPT-5.5	Run 1 (primary)	149 / 124	+12.4	0.0046	29 / 4	25 / 51	26 / 26	0.00
	Run 2	147 / 124	+11.4	0.013	33 / 4	23 / 51	31 / 26	0.18
1, DeepSeek-V4-Pro	Run 1	145 / 124	+10.4	0.022	30 / 3	—	—	—
2, GPT-5.5	Run 3	138 / 130	+4.0	0.46	32 / 3	22 / 55	41 / 16	0.76
	Run 4	134 / 130	+2.0	0.74	30 / 3	21 / 55	46 / 16	0.88
	Run 5	119 / 130	−5.5	0.27	23 / 3	28 / 55	54 / 16	1.41

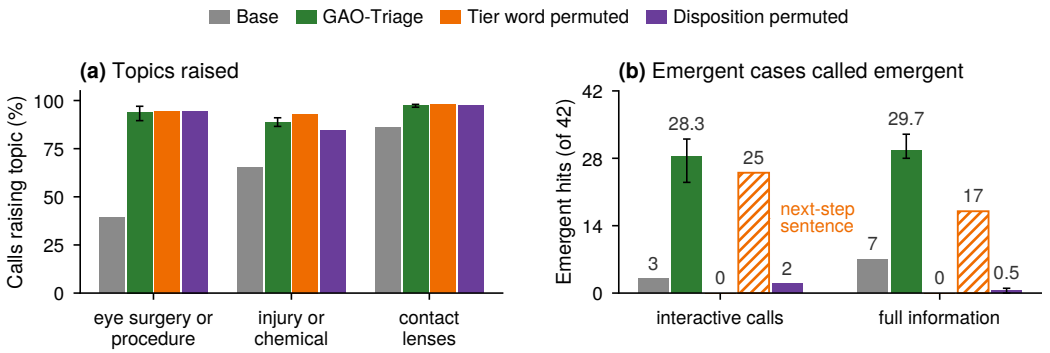


Figure 2: Questioning and disposition output (second environment). **(a)** Share of 201 calls in which a nurse turn other than the verdict raises each topic (keyword coding; Appendix G). **(b)** Emergent cases (of 42) called emergent in interactive calls and with the complete post as the caller's first message. Hatched: the tier-word control's next-step sentence, which kept the rule-derived disposition (post hoc); purple: the control whose entire final message was permuted. Bars for runs 3–5 and for the two purple full-information runs show the mean and range, not a confidence interval.

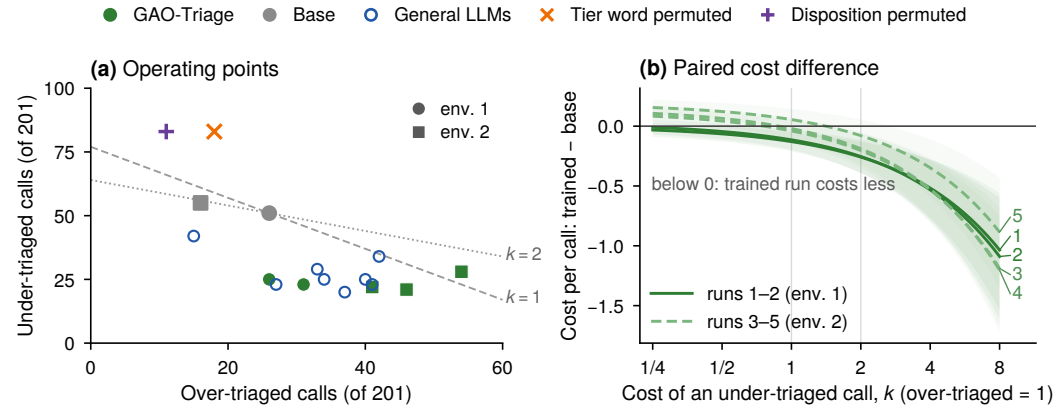


Figure 3: Rule-derived labels move the operating point; both panels count calls without a disposition as routine. **(a)** Under- versus over-triaged calls (of 201) for the GAO-Triage runs (green), their base models (gray), the general-purpose nurses, and the two permutation controls (tier word: cross; entire final message: plus); circles are the first environment and squares the second (Appendix I). Dashed (k=1) and dotted (k=2) lines mark equal cost to the first-environment base model. **(b)** Cost per call of each GAO-Triage run minus that of its within-environment base model, as a function of k, with pointwise paired bootstrap 95% bands; below zero, the trained run costs less.

also improves (9 verified facts never surface with the base model, against 2 with GAO-Triage), but it cannot alone explain the gain.

Remaining misses. In 10 of the 13 emergent cases that GAO-Triage missed, the transcript contains every fact an emergent row requires (Appendix G); most misses, like the gain, lie in mapping stated facts to a tier.

Takeaway 2. The labels teach *when* to escalate, not *what* to ask: controls raising the same topics lose the gain wherever the label is permuted, and the untuned model under-escalates even given every fact.

4.3 An Operating-Point Shift

Agreement weighs a missed emergency and an unnecessary escalation equally. Figure 3a places the nurses by their two error types. Every GAO-Triage run cuts under-triage by 26 to 34 calls relative

to its base model, but the runs differ in over-triage: the first run keeps it at 26 calls, whereas the third, fourth, and fifth runs raise it from 16 to 41, 46, and 54 calls (all $p \leq 1.1 \times 10^{-4}$). Because agreement counts both errors alike, agreement varies across runs of one configuration: the fifth run agrees on 19 fewer cases than the third ($p=0.018$). The break-even cost ratio k^* ranges from 0 to 1.41 (Table 2). Paired cost-difference intervals of the second-environment runs include zero at $k=1$; those of four of the five runs lie below zero at $k=2$, and all five at $k=3$ (Figure 3b; Appendix I). Every run therefore costs significantly less than its base model once an under-triaged call counts as three over-triaged calls or more, even where agreement does not rise. Full-information evaluation and charging per tier of under- or over-triage both leave every k^* below 1.41 (Appendix I).

General-purpose models occupy the same region as GAO-Triage (Figure 3a). Only Qwen3-235B has significantly lower agreement ($\Delta = -11.9$ points, 95% CI -19.7 to -4.2); the other seven intervals include zero (Table 1). Operating points vary widely: DeepSeek-V4-Pro under-triages 42 calls and over-triages 15, whereas Kimi-K2.6 over-triages 37. Paired cost intervals of GPT-5.5 versus GAO-Triage include zero at $k=1$, 2, and 3 (Appendix I). GPT-5.5 also plays the caller, which may favor it; giving it the 70-row table lowers its agreement (67.7% versus 75.1%).

Takeaway 3. Training moves the operating point, which agreement alone misreads: across runs of one configuration, agreement changes by -5.5 to $+12.4$ points, yet every run costs less than its base model once an under-triaged call counts more than 1.41 over-triaged ones.

4.4 Robustness to the Reference

Under the clinician-adjudicated labels, which revise only flagged cases, GAO-Triage gains 11.4 points (95% CI 3.4–19.5, $p=0.0086$), with emergent recall 29/44 against 4/44; Appendix Table 31 rescores the primary comparison under all five reference-robustness views of Section 3. The emergent-recall gain holds under every view (all $p<0.004$), for the second run ($p \leq 0.001$), and for every second-environment run (Appendix I). The overall gain holds under the original labels ($+10.0$ points, $p=0.022$), auditor A ($+12.9$, $p=0.0034$), and the 124-case consensus subset ($+11.3$, $p=0.044$), but not under auditor B, whose labels include fewer emergent (26) and more urgent (49) cases ($+3.0$, 95% CI -5.2 to 11.1). Against GPT-5.5, no view shows a difference larger than 1.5 points (Appendix H). In the second environment no run gains significantly under any view, and the fifth run falls below its base model under two (Appendix I). Who adjudicates the labels changes the size of the overall gain but not its emergent component.

4.5 External Vignettes

Every result so far is scored against a reference that our policy informed. Oxford-40 has 40 expert-reviewed vignettes written by others (Mittal & Aggarwal, 2026); we give each vignette to the nurse as the caller's first message. With this input, the three second-environment runs agree with 25, 26, and 24 of the 40 labels, against 26 for their base model (all $p \geq 0.77$); zero-shot GPT-5.5 reaches 33 ($p=0.039$ against the third run). Every run calls all 23 emergent vignettes emergent (base model 18), but also calls 12 or 13 of the 14 urgent vignettes emergent (base model 6). On these exam-rich vignettes, training moves the operating point toward escalation without raising agreement, and a constant-emergent answer already scores 57.5% (Appendices D.8 and I). The check gives no evidence that the agreement gain transfers beyond our reference.

5 Related Work

Rule-derived supervision. Data programming aggregates noisy labeling functions into training labels (Ratner et al., 2016; 2017), rule distillation moves logic rules into network weights (Hu et al., 2016), and other work derives training data from written principles or model outputs (Bai et al., 2022; Wang et al., 2023; Taori et al., 2023; Yu et al., 2025). GAO assigns each synthetic dialogue its tier directly from cited rows; its learner must elicit the facts that tier depends on.

Clinical dialogue agents. Recent agents learn clinical inquiry and decisions through reinforcement learning (Lai et al., 2026; Cao et al., 2026; Bani-Harouni et al., 2026), route cases through a learned triage agent (Xia et al., 2026), or abstain until evidence suffices (Dang et al., 2026);

guideline-conditioned systems retain the guideline at inference (Gupta et al., 2023; Lu et al., 2024). Interactive settings withhold findings until asked (Schmidgall et al., 2024; Nori et al., 2025; Tu et al., 2024; Li et al., 2024), whereas most medical evaluations provide the full vignette (Singhal et al., 2023; Lee et al., 2025; Mittal & Aggarwal, 2026). GAO supervises from a compiled guideline without per-dialogue human labels or rewards, and we separate asking from deciding.

Evaluating decisions with asymmetric errors. Cost-sensitive learning prices each error type (Elkan, 2001), and cost curves compare classifiers over cost ratios (Drummond & Holte, 2006); we apply both. A reference informed by the method under test is biased toward it (Whiting et al., 2003; 2011), so we rescore under alternative labelings and bound overwritten historical runs (Heitjan & Rubin, 1991; Horowitz & Manski, 1995; Manski, 2003). Appendix J extends this discussion.

6 Discussion

What else the supervision carries. A pre-specified ablation of the 160 urgent boundary dialogues (Section 2.2) detected no significant difference in its target, urgent recall (all pairwise $p \geq 0.22$), or in agreement (all pairwise $p \geq 0.76$), so we claim no benefit for them (Appendix C.10). By contrast, the tier-word control learned the rule-derived disposition from its next-step sentence, and permuting the entire final message removed the emergent gain (Section 4.2): a permutation control must permute every text that co-varies with the label.

Rewarding the outcome. Starting from the original SFT weights, GRPO (Shao et al., 2024) with disposition rewards plus rule, format, and LLM-judged process terms lowers agreement to 91/201 because 80 calls end without a disposition, 72 of them after a single exchange (Appendix F.2); the overwrite heuristic had masked this failure (151/201). We report a failure mode of this configuration, not an estimate of what reinforcement learning can achieve.

Implications for evaluating triage agents. Three checks changed our conclusions without new annotation: direction-aware costs distinguished operating-point shifts from failures; replication across runs and environments separated stable emergent gains from variable agreement; and full-information evaluation with two permutation controls located the labels' effect in decisions rather than questions.

7 Limitations and Conclusion

Limitations. The author-adjudicated reference informed the table and model selection and lacks independent clinician-panel validation. A clinician reviewed only the 25 flagged cases; the resulting labels retain the original tiers elsewhere and differ from the reference in 3 cases. Because this review stays within our workflow, none of the views in Section 4.4 is an independent clinical reference. Rule application also involves judgment: in a separate audit of candidate training dialogues, two rule-conditioned LLM reviewers agreed on 44.1% of 930 reviews (κ=0.285; 58.4% on the 685 where both committed to a tier), and 9 of our 12 label revisions were downgrades, so the reference may under-call emergent cases. Interactive calls use LLM callers, and real callers volunteer and withhold information differently. The cost analysis prices every under-triaged call alike, although calling an emergency routine is worse than calling it urgent; per-tier costing changes no conclusion (Appendix I). Fact positions were keyword-located, and the next-step analysis of the tier-word control was post hoc. A targeted provenance audit found four clusters of 12 to 19 generated rows paraphrasing development cases in historical corpora (Appendix D); the main corpus passed the construction-time 8-gram gate, but we retain no full-corpus semantic-neighbor audit for it.

Conclusion. A compiled specialty guideline can supply every disposition label for an interactive triage agent, and those labels teach when to escalate: across five training runs, under-triage falls by 49 to 62%. Whether agreement also rises depends on how far a run raises over-triage, so triage agents should be compared by direction-aware cost over several runs; clinical value still requires a clinician-labeled held-out reference and real calls.

AI USE STATEMENT

Large language models are components of the studied pipeline: they generated the synthetic training profiles and dialogues (Section 2.3; generator models are listed in Appendix A), simulate the caller during evaluation, act as rule-conditioned auditors that propose but never assign reference-label revisions (Section 3), and serve as the general-purpose nurse baselines of Table 1. Beyond these methodological uses, we used LLM-based assistants to write and modify the data-construction, evaluation, and analysis code, to run and monitor experiments, and to draft and edit the manuscript. The authors reviewed all code and text, verified reported numbers against the retained records, and take full responsibility for the content of this paper.

ETHICS STATEMENT

The system is a research prototype evaluated only on simulated calls; no patient or clinician interacted with it, and it must not guide patient care. Training dialogues are synthetic, but an unrecorded fraction of profile seeds and the 201-case benchmark derive from public, patient-authored Reddit posts, and public posting is not consent to machine-learning use. We retain no platform agreement, acquisition record, or de-identification audit, leaving re-identification risk unassessed; post-derived text reached remote endpoints whose retention policies we did not record, and no IRB review was sought (Appendix D). To limit further disclosure, the released artifact withholds all case text, profiles, transcripts, and source identifiers, and contains only random identifiers and categorical labels. Because the evaluation reference was adjudicated by the authors, with a clinician reviewing the flagged cases but no independent clinician panel, the main results measure agreement with a development reference, not clinical safety.

REPRODUCIBILITY STATEMENT

Section 2 and Appendix B specify the rule table, the construction strategies, the admission gates, the prompts, and the constants of the historical evaluation protocols (Table 6); Section 3 and Appendix C.8 give those of the overwrite-free protocol; Appendix F reports the optimization configurations recoverable from the retained records, and Appendix C reports complete per-arm results and confusion matrices. The supplementary material contains an anonymous artifact with the 70-row rule table, the executable matcher with unit tests, and scripts that recompute the three-view accounting, matched-control intervals, overwrite bounds, and Oxford-40 comparisons from anonymous per-case labels. It also contains the per-arm aggregates with discordant-pair counts, anonymized per-call question flags, and an anonymized per-case table of the verdict of every arm with retained per-case output under the main reference and four alternative labelings, from which Tables 1, 28, and 31, the GPT-5.5-caller rows of Table 2, Figure 3, and Appendices G, H, and I can be recomputed with standard-library scripts; it also includes the training configurations of runs 3 to 5 and of both permutation controls, the construction script of the disposition-permutation corpus, and the per-vignette Oxford-40 verdicts. It does not reproduce training or interactive generation: benchmark narratives, transcripts, and model weights are withheld for the reasons given in the Ethics statement, and immutable snapshots of the remote generator and simulator endpoints were not retained (Appendix A).

REFERENCES

American Academy of Ophthalmology. Preferred practice pattern guidelines. `https://www.aao.org/education/preferred-practice-patterns`, 2026. Accessed 2026.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, et al. Constitutional AI: Harmlessness from AI feedback. *arXiv preprint arXiv:2212.08073*, 2022.

David Bani-Harouni, Chantal Pellegrini, Ege Özsoy, Nassir Navab, and Matthias Keicher. Language agents for hypothesis-driven clinical decision making with reinforcement learning. In *International Conference on Learning Representations (ICLR)*, 2026.

Ruike Cao, Shaojie Bai, Fugen Yao, Liang Dong, Jian Xu, and Li Xiao. ATPO: Adaptive tree policy optimization for multi-turn medical dialogue. In *International Conference on Learning Representations (ICLR)*, 2026.

Xilin Dang, Kexin Chen, Xiaorui Su, Ayush Noori, Iñaki Arango, Lucas Vittor, Xinyi Long, Yuyang Du, Marinka Zitnik, and Pheng Ann Heng. KnowGuard: Knowledge-driven abstention for multi-round clinical reasoning. In *International Conference on Learning Representations (ICLR)*, 2026.

Chris Drummond and Robert C. Holte. Cost curves: An improved method for visualizing classifier performance. *Machine Learning*, 65(1):95–130, 2006.

Charles Elkan. The foundations of cost-sensitive learning. In *Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 973–978, 2001.

Prakhar Gupta, Yang Liu, Di Jin, Behnam Hedayatnia, Spandana Gella, Sijia Liu, Patrick Lange, Julia Hirschberg, and Dilek Hakkani-Tur. DialGuide: Aligning dialogue model behavior with developer guidelines. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 14031–14047, 2023.

Daniel F. Heitjan and Donald B. Rubin. Ignorability and coarse data. *The Annals of Statistics*, 19 (4):2244–2253, 1991. doi: 10.1214/aos/1176348396.

Joel L. Horowitz and Charles F. Manski. Identification and robustness with contaminated and corrupted data. *Econometrica*, 63(2):281–302, 1995. doi: 10.2307/2951627.

Zhiting Hu, Xuezhe Ma, Zhengzhong Liu, Eduard Hovy, and Eric Xing. Harnessing deep neural networks with logic rules. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2410–2420, 2016.

Tenghao Huang, Sihao Chen, Muhao Chen, Jonathan May, Longqi Yang, Mengting Wan, and Pei Zhou. Teaching language models to gather information proactively. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 15588–15599. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-emnlp.843.

Annapoorna Kini, Andriy Vengrenyuk, Derek Pineda, Yuliya Vengrenyuk, et al. Conversational artificial intelligence for pre-procedural patient preparation: Implementation, validation and patient satisfaction. *npj Digital Medicine*, 2026. doi: 10.1038/s41746-026-02959-x.

Florian Kreyssig, Iñigo Casanueva, Paweł Budzianowski, and Milica Gašić. Neural user simulation for corpus-based policy optimisation of spoken dialogue systems. In *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*, pp. 60–69. Association for Computational Linguistics, 2018. doi: 10.18653/v1/W18-5007.

Yunghwei Lai, Kaiming Liu, Ziyue Wang, Weizhi Ma, and Yang Liu. Doctor-R1: Mastering clinical inquiry with experiential agentic reinforcement learning. In *International Conference on Learning Representations (ICLR)*, 2026.

Joseph Lee, Tianqi Shang, Jae Young Baik, Duy Duong-Tran, Shu Yang, Lingyao Li, and Li Shen. From promising capability to pervasive bias: Assessing large language models for emergency department triage. arXiv preprint arXiv:2504.16273, 2025.

Shuyue Stella Li, Vidhisha Balachandran, Shangbin Feng, Jonathan S. Ilgen, Emma Pierson, Pang Wei Koh, and Yulia Tsvetkov. MediQ: Question-asking LLMs and a benchmark for reliable interactive clinical reasoning. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Shuyue Stella Li, Jimin Mun, Faeze Brahman, Pedram Hosseini, Bryceton G. Thomas, Jessica M. Sin, Bing Ren, Jonathan S. Ilgen, Yulia Tsvetkov, and Maarten Sap. ALFA: Aligning LLMs to ask good questions—a case study in clinical reasoning. In *Second Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=12u7diwku0.

Meng Lu, Brandon Ho, Dennis Ren, and Xuan Wang. TriageAgent: Towards better multi-agents collaborations for large language model-based clinical triage. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 5747–5764, 2024.

Charles F. Manski. *Partial Identification of Probability Distributions*. Springer Series in Statistics. Springer, New York, 2003. doi: 10.1007/b97478.

Surina Mittal and Yakshi Aggarwal. Evaluation of large language models in the diagnosis, urgency triage, and initial management of ophthalmic emergencies. *Cureus*, 18(1):e101433, 2026. CC-BY 4.0.

Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. Capabilities of GPT-4 on medical challenge problems. *arXiv preprint arXiv:2303.13375*, 2023.

Harsha Nori, Mayank Daswani, Christopher Kelly, Scott Lundberg, et al. Sequential diagnosis with language models. *arXiv preprint arXiv:2506.22405*, 2025.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, et al. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Qwen Team. Qwen3.5: Towards native multimodal agents. Qwen technical blog, `https://qwen.ai/blog?id=qwen3.5`, February 2026.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Alexander Ratner, Stephen H. Bach, Henry Ehrenberg, Jason Fries, Sen Wu, and Christopher Ré. Snorkel: Rapid training data creation with weak supervision. *Proceedings of the VLDB Endowment*, 11(3):269–282, 2017.

Alexander J. Ratner, Christopher De Sa, Sen Wu, Daniel Selsam, and Christopher Ré. Data programming: Creating large training sets, quickly. In *Advances in Neural Information Processing Systems*, volume 29, pp. 3567–3575, 2016.

Jost Schatzmann, Blaise Thomson, Karl Weilhammer, Hui Ye, and Steve Young. Agenda-based user simulation for bootstrapping a POMDP dialogue system. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Companion Volume, Short Papers*, pp. 149–152. Association for Computational Linguistics, 2007. URL `https://aclanthology.org/N07-2038/`.

Samuel Schmidgall, Rojin Ziaei, Carl Harris, Eduardo Reis, Jeffrey Jopling, and Michael Moor. AgentClinic: A multimodal agent benchmark to evaluate AI in simulated clinical environments. *arXiv preprint arXiv:2405.07960*, 2024.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, et al. Large language models encode clinical knowledge. *Nature*, 620:172–180, 2023.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Stanford alpaca: An instruction-following LLaMA model. `https://github.com/tatsu-lab/stanford_alpaca`, 2023.

Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. Large language models in medicine. *Nature Medicine*, 29:1930–1940, 2023.

Tao Tu, Anil Palepu, Mike Schaekermann, Khaled Saab, Jan Freyberg, et al. Towards conversational diagnostic AI. *arXiv preprint arXiv:2401.05654*, 2024.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. Self-instruct: Aligning language models with self-generated instructions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, 2023.

Zhongyu Wei, Qianlong Liu, Baolin Peng, Huaixiao Tou, Ting Chen, Xuanjing Huang, Kam-fai Wong, and Xiangying Dai. Task-oriented dialogue system for automatic diagnosis. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 201–207, 2018.

Penny Whiting, Anne W. S. Rutjes, Johannes B. Reitsma, Patrick M. M. Bossuyt, and Jos Kleijnen. The development of QUADAS: A tool for the quality assessment of studies of diagnostic accuracy included in systematic reviews. *BMC Medical Research Methodology*, 3:25, 2003.

Penny F. Whiting, Anne W. S. Rutjes, Marie E. Westwood, Susan Mallett, Jonathan J. Deeks, Johannes B. Reitsma, Mariska M. G. Leeflang, Jonathan A. C. Sterne, and Patrick M. M. Bossuyt. QUADAS-2: A revised tool for the quality assessment of diagnostic accuracy studies. *Annals of Internal Medicine*, 155(8):529–536, 2011.

Peng Xia, Jinglu Wang, Yibo Peng, Kaide Zeng, Zihan Dong, Xian Wu, Xiangru Tang, Hongtu Zhu, Yun Li, Linjun Zhang, Shujie Liu, Yan Lu, and Huaxiu Yao. MMedAgent-RL: Optimizing multi-agent collaboration for multimodal medical reasoning. In *International Conference on Learning Representations (ICLR)*, 2026.

Cheng Xu, Shuhao Guan, Derek Greene, and M-Tahar Kechadi. Benchmark data contamination of large language models: A survey. arXiv preprint arXiv:2406.04244, 2024.

Lin Xu, Qixian Zhou, Ke Gong, Xiaodan Liang, Jianheng Tang, and Liang Lin. End-to-end knowledge-routed relational dialogue system for automatic diagnosis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 7346–7353, 2019.

Ailing Yu, Lan Yao, Jingnan Liu, Zhe Chen, Jiajun Yin, Yuan Wang, Xinhao Liao, Zhiling Ye, Ji Li, Yun Yue, Hansong Xiao, Hualei Zhou, Chunxiao Guo, Peng Wei, Junwei Liu, and Jinjie Gu. Medresearcher-r1: Expert-level medical deep researcher via a knowledge-informed trajectory synthesis framework. arXiv preprint arXiv:2508.14880, 2025.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, et al. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2023.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, Zhangchi Feng, and Yongqiang Ma. Llamafactory: Unified efficient fine-tuning of 100+ language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (System Demonstrations)*, 2024.

Kun Zhou, Yutao Zhu, Zhipeng Chen, Wentong Chen, Wayne Xin Zhao, Xu Chen, Yankai Lin, Ji-Rong Wen, and Jiawei Han. Don't make your LLM an evaluation benchmark cheater. *arXiv preprint arXiv:2311.01964*, 2023.

A EXPERIMENTAL CONDITIONS AND PROVENANCE

Appendices A–I provide artifact mappings, protocol specifications, complete result tables, integrity audits, additional diagnostic analyses, reported optimization configurations, the transcript audit, the reference-robustness and triage-safety analyses, and a controlled retraining with a matched label-permutation control. E/U/R denote Emergent/Urgent/Routine; evaluation protocols are defined in Table 6. In the appendices, *clean* denotes the overwrite-free protocol of Section 3; Protocols A, A′, B, and B′ are historical runs with the overwrite heuristic enabled.

We identify each experimental condition by its data construction and optimization schedule.

Construction interfaces are described in Section 2. Table 3 records condition-level provenance; Appendix E retains the negative and null diagnostics without repeating the development chronology.

Generation provenance. The 3,000-dialogue source used for the archival base-model/SFT comparison (the boundary-revised condition in Table 3) is a heterogeneous, filtered pool rather than the output of one generator. Preserved metadata attribute 1,235 rows to MiniMax 2.5, 640 to DeepSeek-V3.2, 635 to GLM-5, 293 to Qwen3-235B, 99 to a rule-template source, 96 to a scripted source, and 2 to Gemini 2.5 Pro. The provider-identity field records the generation source and is not a quality indicator; every row is subject to downstream admission checks. Immutable endpoint snapshots and complete decoding settings were not retained for every historical source, so the mixed corpus cannot be regenerated exactly from the archived configuration records.

Figure 4 shows the primary-corpus composition.

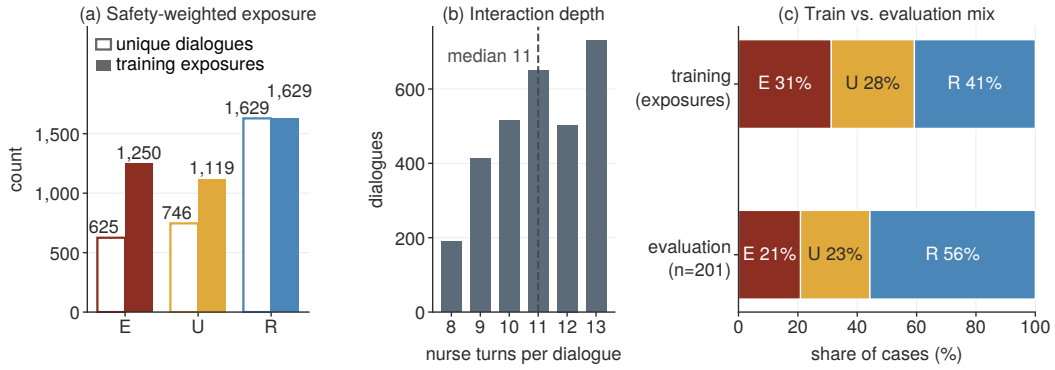


Figure 4: Primary boundary-revised corpus. **(a)** Unique dialogues and safety-weighted training exposures. **(b)** Nurse-turn depth (median 11). **(c)** Class mix of the training exposures against that of the evaluation reference, which differ by design: resampling upweights the severe tiers. E/U/R denote emergent/urgent/routine.

Table 3: Data conditions used in the development analyses.

Condition	Construction
Initial format-normalized	2,250 dialogues (500 E, 500 U, and 1,250 R), after removing role leakage, inconsistent openings, and premature verdicts.
Natural-language	3,000 dialogues generated with an explicit naturalness objective.
No-rationale	Natural-language dialogues after removing the auxiliary rationale field.
Rule-stratified	2,315 dialogues rebalanced over the 24 severity-by-symptom-family strata (rounded E:U:R percentage shares 25:25:49).
Class-ratio pilot	A 1,353-dialogue pilot with rounded E:U:R percentage shares 25:25:51.
Reference corpus	2,635 dialogues assembled under the six admission-gate families in Table 5; the retained record does not certify per-dialogue gate satisfaction. Construction removed 403 E/U dialogues, added 92 boundary-focused dialogues, and revised 221 dialogues under a common output-format standard.
Coverage-expanded	3,274 dialogues produced by the reference construction procedure with broader profile coverage; the number of E profiles increased from 603 to 755.
Safety-balanced	3,360 source dialogues resampled with E/U/R weights 2.0/1.5/1.0, yielding a 4,517-example exposure stream (rounded E:U:R percentage shares 33:27:40).
No-upsampling control	The same 3,360 source dialogues as the safety-balanced condition, without class-dependent resampling.
Gate-integrated expansion	A 1,000-dialogue corpus assembled with admission gates applied during generation. The final corpus contains 83 additional API-generated rows selected from 93 accepted candidates; no lexical overlap was detected by the retained screen.
Provider-concentrated	An unfiltered expansion in which one generator supplied 86.5% of the accepted dialogues (Appendix E: provider identity and admission filtering are confounded in this condition).
Recitative-R	A teacher-generated corpus whose R rationales frequently repeated patient wording rather than discriminating among neighboring dispositions.
Discriminative-R	The recitative-R corpus after applying a discriminative-rationale criterion: 491 dialogues were rejected and all 815 retained R dialogues passed the criterion.
Composite	A combination of safety balancing, continuity constraints, and discriminative R supervision.
Discriminative-E	The composite corpus after replacing scripted E dialogues with teacher-generated, discriminative E dialogues.
Boundary-revised	A 3,000-dialogue source corpus (625 E, 746 U, and 1,629 R) after within-stratum replacement of 160 U-boundary dialogues designed to address both missed escalation and hallucination-induced over-escalation. Class weights of 2.0/1.5/1.0 yield 3,998 exposure examples.
Safety operating point	The boundary-revised source corpus resampled with the U weight increased from 1.5 to 2.0.
Boundary-mix	A fixed-budget augmentation that increases the boundary subset from 160 to 240 dialogues by adding 48 R, 24 E, and 8 U examples.
Boundary-clean	The boundary-mix condition after revising 80 dialogues while preserving its class and stratum structure.
Moderate-U	The boundary-revised source corpus resampled with a U weight of 1.65, yielding 4,110 exposure examples.
Flexibility-revised	5,177 dialogues revised to cover counter-questions, caregiver proxies, and patient deflection.
Reason-grounded	The flexibility-revised corpus after rewriting approximately 2,451 rationale lines while leaving the dialogue turns unchanged.
Dialogue-revised	The reason-grounded corpus after 120 wait-turn splits and 774 repeated-question corrections; no exact repeated questions remained.
Completeness-audited	A family-level audit of 4,328 candidates across eight symptom families, followed by 222 removals and 219 replacements, yielding 6,680 exposure examples.

The trained precursor and dialogue-revised SFT conditions listed below are distinct model states. The former is used for the external Oxford-40 comparison, whereas the latter is used for three-view descriptive accounting.

Table 4: Principal model conditions used in the reported comparisons.

Condition	Training specification	Role in the analysis
Safety-balanced SFT	Three-epoch full-parameter SFT of the 9B backbone on the safety-balanced exposure stream; the epoch-1 model state is used.	Retrospective upsampling contrast.
Boundary-revised SFT	Three-epoch full-parameter SFT of the 9B backbone on the boundary-revised exposure stream; the epoch-3 model state is used.	Historical base–SFT reconstruction and Protocol B/B′ comparison; its same-recipe retrain is the primary clean comparison, and its original weights are the post-training initialization.
Safety-operating-point SFT	Three-epoch full-parameter SFT of the 9B backbone on the safety-operating-point exposure stream; the epoch-1 model state is used.	Alternative class-recall operating point.
Moderate-U SFT	Three-epoch full-parameter SFT of the 9B backbone on the moderate-U exposure stream; the epoch-2 model state is used.	Protocol B′ comparison.
Trained precursor	Three-epoch full-parameter SFT of the 9B backbone on the reason-grounded corpus; the epoch-1 model state is used.	Oxford-40 and family-routing analyses.
Dialogue-revised SFT	Three-epoch full-parameter SFT of the 9B backbone on the dialogue-revised corpus; the epoch-1 model state is used.	Three-view descriptive accounting.

B Prompts, Rule Table, and Protocol Constants

B.1 Construction rule table and executable subset

The 70-row rule table $\mathcal{T}$ is stored as a CSV; the complete file accompanies the source archive as `aao_triage_rules_structured.csv` (`symptom_category`, `triage_level`, `trigger_condition`). In historical construction, the prompt provided the master list and the profile generator cited one or more selected CSV rows verbatim. The fixed training label was the highest cited tier. Validators mechanically checked cited-row membership and that highest tier against the CSV, plus limited explicit consistency checks; no complete semantic interpreter executed all 70 textual triggers over arbitrary profiles.

For retrospective analysis, a mechanically checkable subset is hand-compiled into $\hat{\Phi}$: 39 executable rule entries that access 25 distinct keys in `known_fields`. The rows are author-written operationalizations, not AAO source quotations; a row-to-passage provenance map is not part of this release. Example rows exactly as they appear in the CSV:

```
vision_loss_change, Emergent, "Sudden, painless, severe loss of vision"
Pain, Emergent, "Mild ocular pain if accompanied by redness in a contact lens
wearer"
Pain, Urgent, "Mild ocular pain if accompanied by redness"
Flashes/Floaters, Emergent, "Shadow, curtain or visual field defect
accompanying flashes or floaters, regardless of duration"
```

The following is an English-glossed, line-wrapped rendering of one predicate from `guideline_match.py`. The Boolean test is unchanged; the human-readable citation is translated from Chinese:

```
(EMERGENT, "Flashes/Floaters", "curtain/shadow with flashes or floaters",
  lambda kf:  _T(kf,"flashes_floaters") and _T(kf,"curtain_shadow"))
```

During deterministic execution, `match_triage(known_fields)` returns the tier of the highest-severity rule triggered, alongside the matched rule citations, without invoking an LLM during adjudication. This executable matcher was not the historical mechanism for assigning synthetic training labels and is not an automatic compiler for the 70-row CSV.

The anonymous review artifact includes the 70-row CSV, a cleaned copy of all 39 rule entries with tests, a standalone implementation of the descriptive accounting, anonymous categorical labels for the 201-case three-view analysis and 12 matched-control runs, and byte-level manifests. It reproduces the reported three-view totals, paired transitions, rewrite counts, and three-seed resampling intervals. Restricted narratives, model weights, historical generation runners, and raw overwritten verdicts remain excluded, so it does not reproduce training or interactive generation.

B.2 NURSE SYSTEM PROMPT AND THE AUDITED TRAIN–EVALUATION MISMATCH

Each training sample carries a symptom-family-aware system prompt; the family block varies per case. Verbatim example (`lid` family):

```
You are an ophthalmology triage nurse speaking by phone.  Ask one focused,
natural question at a time and adapt to the caller's main complaint; do not
run the same fixed checklist on every patient.
Symptom focus for this training case:  eyelid problems (droop, swelling,
bump).
Decision cues:  new lid droop with pupil/vision/neurologic symptoms ⇒
Emergent; lid droop/swelling/cannot open or blocked vision without
pupil/neurologic symptoms ⇒ Urgent; small stable stye without red flags ⇒
Routine.
Final output must be exactly:  a first line that starts with "Triage Result:
<Emergent|Urgent|Routine>"; immediately a "Reason:"  line citing only 2-4
facts actually stated in the transcript (no invented or unasked facts); then
a final "[Hang Up]" line.  Do not output hidden reasoning, markdown, or
<think> blocks.
```

The available evaluation code and configuration history indicate that the reported evaluations used a generic, non-family-specific block. Under Protocol A, the fixed nonspecific opening does not match the family regular expressions. Under Protocols B and B′, the runner constructs and freezes the nurse prompt before the caller's first reply, while family inference occurs only afterward. The same records indicate generic routing for the diagnostic and Oxford scripts. Per-case artifacts do not retain the exact prompt text, and the exact historical runner revision is unavailable; the route is therefore inferred from the available code rather than independently verified for every retained dialogue.

We audited the mismatch on the 201 cases using the trained precursor, not the distinct dialogue-revised SFT condition (Table 4). The audit does not invoke the historical post-interaction heuristic and is not a direct replay of Protocol B. Three post hoc arms use (A) the generic block, (B) hidden-profile family inference, or (C) visible-complaint family prediction. With 42–46 cases per arm triggering simulator fallback or a quality violation, we report full-sample descriptive metrics in prose rather than a table.

Full-sample agreement/balanced accuracy/E_{rec}/U_{rec} are 0.756/0.749/0.714/0.766 for the generic replay, 0.677/0.625/0.524/0.574 for hidden-profile routing, and 0.701/0.671/0.714/0.532 for predicted-family routing. The available code contains no primary route that used hidden-profile family routing.

Among cases without fallback or quality violations in either arm, hidden-profile minus generic agreement is -0.071 (95% CI $[-0.143, 0.000]$, 10 hidden-profile-only versus 21 generic-only, McNemar $p = 0.071$, 154 pairs), and predicted minus generic is -0.052 (CI $[-0.124, 0.020]$, 12 predicted-only versus 20 generic-only, $p = 0.215$, 153 pairs). Neither comparison supports an agreement difference. The predicted-family classifier matches the hidden-profile-derived family on 53.2% of full-sample cases; that derived family is a routing reference, not a clinically validated label. These results document the reconstructed evaluation route and do not justify routing on hidden patient information.

B.3 TRAINING-ROW ADMISSION GATES

Figure 6 shows how a single dialogue moves through the construction pipeline. The pipeline groups its admission checks into the following final intended taxonomy: the label validator of Section 2.3 plus the seven gates named in Figure 1, grouped into six families. Gate families and thresholds evolved across the adaptive data versions; historical corpora did not all implement every entry.

Cited row (family *redness/discharge*): "Redness or discharge in a contact lens wearer" ⇒ Emergent; label $\Phi(J) =$ Emergent.

Profile (written by DeepSeek): microbial keratitis after extended contact-lens wear; hidden facts include a red right eye with sticky discharge, sharp worsening pain (7/10), light sensitivity, blurred vision, contact-lens wear, onset 36 hours ago; behavior: irritable, volunteers little. Opening line: "Look, my eye's killing me and there's like a white dot on it, I need this sorted fast."

Dialogue (realized by MiniMax-2.5; 3 of 11 nurse questions shown):
NURSE: Is your eye red, and is there any discharge or tearing?
CALLER: Yeah, it's red as hell and there's sticky white-yellow gunk that makes my eyelids stick together. [...]
NURSE: Sure. Do you wear contact lenses?
CALLER: Yeah, I've been wearing monthly contacts for five years. [...] I've been sleeping in them. [...]
NURSE: Have you had any eye surgery or an eye procedure within the past six weeks? CALLER: No, nothing like that. [...]
NURSE: Triage Result: Emergent. Reason: the caller reported contact lens use with a red eye and other warning symptoms. Next step: Please seek emergency eye care now [...]

Gates passed: final tier equals $\Phi(J)$; redness and contact-lens use are stated by the caller before the verdict; the reason cites only caller-stated facts; all five safety items are covered.

Figure 5: One training example traced from a table row to a labeled dialogue. All text is verbatim from the training corpus, abridged at [...].

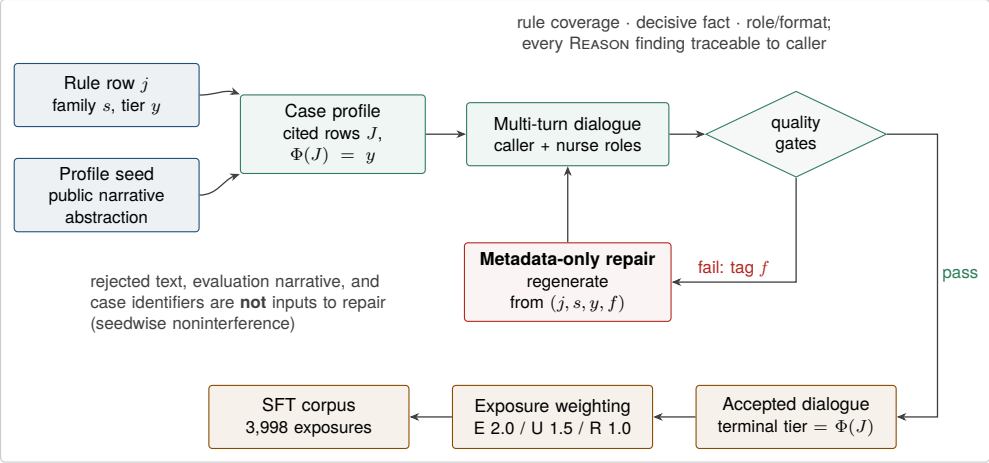


Figure 6: Life cycle of one training dialogue under the construction contract. A cited rule row and a profile seed determine the case profile and its immutable label $\Phi(J)$; the generated multi-turn dialogue is admitted only if it passes the gate families in Table 5. A failed dialogue is not edited: it is regenerated through the metadata-only repair interface, whose signature receives only the rule identifier, symptom family, target tier, and failure-mode tag. Accepted dialogues enter the corpus with deterministic class exposure weighting. The intended information restriction shown here is a contract, not a verified property of every historical batch; audited departures are reported in Section D.

B.4 PATIENT-SIMULATOR CONSTRAINTS

The simulator receives the complete profile but is instructed to answer only the question asked, disclose hidden facts progressively, avoid volunteering the terminal template, and emit no role or hidden-reasoning tokens. Protocol A is an exception to the usual profile-derived opening: its first caller turn is the fixed sentence "Hi, I am calling about a problem with my eye." Protocols A′, B and B′ instead use a neutral nurse-first opening. Protocol B′ further limits each patient reply to 45 words and permits up to three generation attempts under a validator whose exact rubric was not retained. The two historical Protocol B executions produced different first caller replies on all 201 paired cases despite temperature 0; this comparison does not include Protocol B′ and cannot distinguish version from realization differences.

Table 5: Final intended admission-gate taxonomy for synthetic training dialogues.

Family (gates)	Acceptance criterion
Rule and label (label validator)	Final tier parses, matches the immutable profile label, and is supported by the profile's highest-severity triggered rule.
Dialogue structure (format)	Roles alternate; the final tier and hang-up occur once, at the final nurse turn; no prompt, tool, role, or thinking tokens appear.
Profile consistency (consistency)	Caller answers do not contradict explicit profile facts or invent high-risk modifiers; hidden facts are disclosed only when asked.
Grounding (decisive fact; grounding)	Decisive facts are present in caller turns; every clinical finding cited in `Reason` is transcript-locatable, with no unasked negative or contradiction.
Coverage and diversity (coverage; diversity)	Rule/family/tier coverage is checked; length bounds, exact deduplication, and repeat caps limit template dominance.
Contamination (leakage; revised prospective procedure)	Normalized exact and 8-gram matches are rejected; candidate semantic-neighbor hits are audited and regenerated from rule identifiers rather than benchmark text.

B.5 Evaluation-protocol specification

Across A, A′, B, and B′, the common parser selects the last regex-matched *Triage Result* or *Triage Level* and applies a fixed synonym map. A parseable disposition terminates the dialogue. A malformed hang-up permits one retry, and a terminal verdict is required during the final two permitted turns. These constants describe the recorded historical evaluation procedure; in the absence of immutable remote-model snapshots, they do not make the procedure exactly reproducible.

The matched Protocol-A reruns were not interleaved. Archived timestamps place the seed-42 reference-side executions on June 12 or 14 and the corresponding comparison-side executions on July 11, a 27–29 day gap. Seed-43 and seed-44 arm pairs were run approximately 16–17 hours apart on July 12–13. All used the same recorded provider model identifier but no immutable snapshot. These timing differences remain a procedure-level confound in the descriptive three-seed contrasts.

B.6 Historical post-interaction heuristic audit

Several historical base–SFT comparisons carry this defect. In one archived protocol (Protocol B) we detect 31 base and 18 SFT overwrites, and an exact-template audit spanning 5,427 case-runs finds 19 to 44 overwritten cases per run. The same runs share the simulator's identifier and temperature but sometimes ran days apart without an immutable provider snapshot. Because the heuristic fired only on disagreement, an overwritten case that now agrees must previously have been wrong, whereas an overwritten disagreement may have been right. Writing C for stored agreement and M_+, M_- for the overwritten agreements and disagreements, direct agreement lies in $[C - M_+,\ C - M_+ + M_-]$, assuming that overwriting was the only record change, that detection is exact, and that cases factorize (Appendix B.7). Training used symptom-family-specific prompts, while the retained evaluation prompts of these runs were generic and unversioned.

Available evaluation code and retained outputs indicate that Protocols A, A′, B, and B′ enabled a benchmark-adapted deterministic heuristic after the dialogue terminated. This post-interaction mechanism is not part of the learned policy or $\hat{\Phi}$. It operates as follows: (i) parse the last model-generated disposition z_0; (ii) concatenate the generic opening and hidden complete profile, then scan an ordered list of regular expression branches; (iii) assign the tier z_H from the first matching branch; and (iv), whenever $z_H \neq z_0$, overwrite both the final nurse turn and the stored final-verdict field with z_H. The trigger list combines general symptom patterns with strings adapted to particular

Table 6: Protocol constants recoverable from evaluation code and retained artifacts. "Unavailable" denotes an unretained constant, rubric, immutable provider snapshot, or historical runner revision.

Protocol	Patient simulator	Opening and nurse configuration	Output handling	Primary use and limitation
A	Recorded provider model identifier: Qwen3-235B; temperature 0.9; top-p 0.9 and 256-token cap from the retained code defaults; immutable snapshot unavailable.	Fixed caller-first opening; generic prompt route indicated by the available code; greedy nurse decoding.	Common parser and benchmark-adapted post-interaction heuristic; at most 20 turns.	Adaptive trace and matched data controls evaluated June 12 or 14 through July 13, 2026 (earliest archived timestamp ambiguous); exact evaluation-code revision unavailable.
A′	Recorded provider model identifier: Qwen3-235B; temperature 0.3; top-p 0.9 and 256-token cap from the retained code defaults; immutable snapshot unavailable.	Neutral nurse-first opening; generic route indicated by the available code; greedy nurse decoding.	Common parser and benchmark-adapted post-interaction heuristic; at most 20 turns.	Model selection and sensitivity analyses; exact evaluation-code revision unavailable.
B	Recorded provider model identifier: GPT-5.5; temperature 0; top-p and token cap were not retained in the producer metadata; immutable snapshot unavailable.	Neutral nurse-first opening; generic route indicated by the available code; greedy nurse decoding.	Common parser and benchmark-adapted post-interaction heuristic; at most 20 turns.	Case-paired base-model/SFT comparison. SFT was evaluated June 22 and the base model July 10, 2026; first caller replies match exactly in 0/201 cases.
B′	Recorded provider model identifier: GPT-5.5; temperature 0; top-p 1; 128-token cap; immutable snapshot unavailable.	Neutral nurse-first opening; generic route indicated by the available code; patient replies limited to 45 words and passed through a validator with at most three attempts; exact rubric unavailable.	Common parser and benchmark-adapted post-interaction heuristic; at most 20 turns.	Naturally phrased grounding and ablation studies; exact evaluation-code revision unavailable.
L	Simulator identity, sampling parameters, opening, prompt route, and stopping rules vary across experiments or are unavailable.	Heterogeneous historical configurations.	Heterogeneous parsing and termination behavior.	Historical diagnostics only; not used for direct numerical comparison with A, A′, B, or B′.

development cases. Companion source comments explicitly distinguish generalizable rules from development-case-specific strings. The mechanism therefore violates the interactive information constraint and must not be interpreted as model performance. Because retained outputs preserve the replacement but not z_0, exact pre-heuristic policy-reference agreement cannot be recovered.

Exact-template output audits flag replacement-consistent matches for Protocols A and B. For the Protocol A matched controls, output-pattern analysis flags 27/27/29 matches across the three safety-balanced runs, 32/27/28 in the no-upsampling controls, 28/28/31 in the reference-corpus controls, 27/31/31 in the size-matched content controls, and 32/28/27 in the full coverage-expanded runs.

Table 7: Selected archived runs with identifiable exact-template flags ($n = 201$ each). C is stored policy-reference agreement; M_+/M_- partition detected rewrites, and Bounds are conditional sharp direct-agreement counts. Boundary-mix and boundary-clean have no designated operating checkpoint, so we retain their best recorded agreement and mark it as retrospective. Exact runner revisions and per-case transcripts are unavailable; the counts can be recombined but not independently regenerated.

Protocol	Displayed condition	Epoch	C	M_+/M_-	Bounds
A	Rule-stratified	3	123	23/18	100–118
	Class-ratio pilot	1	144	20/13	124–137
	Reference corpus	3	145	15/13	130–143
	Coverage-expanded	3	147	18/14	129–143
	Safety-balanced	1	150	16/11	134–145
A′	Frontier nurse, no in-prompt table	—	123	23/10	100–110
	Frontier nurse, 70-row in-prompt table	—	144	14/11	130–141
B′	Boundary-revised	3	148	13/12	135–147
	Boundary-mix (best recorded)	1	147	12/7	135–142
	Boundary-clean (best recorded)	3	144	14/11	130–141
	Moderate-U	2	151	15/12	136–148

Each count is out of 201 cases. Protocol A metrics therefore describe the complete evaluation stack, not raw model verdicts.

We additionally screened 29 paper-table rows across Protocols A, A′, and B′. Twenty-seven admit the same exact-template audit; two reproduce matched-control seed-42 rows, leaving 25 newly released runs and 5,025 case-runs. Table 7 retains one selected or explicitly best-recorded checkpoint per condition rather than the full trajectories. The initial format-normalized epoch-3 row lacks compatible calibration metadata and terminal formatting. The no-rationale epoch-3 contract omits the rationale, so its two-line verdict is indistinguishable from unmodified model output. Neither row receives a rewrite count or direct-agreement bound.

For Protocol B, output-pattern analysis detects 31 affected base-model dialogues. Among the 119 post-heuristic agreements, 21 are detected rewrites, leaving 98 unaffected agreements. Those 21 rewrites must have disagreed before the tier changed; only the 10 detected rewrites that remain in disagreement could previously have agreed. The pre-heuristic count is therefore 98–108/201. For SFT, 10 of 151 post-heuristic agreements are detected rewrites, leaving 141 unaffected agreements; at most the eight rewrites that remain in disagreement could previously have agreed, giving 141–149/201. Conditional on template matching detecting every and only a post-interaction heuristic overwrite, the bounds are 0.488–0.537 and 0.701–0.741. Coincidental emission of an identical template by the model would be a false positive and violate this assumption. As a direct check, we applied the same 31-template exact-match detector to 2,010 archived case-runs from ten later, unrelated interactive evaluations—including the guards-off arm of the interaction-constraint sensitivity analysis (Section C.13)—drawn from a harness generation whose output records carry no legacy-calibrator tag: zero final nurse turns match any template. This bounds the detector's false-positive rate at 0/2,010 on runs structurally unlikely to invoke the calibrator, but is not a guarantee for the two historical runs under audit, whose model states, decoding, and evaluation code revision differ. Undetected overwrites act in the opposite direction: closing the 33-case separation between the two configurations would require at least 33 overwrites to escape detection across the two runs (for example, 22% of the 151 stored SFT agreements being undetected rewrites). The recorded point estimates characterize outputs after the heuristic, whereas the exact pre-heuristic confusion matrices and paired test are unrecoverable.

B.7 General bound and proof

The general bound behind Section B.6 states that, for a finite label set $\mathcal{Y}$, fixed references y_i, a case-additive score $\Psi_\ell = n^{-1}\sum_i \ell(d_i, y_i)$ with $\ell : \mathcal{Y}^2 \to \mathbb{R}$, and nonempty per-case fibers Γ_i whose joint compatibility set is their product, the sharp bounds are $\underline{\Psi}_\ell = n^{-1}\sum_i \min_{d\in\Gamma_i} \ell(d, y_i)$ and $\overline{\Psi}_\ell = n^{-1}\sum_i \max_{d\in\Gamma_i} \ell(d, y_i)$, with point identification exactly when $\ell(\cdot, y_i)$ is constant on every

fiber. Every compatible execution has $d_i \in \Gamma_i$, so each score contribution lies between its fiberwise minimum and maximum. Summing gives the stated bounds. Because the joint compatibility set factorizes across cases, choosing a minimizing or maximizing compatible verdict independently for each case attains both endpoints. If a fiber contains verdicts with different scores, two completions differing only on that case give different aggregate scores; the score is not point identified. Conversely, if the score is constant on every fiber, its aggregate is determined by the retained records.

B.8 Exact changed-output identification set

Let $|\mathcal{Y}| = k \geq 2$ and $N^r_{as} = \sum_i \mathbf{1}[y_i = a, z_i = s, o_i = r]$. Assume overwriting is the only record-changing mechanism, $o_i = 0$ if and only if $d_i = z_i$, $o_i = 1$ if and only if $d_i \neq z_i$, every label in $\mathcal{Y} \setminus \{z_i\}$ is feasible for an overwritten record, and compatibility factorizes across cases. For each (a, s), allocate the N^1_{as} overwritten records through nonnegative integers x_{asb} satisfying

$$\sum_{b \neq s} x_{asb} = N^1_{as}, \qquad Q_{ab} = N^0_{ab} + \sum_{s \neq b} x_{asb}, \tag{5}$$

where $Q_{ab} = \sum_i \mathbf{1}[y_i = a, d_i = b]$. Every compatible completion induces such an allocation. Conversely, any allocation can be realized by assigning the indicated number of overwritten cases in each observed cell to each alternative label. Equation (5) is therefore the exact integer identified set, not merely a collection of cellwise intervals.

For a linear confusion functional $L_w = n^{-1} \sum_{a,b} w_{ab} Q_{ab}$ with weights w_{ab}, independent allocation to a minimum- or maximum-weight alternative within each (a, s) cell yields the sharp endpoint

$$B_{\text{ext}} = \frac{1}{n}\Big\{ \sum_{a,s} w_{as} N^0_{as} + \sum_{a,s} N^1_{as} \operatorname*{ext}_{b \neq s} w_{ab} \Big\}, \qquad \text{ext} \in \{\min, \max\}. \tag{6}$$

The functional is point identified exactly when its weight is constant over feasible alternatives for every nonempty overwritten cell. With $k = 2$, each complement is a singleton, so the complete direct confusion matrix is identified. With $k \geq 3$, exact agreement (the special case $w_{ab} = \mathbf{1}[a = b]$) has the sharp interval $[C - M_+, C - M_+ + M_-]$: an overwritten stored agreement must have been wrong, whereas an overwritten stored disagreement may have been correct or may have taken another nonreference label. Independent allocations attain every integer in this interval, matching the exact-agreement bound reported in Section B.6.

The separate three-view accounting analysis uses the model-generated disposition directly, without this historical post-interaction heuristic or an additional escalation rule. Thus, 147/201 is direct model-disposition policy-reference agreement within an environment-constrained interaction.

B.9 Judge constraints

Process/quality judges are rubric-based, run with thinking disabled, and are *blind to the reference disposition*. The grounding penalty is not a judge: it is a deterministic transcript check (assertion must be locatable in patient turns; negation only if asked).

C Additional Results and Sensitivity Analyses

Unless explicitly stated otherwise, measurements under Protocols A, A′, B, and B′ use a historical evaluation procedure that applied a benchmark-adapted heuristic to the complete case profile after interaction. They characterize the recorded procedure and must not be interpreted as pre-heuristic model policy-reference agreement. Throughout this section, archival table headings such as "acc" and narrative uses of "correct" mean exact agreement with the operational reference, not clinical accuracy. The three-view accounting analysis instead evaluates the model-generated disposition directly, without the post-interaction heuristic or an additional escalation rule, although deterministic environment constraints remain active.

C.1 Protocol A: Adaptive Data Development

Table 9 retains one selected checkpoint per condition. Corpus construction and checkpoint provenance are recorded separately in Tables 3 and 7.

C.2 Retrospective Robustness Reruns for Data Contrasts

Exposure reweighting: no-upsampling control. We retrained the 3,360-row safety-balanced source with deterministic class upsampling disabled (3,360 rather than 4,517 exposures). Table 8 reports the fixed epoch-1 comparison across all retained seeds. At seed 42, no-upsampling versus safety-balanced agreement is 0.662 versus 0.746 and the Emergent/Urgent→Routine rate is 0.070 versus 0.090.

We reran both arms with seeds 42/43/44. At the fixed epoch-1 comparison, each upsampled seed exceeds its paired no-upsampling seed in κ, but the reference arm itself varies across seeds. Our primary uncertainty analysis therefore independently resamples one reference-arm seed and one control-arm seed, then jointly resamples all 201 paired cases (B=4,000). The resulting seed-pair-and-case resampling estimates are $\Delta\kappa = 0.057$ (95% RI [-0.052, 0.166]) and $\Delta E_{\text{rec}} = 0.103$ (RI [-0.146, 0.333]), upsampling minus no upsampling. These are descriptive variability intervals, not confidence intervals for a population mean over training seeds. Both include zero. The comparison holds source dialogues fixed but changes both class exposure and epoch length (approximately 105 vs. 142 optimizer updates); it therefore estimates deterministic upsampling as implemented rather than class weighting at a fixed update budget.

Coverage-expanded content mix: scale-matched control. The coverage-expanded corpus is a sample-ID superset of the reference corpus: all 2,635 IDs remain and 639 are added across all three tiers, families, providers, and styles; among the shared IDs, 105 dialogues were also revised. The stored data contain no auditable row-level `boundary` marker. We therefore describe the intervention as a coverage-expanded content mix, not as identified boundary content. We downsampled the expanded corpus to 2,635 rows using deterministic within-class sampling and retrained with the same recipe. The resulting control retains 2,121 reference IDs, substitutes 514 new IDs, and revises 82 dialogues among the retained IDs. Its E/U/R counts are 608/603/1,424 versus 603/604/1,428 in the reference, a near-ratio match rather than an exact content match.

Both the reference and control use seeds 42/43/44. The primary comparison holds the epoch fixed at 3, independently resamples one seed per arm, and jointly resamples all 201 paired cases. Expanded minus reference is $\Delta E_{\text{rec}} = 0.151$ (95% RI [-0.074, 0.372]) and $\Delta\kappa = 0.011$ (RI [-0.112, 0.142]). At source precision, E_{rec} increases in all three seed pairs (0.476→0.643, 0.548→0.762, 0.476→0.548); at table precision, U_{rec} decreases in two pairs (0.45→0.38, 0.53→0.55, 0.60→0.45). The resampling interval nevertheless includes zero.

Comparing the full-scale coverage-expanded corpus with the reference corpus at epoch 3 changes both content and size (2,635→3,274). This joint intervention gives $\Delta E_{\text{rec}} = 0.206$ (95% RI [0.022, 0.393]) and $\Delta\kappa = 0.023$ (RI [-0.083, 0.132]). It identifies a joint content-plus-scale recall contrast but cannot assign it to content, scale, or their interaction.

Under exact rewrite detection and factorized compatibility, Panel (b) of Table 8 implies sharp finite-record difference intervals of $[-8, 68]/603$ case-runs for safety-balanced minus no-upsampling and $[-33, 51]/603$ for scale-matched expansion minus reference. Both include zero. These intervals apply the same changed-output audit independently of the seed-and-case resampling analysis; they do not reproduce unrecorded direct verdicts or identify intervention effects.

C.3 Additional SFT analyses

All historical SFT conditions fine-tuned every parameter of the 9B base with DeepSpeed ZeRO-3, learning rate 5×10^{-6} with cosine decay, effective batch size 32, three epochs, and one seed per data condition. They changed several factors at a time and were scored with the overwrite heuristic, so Table 26 summarizes them as configuration-level observations. Two of them bear on the main paper. First, benchmark choice mattered: the configuration trained on the first format-normalized SFT corpus gained 21.8 points of recorded agreement on its initial benchmark but only 6 points on a

Table 8: Selected Protocol-A matched controls. Panel (a) reports $\kappa/E_{\text{rec}}/U_{\text{rec}}$ for seeds 42/43/44 at the first checkpoint for the safety-balanced pair and the third checkpoint for the expansion pair; all values include the historical post-interaction heuristic. Panel (b) pools rewrite flags over 603 case-runs per arm; bounds are conditional direct-agreement counts, not population intervals.

(a) Fixed comparison checkpoints

Condition	seed 42	seed 43	seed 44
Safety-balanced	0.574/0.81/0.51	0.526/0.69/0.40	0.531/0.74/0.47
No upsampling	0.455/0.76/0.43	0.498/0.55/0.64	0.507/0.62/0.53
Reference corpus	0.494/0.48/0.45	0.509/0.55/0.53	0.488/0.48/0.60
Scale-matched expansion	0.478/0.64/0.38	0.566/0.76/0.55	0.478/0.55/0.45
Coverage-expanded, full	0.537/0.76/0.38	0.507/0.64/0.43	0.514/0.71/0.40

(b) Pooled changed-output audit

Arm	C	M_+	M_-	Direct agreement
Safety-balanced	441	45	38	396–434
No upsampling	415	49	38	366–404
Reference corpus	428	48	39	380–419
Scale-matched expansion	430	44	45	386–431

Table 9: Selected Protocol-A checkpoints (201-case development benchmark, n=201). Mix is the rounded E:U:R percentage share; safety-balanced uses the exposure-weighted 4,517-row stream while Rows gives its 3,360-row source. Values are historical post-interaction outputs, not pre-heuristic model metrics.

Condition	Rows	Mix	acc	κ	E_{rec}	U_{rec}	E→R
Initial format-normalized	2,250	—	0.687	0.497	0.55	0.70	4
No-rationale	3,000	—	0.552	0.350	0.83	0.49	1
Rule-stratified	2,315	25:25:49	0.612	0.420	0.50	0.83	1
Class-ratio pilot	1,353	25:25:51	0.716	0.504	0.40	0.55	6
Reference corpus	2,635	23:23:54	0.721	0.494	0.48	0.45	11
Coverage-expanded	3,274	23:23:54	0.731	0.537	0.76	0.38	4
Safety-balanced ⋆	3,360	33:27:40	0.746	0.574	0.81	0.51	3

Checkpoint identities are recorded in Tables 7 and 4; the initial format-normalized and no-rationale rows use their retained third states.
⋆ Distinct from the safety-operating-point condition.

distribution-shifted one, with U_{rec} down 17.7 points and questioning length nearly doubled, a fixed-checklist behavior that the initial benchmark did not reveal. Both sets were reused during development, use superseded labels, and are distinct from the current 201-case development benchmark. Second, the coverage-expanded-to-safety-balanced transition revised source content (834 of 3,091 overlapping dialogues changed) and introduced class upsampling at the same time, so its single-seed shift (balanced accuracy 0.670→0.714, E_{rec} 0.76→0.81, U_{rec} 0.38→0.51) is descriptive only; the no-upsampling control of Appendix C.2 holds the 3,360 source dialogues fixed.

C.4 Protocol A: Comparative SFT Results

These adaptive development observations are recorded after the historical heuristic and are not controlled contrasts.

C.5 Protocol B′: Boundary and Class-Weight Ablations

Table 10 reports the boundary and class-weight conditions run under the reconstructed Protocol-B′ routing. The primary boundary result is instead the prospective ablation in Section C.10; these historical arms predate it and carry the heuristic.

Table 10: Selected Protocol-B′ boundary and class-weight conditions (generic non-family-specific nurse prompt, reconstructed routing; n=201; reference E/U/R totals 42/47/112). Selected-state provenance is in Table 7. All entries include the historical post-interaction heuristic. E/U/R are classwise correct counts; bal is macro recall; und/over count all tier-lowering/raising disagreements; pRoutine counts predicted Routine outputs; turns is the mean.

Condition	exact	acc	bal	κ	E	U	R	und	over	pRoutine	turns
Moderate-U	151/201	75.1%	0.709	0.586	30	27	94	23	27	106	22.1
Boundary-revised	148/201	73.6%	0.691	0.559	34	20	94	19	34	109	21.8
Boundary-mix (best recorded)	147/201	73.1%	0.695	0.555	37	18	92	19	35	106	21.1
Boundary-clean (best recorded)	144/201	71.6%	0.679	0.532	24	32	88	27	30	102	22.0

Table 11: GPT-5.5 as nurse with and without the in-prompt rule table under Protocol A′ (n=201). Values follow the post-interaction heuristic; the 9B reference rows are aggregate-only (see note).

Nurse configuration	acc	κ	E_{rec}	U_{rec}	under	over
GPT-5.5, no rules	0.612	0.404	0.619	0.702	0.050	0.264
GPT-5.5 + 70-row table	0.716	0.552	0.762	0.660	0.030	0.204
9B safety-operating-point	—	0.574	0.74	0.60	—	—
9B boundary-revised SFT	—	0.578	0.81	0.40	—	—

Under is Emergent/Urgent → Routine; over is Routine → Emergent/Urgent. Both exclude Emergent↔Urgent, unlike the broad counts in Table 10. For the two 9B rows, exact match and these rates were not retained; their κ and recalls are aggregate-only same-protocol (A′) SFT-selection values, not recomputable from per-case outputs. The safety-operating-point condition is distinct from the Protocol-A safety-balanced condition in Table 9.

C.6 Rule-table prompting baselines

Table 11 reports both arms under the same interactive protocol with the Qwen3-235B patient at temperature 0.3 (the Protocol A′ examiner family; comparisons are restricted to this table and its reference configurations). The historical evaluation procedure with GPT-5.5 as a zero-shot *nurse* yielded $\kappa = 0.404$ without the rule table and $\kappa = 0.552$ when the complete 70-row table was included in the system prompt. The same-protocol stack using the fine-tuned 9B safety-operating-point yielded $\kappa = 0.574$ without an in-prompt rule table. U_{rec} falls here. Table 15 shows the opposite agreement direction under a different patient, protocol, temperature, and session; neither contrast identifies a portable rule-table or training effect.

C.7 Historical Base–SFT Reconstruction

For the paired base–SFT comparison in Table 12, 101 cases agree with the reference under both configurations, 18 under base only, 50 under SFT only, and 32 under neither. Thus the difference in recorded agreement is 0.159 (paired case-bootstrap 95% percentile interval [0.080, 0.239], $B = 4{,}000$), with exact two-sided McNemar $p = 0.00013$. The two evaluations share case identities and nominal recorded settings but were executed 18 days apart; their first caller replies differ on all 201 cases, and neither an immutable provider snapshot nor the exact historical runner revision is available. The paired statistics therefore describe the recorded evaluation-procedure outputs; they neither isolate the effect of SFT from uncontrolled simulator variation nor constitute an independent confirmatory test.

C.8 Clean Same-Day Reevaluation of a Same-Recipe Retrain

Because the original boundary-revised SFT weights were not retained, we retrained the condition from the archived configuration file and its SHA-256–verified 3,000-dialogue corpus (identical base model, dataset registration, class weights, optimizer settings, sequence length, and three epochs; one training seed), and then evaluated the retrained model and the pretrained base configuration on the same day under the historical interactive protocol with the post-interaction heuristic disabled

Table 12: Historical Base–SFT reconstruction under nominal Protocol B settings with the generic non-family-specific nurse prompt (n=201). Entries are metrics computed from recorded outputs after the post-interaction heuristic and do not estimate pre-heuristic model agreement. Detected rewrites differ: base 31 ($M_+ = 21$, $M_- = 10$) versus SFT 18 ($M_+ = 10$, $M_- = 8$), with conditional direct-agreement bounds 98–108 and 141–149, respectively. Under and over are total tier-lowering and tier-raising disagreement rates over all 201 cases.

Model type	Configuration	acc	bal. acc	κ	E_{rec}	U_{rec}	R_{rec}	under	over
Base model	Qwen3.5-9B-Base	0.592	0.548	0.343	0.357	0.617	0.670	0.199	0.209
SFT model	Boundary-revised	0.751	0.698	0.576	0.857	0.362	0.875	0.109	0.139

Table 13: Clean same-day reevaluation (n=201 per arm; overwrite heuristic disabled). The SFT arm is a same-recipe retrain, not the original weights. Wilson 95% intervals for exact agreement; McNemar tests are exact and two-sided, with class rows exploratory. Severe under-triage counts are descriptive; the one SFT dialogue that ends without a parseable disposition is counted as incorrect, and the base arm emits a disposition on all 201.

Quantity	Retrained SFT	Base	Discordant[a]	McNemar p
Exact agreement	149/201 (74.1%)	124/201 (61.7%)	49/24	0.0046
Wilson 95% CI	67.7–79.7%	54.8–68.1%		
Emergent recall	29/42	4/42	26/1	4.2×10^{-7}
Urgent recall	29/47	32/47	6/9	0.61
Routine recall	91/112	88/112	17/14	0.72
Severe under-triage[b]	17	24		

[a] SFT-only / base-only discordant pairs.
[b] Emergent→Routine plus Urgent→Routine.

(`calibration=none`; every archived case record carries this tag). Both arms used the recorded Protocol B simulator settings (GPT-5.5 identifier, temperature 0, nurse-first opening, generic prompt route, greedy nurse decoding, at most 20 turns) within a several-hour window, removing the 18-day gap and the overwrite mechanism simultaneously; the provider endpoint remains unversioned.

Table 13 reports the paired outcome. The retrained configuration agrees with the operational reference on 149/201 cases versus 124/201 for the base configuration, a 25-case (12.4-point) difference with exact two-sided McNemar $p = 0.0046$; an asymptotic paired risk-difference 95% interval is 4.3–20.6 points. The entire net 25-case gap is attributable arithmetically to Emergent-reference recall (26 SFT-only versus 1 base-only correct; exploratory class-level $p = 4.2 \times 10^{-7}$, which remains significant after Holm or Bonferroni correction across the three class-level tests); the net Urgent and Routine differences cancel, though 23 SFT-only and 23 base-only discordances lie outside Emergent, and urgent recall is nominally higher for the base arm. The severe under-triage counts (17 versus 24) are descriptive, with no paired test attached. The base configuration predicts Emergent for only 6/201 cases overall, parking most emergent-reference cases in Urgent (27/42) or Routine (11/42); its aggregate agreement therefore rests on Routine/Urgent outputs while missing 38/42 emergent-reference cases. The base arm commits on all 201 dialogues; the single missing disposition in this comparison belongs to the SFT arm and is counted as incorrect.

A second training run. Because the primary contrast rests on one retrain, we trained the same configuration a second time and scored it under the same protocol. It agrees on 147/201 (Emergent 33/42, Urgent 29/47, Routine 85/112, no missing dispositions), which is indistinguishable from the first run (22 versus 20 discordant, exact McNemar p=0.88) and retains the advantage over the base model (51 versus 28 discordant, p=0.013). The two runs share the default random seed, so they measure run-to-run nondeterminism rather than seed variation; they show that the base–SFT difference is not an artifact of a single training run.

Paired Case Transcript. The paired clean-protocol transcript and its evidence-acquisition interpretation appear as Figure 8. One case illustrates the mechanism but is not itself evidence for the recall difference.

A second patient simulator. The primary protocol uses a GPT-5.5 caller. We replicated the base–SFT pair with the caller driven by DeepSeek-V4-Pro at temperature 0, holding the checkpoints, reference, prompts, turn limit, and `calibration_mode=none` fixed. Before either arm was scored we wrote a plan naming the paired exact McNemar on the 201 cases as this replication's designated endpoint, together with four interpretation rules covering the same-direction, non-significant, reversed, and instrument-failure cases. The authors retain the plan; it carries no external timestamp, so we describe the analysis as pre-specified, not registered. The SFT arm agrees on 145/201 (Emergent 30/42, Urgent 21/47, Routine 94/112) against 124/201 for the base arm (Emergent 3/42, Urgent 29/47, Routine 92/112), 49 versus 28 discordant, exact McNemar p=0.022. Neither arm produced a missing disposition, and mean interaction lengths are comparable (21.1 versus 22.4 messages), so the instrument behaved like the primary one rather than degrading. The direction, the concentration in emergent recall, and the significance all carry over. Two caveats bound the replication. The reference is unchanged, so this addresses simulator dependence and nothing else. And the shared-family coupling is relocated rather than removed: DeepSeek-V4-Pro is itself one of the general-purpose nurse arms, so under this simulator it is that arm, not the GPT-5.5 arms, whose independence is compromised.

Verdict extraction is more permissive than the format audit. Scoring reads the disposition recorded by the evaluation harness, which accepts any final turn from which a tier can be recovered; a separate format audit applies a stricter template check. The two disagree often, and only the former determines the reported totals. Under the strict check, every scored arm shows template violations—predominantly a line-count mismatch, because the trained models emit a five-line closing (tier, reason, next step, safety net, hang-up) whereas the audit script encodes a three-line template—and the audit additionally marks 30 base and one SFT dialogue as unparseable even though the harness recovers a tier for all but the one SFT case noted above. We report the harness parse because it is the quantity the paired tests operate on, but the gap means our totals are not robust to a stricter output contract, and a deployment that rejected off-template closings would score the base arm far lower than 124/201. No result in this paper depends on the strict audit.

Four caveats bound the claim. First, the SFT arm is a same-recipe reconstruction: the corpus is SHA-256–verified identical and the archived training configuration is byte-identical apart from the output path, and training starts from the same archived base checkpoint, but the retrained parameters are not the historical parameters and the run uses one training seed. Second, the paired units are the 201 profiles, not training runs or simulator realizations: the McNemar test quantifies a checkpoint-level policy-conformance difference on the reused development benchmark and does not propagate uncertainty over training seeds or provider realizations. Third, first caller replies match in only 2/201 cases across arms: under nurse-first openings the simulator conditions on each model's own questions from the first turn, so the comparison is between complete same-day interactive procedures on identical cases, not turn-matched transcripts. Fourth, the four evaluation runs (two arms $\times$ two benchmark halves) were submitted concurrently and three executed in the same window, but one SFT half reran roughly thirty minutes later after a scheduler failure, and cases were not interleaved across arms within a run, so within-day endpoint drift is reduced rather than eliminated. Agreement throughout remains policy-conformance on the adaptively reused benchmark, not clinical accuracy or held-out generalization.

These new-run counts are not estimates of the missing historical verdicts and do not update the historical identified sets, which describe particular historical executions conditional on exact detection. The comparison is still informative about transportability: the retrained SFT count (149) coincides with the upper endpoint of the historical pre-overwrite SFT set (141–149), whereas the clean base count (124) exceeds both the historical stored base count (119) and the historical pre-overwrite base set (98–108), and the clean 25-case gap lies below the historical 33–51 set. These differences illustrate nontransportability across checkpoints and simulator executions – cross-date variation between the June 22 and July 10 historical runs and the July 18 reevaluation, retrain-versus-original parameter differences, or undetected historical overwrites could each contribute – and they reinforce reading the historical sharp sets as conditional statements about destroyed records, not recovered ground truth.

Table 14: Overwrite-free confusion counts for retained first-environment arms with the GPT-5.5 caller. Rows are nurse arms; each column block is one reference tier (42 emergent, 47 urgent, 112 routine) and splits its cases by predicted tier, with ∅ for no disposition. Agr. sums the three correct cells. Generated from the retained per-arm records.

	Ref. Emergent				Ref. Urgent				Ref. Routine				
Nurse	E	U	R	∅	E	U	R	∅	E	U	R	∅	Agr.
GAO-Triage	29	8	5	0	6	29	12	0	6	14	91	1	149
second run	33	8	1	0	4	29	14	0	3	24	85	0	147
Base 9B	4	27	11	0	2	32	13	0	0	24	88	0	124
Outcome GRPO	18	4	0	20	2	25	10	10	3	11	48	50	91
Boundary removed	29	10	3	0	4	33	10	0	1	26	85	0	147
Boundary swapped	35	5	2	0	7	28	12	0	6	19	87	0	150
GPT-5.5	27	12	3	0	2	37	8	0	4	21	87	0	151
given the rule table	28	10	4	0	10	26	11	0	8	22	82	0	136
GPT-5.4	27	14	1	0	6	31	10	0	6	22	84	0	142
DeepSeek-V4-Pro	17	21	4	0	2	28	17	0	0	13	99	0	144
GLM-5.2	19	22	1	0	1	40	6	0	0	32	80	0	139
Qwen3.5-397B	31	11	0	0	7	28	12	0	4	30	78	0	137
Kimi-K2.6	32	8	2	0	9	28	9	1	6	22	83	1	143
Qwen3-235B	15	26	1	0	1	39	7	0	1	40	71	0	125
Unmatched permutation	0	0	42	0	0	0	47	0	0	0	112	0	112

Table 15: General-purpose models driving the nurse role (n=201, clean protocol; GPT-5.5 also serves as patient in all arms; single-seed retrain; run-specific and nontransportable). McNemar compares each arm against the retrained SFT checkpoint (149/201); class recalls are descriptive.

Nurse	Agreement	Wilson 95% CI	Emergent	Urgent	Routine	p vs. retrain
GPT-5.5 zero-shot	151/201 (75.1%)	68.7–80.6%	27/42	37/47	87/112	0.899
GPT-5.5 + 70-row rules	136/201 (67.7%)	60.9–73.7%	28/42	26/47	82/112	0.130
GPT-5.4 zero-shot	142/201 (70.6%)	64.0–76.5%	27/42	31/47	84/112	0.464
DeepSeek-V4-Pro zero-shot	144/201 (71.6%)	65.1–77.4%	17/42	28/47	99/112	0.590
GLM-5.2 zero-shot	139/201 (69.2%)	62.5–75.1%	19/42	40/47	80/112	0.295
Qwen3.5-397B zero-shot	137/201 (68.2%)	61.4–74.2%	31/42	28/47	78/112	0.141
Kimi-K2.6 zero-shot	143/201 (71.1%)	64.5–77.0%	32/42	28/47	83/112	0.532
Qwen3-235B zero-shot	125/201 (62.2%)	55.3–68.6%	15/42	39/47	71/112	0.0043

Retrain−arm paired difference for GPT-5.5 zero-shot is -1.0 points (95% CI -8.7–6.7) and for the rule-prompted arm $+6.5$ points (-1.2–14.2). The Qwen3-235B arm is the only one whose paired difference excludes zero; GLM-5.2 gives $+5.0$ points (-3.4–13.3), Qwen3.5-397B $+6.0$ points (-1.3–13.2), and Kimi-K2.6 $+3.0$ points (-4.8–10.8). Dialogue lengths differ markedly across arms (mean total messages 3.8 for Qwen3-235B, 4.7 for Qwen3.5-397B, 7.9–8.7 for GPT-5.4, Kimi-K2.6, DeepSeek-V4-Pro and GLM-5.2, 12.3 for GPT-5.5, and 20.2 for the retrain), so these arms are not matched on interaction length.

C.9 General-Purpose Models in the Nurse Role

Table 15 reports every general-purpose model we ran in the nurse role on the same 201 cases with `calibration_mode=none`, the same reference, and the same gpt-5.5@temp0 patient protocol. Each arm ran in its own endpoint session rather than concurrently with the clean pair (Section C.8). GPT-5.5 also serves as the patient simulator in every arm, so none of these is an independent cross-model comparison: the patient's dialogue style, and any priors shared between the GPT-5.5 nurse and patient roles, are uncontrolled confounds that plausibly favor the GPT-5.5 nurse arms specifically.

Only the Qwen3-235B arm's paired difference against the retrain excludes zero. The GPT-5.5 zero-shot comparison is close to a coin-flip discordance pattern ($p = 0.899$), and the GPT-5.4 and DeepSeek-V4-Pro comparisons likewise detect no difference. The rule-prompted GPT-5.5 arm numerically favors the retrain by 6.5 points but does not reach significance at $n = 201$ ($p = 0.130$); injecting the rule table did not improve GPT-5.5's agreement over its zero-shot condition (67.7% versus 75.1%), with the Urgent-recall drop (26/47 versus 37/47) accounting for most of the differ-

Table 16: Clean-protocol confusion for the 42 emergent-reference cases. E→E is correct, E→U is one-tier under-triage, E→R is a catastrophic miss, and E→ ∅ is no verdict; E rec. is E→E/42. Counts for all 14 evaluated arms were recomputed from the full case-derived confusion matrices; ROUTINE-only is a separate definitional reference. Run-to-run variation is larger on the catastrophic-miss axis than in aggregate agreement: the two same-recipe SFT runs have 5 versus 1 E→R misses despite 74.1% versus 73.1% agreement.

Nurse	E→E n	E→U n	E→R n	E→ ∅ n	E rec. % (hit/42)
GAO-Triage (ours)	29	8	5	0	69.0 (29)
↪ second run	33	8	1	0	78.6 (33)
Base 9B	4	27	11	0	9.5 (4)
ROUTINE-only (ref.)	0	0	42	0	0.0 (0)
Outcome GRPO	18	4	0	20	42.9 (18)
GPT-5.5	27	12	3	0	64.3 (27)
↪ given the rule table	28	10	4	0	66.7 (28)
GPT-5.4	27	14	1	0	64.3 (27)
DeepSeek-V4-Pro	17	21	4	0	40.5 (17)
GLM-5.2	19	22	1	0	45.2 (19)
Qwen3.5-397B	31	11	0	0	73.8 (31)
Kimi-K2.6	32	8	2	0	76.2 (32)
Qwen3-235B	15	26	1	0	35.7 (15)
Boundary swapped	35	5	2	0	83.3 (35)
Boundary removed	29	10	3	0	69.0 (29)

ence. We have no explanation for that direction in the retained records and flag it as an unresolved observation.

Two patterns are worth separating from the totals. First, the arms differ sharply in emergent recall at similar overall agreement: DeepSeek-V4-Pro reaches 144/201 overall with 17/42 Emergent recall, whereas the retrain reaches 149/201 with 29/42, so arms that look comparable on exact agreement are not comparable on the safety-asymmetric axis. The ordering also reverses. Qwen3.5-397B agrees on 137/201, the lowest of any zero-shot arm besides Qwen3-235B, yet reaches 31/42 Emergent recall; Kimi-K2.6 agrees on 143/201, six fewer than the retrain, yet attains the highest Emergent recall of any general-purpose arm (32/42). Ranking these systems by total agreement inverts their ranking on the axis that matters for under-triage. Second, Qwen3-235B terminates after 3.8 messages on average and escalates 41/112 reference-Routine cases above routine; the records do not isolate why.

We read these comparisons as a deployment-cost motivation, not a capability claim: the retrained checkpoint reaches policy agreement with no detected difference from six general-purpose nurses, while the nurse policy itself issues no frontier-model call. The evaluation as a whole still does: GPT-5.5 drives the patient simulator in every arm, so this is a claim about the deployed policy, not about running the pipeline frontier-free. It does not demonstrate that the checkpoint is more accurate, more robust, or preferable on any axis besides the nurse-side inference cost; the single-seed retrain, the adaptively reused reference, the unmatched interaction lengths, and the shared-model nurse/patient confound all apply here exactly as in Section C.8.

How deterministic is the rule table in practice? The construction interfaces are deterministic as *signatures*: given a set of cited rows, Φ returns the maximum tier. Whether two competent readers cite the same rows for the same dialogue is a separate question, and the retained record answers it. A semantic audit of the training corpus had two rule-conditioned reviewer models independently recompute a tier for each dialogue from the same 70-row table; 930 dialogues carry both reviews. They agree on 410 of 930 (44.1%), Cohen's $\kappa = 0.285$. Excluding the cases where either reviewer declined to commit leaves 685 dialogues at 400 agreements (58.4%, $\kappa = 0.374$), and the residual disagreement is directional rather than symmetric: one reviewer assigns Emergent to 266 dialogues the other calls Urgent or Routine, against five in the reverse direction.

Table 17: Contingency table for the two rule-conditioned LLM reviewers recomputing a tier from the same 70-row table (n=930 training dialogues with both reviews). Rows: Qwen3.6-35B-A3B-FP8; columns: deepseek-v4-flash. "Uns." is a declined commitment, not a tier.

	Emergent	Urgent	Routine	Uns.
Emergent	218	3	2	11
Urgent	102	62	2	87
Routine	164	12	120	137
Uns.	0	0	0	10

Provenance. The audit drew 967 dialogues from the training pool; 930 carry a review from both models and are the denominator here, and in 685 of them both committed to a tier. The reviewers are Qwen3.6-35B-A3B-FP8 and deepseek-v4-flash, each given the rule table, the dialogue, and the assigned symptom family, and each prompted to recompute the tier without seeing the stored label or the other reviewer's output. The agreement analysis uses only the 930 double-reviewed dialogues; none of them entered the published 3,000-dialogue corpus through this audit route. Table 17 gives the full contingency table, from which both figures can be recomputed; the unfiltered review log contains dialogue text and is not released.

We read this as a bound on what "guideline-derived" can mean here. The label that reaches training is deterministic given the cited rows, but the citation step is a judgment these two readers make differently more often than not. The determinism claim of Section 2.2 is therefore about the interface, not about the semantics of applying the table. The estimand is LLM–LLM agreement on policy application: it confounds table ambiguity with reviewer prompting and model behavior, and it is *not* a measured disagreement rate among clinicians, which the record does not contain.

A harness defect that mimicked a weak model. One arm required a fix before it could be scored at all, and the failure mode generalizes. Our gateway adapter fell back to the tail of a model's `reasoning` field whenever `content` came back empty. For Kimi-K2.6—the one evaluated model whose reasoning trace the gateway would not disable—a 512-token budget was consumed entirely by reasoning, so `content` was always empty and partial chains of thought were written into the transcript as though they were nurse turns. The arm scored 82/201 with 99 dialogues lacking any disposition, which reads exactly like a model that cannot triage. The diagnostic that separates the two is cheap: 71% of that arm's nurse turns began mid-sentence, against 0% for every other arm. Removing the fallback and raising the budget yields the 143/201 reported above with 0% mid-sentence turns. We report this because a silent degradation of this kind is indistinguishable from a genuine capability result in a table of totals, and because the same adapter served the GLM-5.2, Qwen3.5-397B, and DeepSeek-V4-Pro arms; we re-audited all three and found 0% mid-sentence turns, so their reported numbers are unaffected.

C.10 Boundary-Interface Ablation

The construction claim names three interfaces, and no *historical* condition isolates any one of them. The boundary interface, however, admits a prospective ablation, because the generation records tag the dialogues it produced: 160 urgent rows, all replacements within their stratum. We therefore built two corpora that differ from the reported one only in those slots. The *removed* arm drops the 160 dialogues (2,840 dialogues, 3,758 exposures). The *swapped* arm replaces each with a same-family non-boundary urgent dialogue that inherits its exposure count, preserving dialogue count, exposure count, and the emergent/urgent/routine balance exactly; without this second arm a deletion would confound boundary content with corpus size. Both are cut from the published weighted corpus rather than re-exported, so prompts and per-dialogue exposure counts carry over verbatim, and both were trained with the archived recipe, one seed, differing from it only in the dataset path.

We score both arms under the primary protocol—the same overwrite-free, live-simulator setting as the headline comparison—so the ablation is read on the scale the paper reports.

Neither ablation separates from the reported arm (swapped, exact McNemar p=1.000; removed, p=0.885), and they do not separate from each other (p=0.76). Removing the boundary dialogues

Table 18: Boundary-interface ablation under the primary protocol (n=201, one seed per arm, `calibration_mode=none`, live simulator). p is exact McNemar against the reported SFT arm.

Arm	Agree	Emergent	Urgent	Routine	p
Reported SFT	149/201	29/42	29/47	91/112	–
boundary swapped	150/201	35/42	28/47	87/112	1.000
boundary removed	147/201	29/42	33/47	85/112	0.885

Table 19: One-factor-at-a-time clean-protocol checks on 201 cases. Exact McNemar p uses the named comparator ("pri." = primary); removal and swapping do not differ (p=0.76). (†) Under the new caller, SFT is compared with base (72.1% vs. 61.7%), and the trained–untrained gap survives.

What varies	Condition	Agr. (%)	p vs.
—	**Primary SFT (9B)**	**74.1**	—
Training run	second run	73.1	0.878 pri.
Patient sim.	DeepSeek caller	72.1	0.022 base†
Boundary iface	160 dlg. swapped	74.6	1.000 pri.
Boundary iface	160 dlg. removed	73.1	0.885 pri.
Repair iface	*no on/off arm exists*	—	—
Rules at inference	GPT-5.5 + 70 rows	67.7	0.130 pri.
its own zero-shot	GPT-5.5 zero-shot	75.1	0.899 pri.

entirely costs two cases of agreement out of 201. We therefore detect no separate contribution of the boundary interface on the protocol the paper reports.

Two caveats apply. The analysis was specified before the arms were scored and named urgent recall as the targeted quantity, because the interface targets urgent boundaries; urgent recall is nominally *higher* without the boundary dialogues (33/47 versus 28/47 and 29/47), the opposite of the intended effect, but these are single-seed class-level contrasts and we read no direction from them. And a null at n=201 with one seed per arm does not exclude an effect: an overall difference smaller than roughly ten points (about 20 cases) would not be resolved here, and the urgent-recall target rests on 47 cases. A deterministic-patient replication of the same three arms reaches the same conclusion, though on an instrument whose repeated caller turns make it poorly suited to elicitation differences.

For the claims of the paper: the three interfaces describe how the corpus was constructed; label assignment supplies every training disposition, no separate contribution is established for the boundary interface, and repair is not separately ablated. We therefore claim no separate contribution for either—which is not evidence that either is unnecessary, only that this record cannot establish one at n=201.

C.11 Three-View Descriptive Accounting by Class

A 39-entry subset of the table over a 25-key schema forms an executable matcher $\hat{\Phi}(x) = r(g(x))$: g extracts whichever schema fields a case narrative x mentions, and r runs the corresponding rule entries. $\hat{\Phi}$ never labels a training row or an evaluation case; it lets one case be read three ways: the disposition $v(h)$ a system states from the transcript h it elicited, $\hat{\Phi}(h)$, and $\hat{\Phi}(x)$ on the complete narrative. We score all three on a dialogue-revised condition (an earlier corpus revision, n=201, on which the model agrees with 147 cases), chosen because it is free of the overwrite defect. The two rule views disagree on 52 cases. Agreement on urgent-reference cases is higher from elicited transcripts than from complete narratives (42.6% versus 38.3%), because extra fields recovered from the full narrative can trigger emergent rules. Of the 72 matcher–reference disagreements, the model is right on 29 of 33 extraction faults but on only 19 of 39 compiled-policy shortfalls (Appendix C.12). Because the reference is policy-coupled, this audits a lossy matcher under two information views; it estimates neither an SFT effect nor generalization.

Table 20: Three-view descriptive accounting for dialogue-revised SFT (201-case development benchmark, $n = 201$). This analysis uses the model-generated disposition directly, without the benchmark-adapted post-interaction heuristic or an additional escalation rule. The model column therefore reports direct model-disposition policy-reference agreement within an environment-constrained interaction.

Reference class	n	agree(model)	agree($\hat{\Phi}$ elic.)	agree($\hat{\Phi}$ full)
Emergent	42	0.810 (34/42)	0.500 (21/42)	0.571 (24/42)
Urgent	47	0.638 (30/47)	0.426 (20/47)	0.383 (18/47)
Routine	112	0.741 (83/112)	0.777 (87/112)	0.777 (87/112)
Overall	201	0.731 (147/201)	0.637 (128/201)	0.642 (129/201)

Agreement outcomes for the model and $\hat{\Phi}$(elicited): both agree on 102/201 cases, both disagree on 28/201, only the model agrees on 45/201, and only $\hat{\Phi}$(elicited) agrees on 26/201.

Table 21: Model-tier confusion matrix against the reference (dialogue-revised SFT, 201-case development benchmark, $n = 201$; one training seed; patient simulated by GPT-5.5). Rows are references and columns predictions. Direct-disposition counts are run-specific and nontransportable, from a constrained interaction without the post-interaction heuristic or an additional escalation rule.

Ref. \ Pred.	Emergent	Urgent	Routine	n
Emergent	34	6	2	42
Urgent	8	30	9	47
Routine	4	25	83	112

Dangerous under-triage
Emergent $\rightarrow$ Routine $= 2$,
Urgent $\rightarrow$ Routine $= 9$ (11/201 $= 0.055$ combined). Emergent $\rightarrow$ Urgent (a delayed-escalation risk this scalar does not count as "dangerous") $= 6$. Over-triage cells:
Urgent $\rightarrow$ Emergent $= 8$,
Routine $\rightarrow$ Emergent $= 4$,
Routine $\rightarrow$ Urgent $= 25$.

Three-view accounting. Agreement with the operational reference is 147/201 for the model disposition, 128/201 for $\hat{\Phi}$(elicited), and 129/201 for $\hat{\Phi}$(full). These are three descriptive measurements on the same cases, not additive error components. Differences may reflect predicate coverage, incomplete or spurious extraction, and model behavior. We report paired transitions without causal labels; the model confusion matrix is in Table 21.

Comparing the complete-narrative and elicited-transcript rule views, $\hat{\Phi}$'s verdict differs on 52 of 201 cases: 22 move from disagreement to agreement, 23 from agreement to disagreement, and 7 between two distinct nonreference tiers. The one-case reduction in elicited-transcript agreement therefore conceals 45 transitions in opposing agreement directions. Additional complete-narrative facts can activate severe rules on cases whose reference label is lower (visible per class above, where complete information *lowers* Urgent agreement 0.426→0.383 while raising Emergent 0.500→0.571). Manual review of five cases identified rules that are duration/degree-insensitive by construction (e.g. a curtain/shadow rule fires "regardless of duration"); fields with no reliable extraction path (e.g. neurological signs); and cases requiring clinical inference not represented by the current predicate set. One traced disagreement (flashes/floaters + myopia $\rightarrow \hat{\Phi}$ returns Emergent, while the reference label is Urgent) also matches a pattern flagged in our reference-label audit (Appendix D): systematic flashes/floaters over-triage.

Table 22: Manual categorization of all 72 $\hat{\Phi}$(full) disagreements with the reference label. "Model agrees" uses the model-generated disposition from the same evaluation as Table 20; it was not modified by the historical post-interaction heuristic.

Review category	n	under	over	model
A concept outside compilation	22	22	0	12/22
B extractor recall miss	14	14	0	13/14
C1 context-blind rule fire	12	0	12	5/12
C2 extractor false positive	19	0	19	16/19
D policy/reference disagreement	5	4	1	2/5
Total	72	40	32	48/72

A: the deciding concept has no predicate/field (pediatric cataract, metamorphopsia, chronicity, growing lesion, only-seeing-eye). B: field exists, narrative states it, extraction missed the phrasing ("multiple images"→`double_vision`, "still dilated"→`pupil_abnormal`). C1: extraction correct, trigger literally matches, context negates (examined-and-cleared, blink-transient, chronic-for-years). C2: historical / hypothetical / negated / non-ocular text asserted as a current positive. D: the source policy or a fired rule assigns a tier different from the adjudicated reference.

C.12 Manual Categorization of Full-Information Disagreements

We recomputed the fired-rule set for every case where $\hat{\Phi}$(full) disagreed with the reference label (72/201), then manually classified all 72 from the source profile, extracted fields, and fired rules. Table 22 reports five categories covering all observed disagreements.

Four findings. **(1) Under-triage disagreements primarily reflect absent rule activation rather than priority misranking:** 33 of the 40 under-triages fired *no rule at all* and defaulted to Routine—21 of the 22 class-A concept gaps plus 12 of the 14 class-B extraction misses—while the remaining 7 (the other 1 and 2, plus the 4 under-triaging class-D cases) fired a rule one tier below the reference. **(2) Low urgent-class agreement reflects missing rule paths:** 17 of the 22 class-A cases have reference tier Urgent. Isolated severe pain, monocular diplopia (a field with no rule that consumes it), chronicity, and progressive-but-minor decline have no path to Urgent in the table; consequently, complete information does not improve $\hat{\Phi}$'s Urgent recall (0.383, Table 20). **(3) Over-triage disagreements are dominated by missing temporal and pragmatic grounding:** of the 19 extractor false positives, 9 interpreted remote surgery or injury history as recent (LASIK 2022 or a 2015 trauma firing "within 6 weeks post-op" rules), 4 assert negated or hypothetical text ("thought it might be drooping but it looks normal"→`lid_droop`), 3 import non-ocular findings ("chronic shoulder and neck pain"→eye pain), and 3 confuse entities (a "flashlight" or a welder's arc →photopsia). **(4) Model agreement is highest on the extraction-side classes:** the dialogue model agrees with the reference on 29/33 (0.879) of B+C2 cases—tense, negation, and relevance judgments a deterministic extractor lacks—versus 19/39 (0.487) on A+C1+D, where the failure is conceptual rather than linguistic. This pattern suggests that the observed model–elicited agreement contrast concentrates on linguistic-grounding cases; it does not establish that triage-specific training, rather than pretrained capability, supplied that behavior.

Independent second-rater check. The categorization above is a single author's manual reading. We additionally scored all 72 cases with an independent rater: GPT-5.5 (temperature 0), given the same category definitions verbatim, the source profile, extracted fields, fired rules, $\hat{\Phi}$(full), and the reference tier, blinded to the author's assigned category. Raw five-way agreement is 47/72 (65.3%), Cohen's $\kappa = 0.553$. Disagreement concentrates at two boundaries: 12 of 22 class-A cases were rated B and 3 of 14 class-B cases were rated A (the concept-inexpressible/extraction-miss distinction), and 4 of 19 class-C2 cases were rated C1 while 2 of 12 class-C1 cases were rated C2 (the false-positive/context-negation distinction). Collapsing to the field-extraction (B+C2) versus rule-or-policy (A+C1+D) split used above does not improve reliability: raw agreement rises to 51/72

(70.8%) but κ falls to 0.422, because most A/B disagreement crosses that same family boundary. The extraction/conceptual contrast on model agreement (finding 4 above) should therefore be read as sensitive to where a case-by-case categorization draws the A/B line, not as a distinction that a second rater reproduces cleanly. This is moderate agreement, not independent validation: an LLM rater is not a human annotator, and GPT-5.5 is also used elsewhere in this paper as an evaluated system, though it played no role in constructing the original categorization or reference tiers. We report this as a bound on the categorization's inter-rater reliability, not as confirmation of its correctness.

One case instead exposes a policy-reference disagreement. The case describes repeated unprotected welding-arc exposure but no eye pain, so the source table's UV-specific Urgent row is not triggered. The compilation nevertheless returns Urgent by mapping welding-arc language to a generic flashes predicate, whereas the medium-confidence operational reference is Emergent and neither blinded LLM supported Emergent. We therefore count this case under policy/reference disagreement, not as evidence that an Emergent rule was lost during compilation. The class-C2 mechanisms also match the two extractor limitations observed separately while implementing the diagnostic (Appendix D.6), and class C1 matches the duration/degree-blind safety rules traced in the five initial case reviews above.

C.13 Interaction-Constraint Sensitivity Analysis

A paired re-run on the same 201 cases, executed separately from the three-view run reported above, toggles a bundle of three deterministic interaction constraints: repeat-question blocking and two checks governing final-answer intent and follow-up clarification. Their exact operational definitions were not retained, so their effects cannot be separated. Model weights are unchanged. Exact policy-reference agreement is 0.731 in both arms, but 41 case-level verdicts differ; the paired agreement-difference 95% CI is $[-0.060, 0.060]$. Class recalls also shift (guards on/off: E_{rec} 0.833/0.762, U_{rec} 0.638/0.681), so equal aggregate agreement is cancellation rather than behavioral equivalence. Repeated-question dialogues are 0/201 versus 1/201, and mean turns are 19.99 versus 20.20; bounded termination is active in both arms. For this model condition, the constraints are rarely activated. Earlier model conditions motivate retaining them as experimental controls: one soft-penalty condition recorded more than 100 repeat loops in a single evaluation (Appendix E). These results distinguish interaction legality from policy-reference agreement but do not establish an agreement benefit from the constraints.

C.14 Construction Strategy Details

This section records the complete measurements behind the construction strategies named in Figure 1 and summarized in Table 23. The corpus was developed adaptively, so these records form a development trace rather than a set of controlled interventions.

Compiled labels, boundary pairs, and repair. The base recipe compiles the operational rule table into traceable labels: a profile stores a nonempty set J of rows cited by its generator and receives $\Phi(J)$, the maximum severity among those cited rows. One-fact boundary pairs then hold the case, time course, and expression fixed while changing one decision-critical fact ($p^+ \sim_1 p^-$). They were designed to concentrate supervision near tier boundaries; a single-seed ablation (Appendix C.10) detected no effect. Metadata-only regeneration likewise defines an intended provenance-preserving repair interface: replacements are conditioned on (j, s, y, f), not the rejected dialogue, evaluation narrative, or case identifier. Historical batches violated this intended contract, so we treat it as a construction specification rather than a verified property.

Phrasing diversity, replacement, and evidence gating. The phrasing-diversity increment added 309 samples in six caller styles (plain speech, metaphor, half-understood jargon, third-person reporting, consequences instead of symptoms, and broken or non-native language) across nine logged categories; because the same increment added 136 emergent examples, its effect is confounded, and the log gives no per-style counts. Within-stratum replacement then substituted failing samples one for one, holding corpus size and class mix fixed; no isolated effect was measured. The evidence-gated configuration changed two things at once, activating family-specific control only under sufficient posterior support and training with the deployment prompt, which does not name the family. It was evaluated under a historical protocol on filtered subsets of 111 to 161 cases, and the

Table 23: The eight construction strategies of Figure 1 and what the retained records establish about each.

Strategy	Purpose	What the records establish
1 Cited-row tier assignment	Fix each label by the rows a profile cites	Labeling mechanism, not ablatable; an unmatched permutation control collapses to the majority class
2 One-fact boundary pairs	Concentrate supervision at tier boundaries	Single-seed pre-specified ablation: 147 and 150 versus 149 of 201; no effect detected, interval still wide
3 Metadata-only repair	Regenerate failures without seeing rejected or evaluation text	Specification only; historical batches violated it; not ablated
4 Phrasing diversity	Vary caller language across six styles	Confounded: the same increment added 136 emergent examples
5 Within-stratum replacement	Replace failing samples one-for-one to keep the class mix fixed	No isolated effect measured
6 Evidence gating	Use family-specific control only under sufficient support	Confounded with strategy 7
7 Train–serve prompt alignment	Train with the deployment prompt	Confounded with strategy 6
8 Label repair	Neutralize or correct contradictory labels	Package-level development evidence only (Figure 7)

retained records lack the scoring rubric needed to interpret its reported question-quality gains, so we attribute no effect to either change.

C.15 Extended Diagnostic Findings

This section preserves the full provenance, version-level trajectories, and contextual diagnostics condensed in Section 6. Unless stated otherwise, the historical observations below are development records rather than same-denominator controlled comparisons.

Shuffled-label control. On the main clean evaluation protocol at the full n=201, the evaluated shuffled-label model predicts Routine on every case, with no unparsed outputs. Agreement is $112/201$=55.7%, with recall $0/42$ for Emergent, $0/47$ for Urgent, and $112/112$ for Routine—exactly the majority-class rate. The permuted training corpus was built from the 2,635-sample `v31final3000_v321` corpus. Label shuffling preserves its class marginals exactly and changes only which dialogue receives which final tier, so the collapse indicates that dialogue format, length, and marginal tier frequency are insufficient without the guideline-derived assignment. The control shares the backbone and training pipeline with the headline SFT arm, but it is not corpus-matched: that arm uses 3,998 weighted exposures from the 3,000-dialogue v52 corpus. On its own, the result is therefore evidence about where the signal lives, not a one-variable control for the headline checkpoint; Appendix I reports a matched control on the v52 corpus, trained and evaluated in the second environment. The authors retain the audit record of this earlier control. The untuned 9B backbone reaches $124/201$=61.7%, only 12 cases above the majority reference, whereas SFT reaches $149/201$=74.1%, 37 cases above it. Thus SFT learns more than the class prior, but much of the backbone's apparent agreement requires no case discrimination.

Safety and the historical protocol boundary. The clean-protocol confusion matrices in Table 16 show that similar agreement can conceal very different error geometry: catastrophic Emergent→Routine counts vary across model arms at comparable agreement and between two same-recipe training runs. The aggregate base–SFT difference is therefore better pinned down than the safety operating point of any particular checkpoint, and total agreement does not order systems by safety.

Historical development observations. Table 24 summarizes the historical development record. These runs used the overwrite heuristic, evolving settings, and, for the last four rows, evaluation subsets filtered after 19–44% patient-gateway contamination (n=111–161); each row changed sev-

eral factors at once. The table therefore records observations rather than intervention effects, and Figure 7 shows the late-checkpoint sequence. Within the added-cases variant, epoch 2 had the highest filtered agreement of the study (72.7%) yet 3/24 Emergent→Routine misses, while epoch 1 reached 63.7% with 0/23: maximizing overall agreement and minimizing catastrophic under-triage can select different checkpoints.

Table 24: Historical development observations (overwrite heuristic enabled). E→R totals count emergent cases called routine over three filtered epochs. Rows are successive variants, not controlled replications.

Change	Observation	Limitation
Aggressive pruning of an early cleaned corpus	Red-flag density rose from about 10% to 36%; routine recall fell to 0.30–0.44 with over-triage	Co-movement only; pruning not isolated
Capability packs (red flags, postoperative, trauma) over four versions	Each pack improved its targeted behavior alone, but their union did not compose monotonically; a long interval since surgery became a shortcut to urgent, later removed	Version traces, not a controlled component attribution
Per-turn structured state (JSON or checklist) instead of prose	Structured-output checkpoints stalled at 0.63–0.68 agreement; the prose line reached about 0.70	Tracks also differed in state contracts and parsing
Stricter family-lock threshold plus a keyword-order fix	No consistent improvement; E→R total 6	Threshold confounded with the fix
Keyword-order fix alone (routing fix)	Neutral to mildly positive; E→R total 2 (2/64)	Filtered, unequal subsets
236 added guideline-derived case-bank examples	Epoch 1: emergent recall 82.6%, agreement 63.7%, over-triage 27.4%; E→R 0/23, 3/24, 1/27	Unequal filtered subsets
Label repair: 491 contradiction rows neutralized, 91 poison rows corrected, oracle leakage removed	Epoch 2: agreement 71.8%, κ=0.537, emergent recall 76.0%, over-triage 18.8%; E→R 0/25, 0/25, 1/21 (1/71)	Three edits not separately ablated; denominators differ from the routing fix

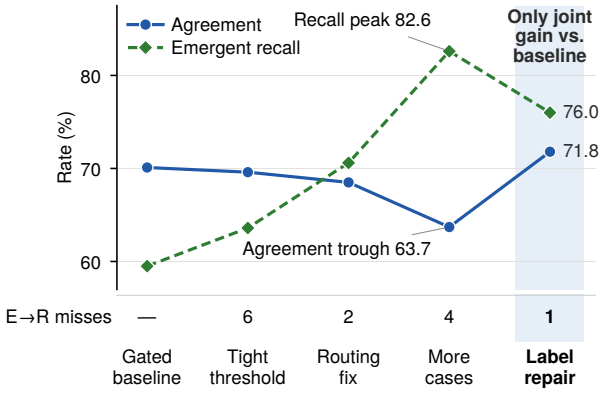


Figure 7: Late-checkpoint diagnostics across five successive development variants; E→R totals span three epochs (a dash marks one not recorded). Only label repair lowers the miss total and raises agreement together. These are successive variants on unequal filtered sets under the historical protocol, not controlled replications.

Literal versus contextual grading. A dual-track audit suggests another plausible contributor beyond row-level noise. A literal application of the rule table marked 39 of 201 reference decisions as disagreements, whereas a contextual audit marked only 2; the difference largely concerned chronic course, prior clinical assessment, or a known cause that led the reference to downgrade urgency.

Training labels instead applied the highest tier whenever a rule fired. The model is thus trained toward a literal convention and scored against a more contextual one. The audit does not decide which convention is clinically correct, and this mismatch was not isolated experimentally. It could plausibly contribute to the roughly 70% development plateau and persistent over-triage; if operative, row-level repairs alone would not reconcile the two grading philosophies.

A discriminability limit of the retained family router. The original 813-case bank contains asymmetric overlap: 40 of 110 `redness_discharge` cases also carry pain (36%), whereas only 16 of 141 `eye_pain` cases carry redness (11%). This bank-level co-occurrence is consistent with `redness_discharge` absorbing `eye_pain` cases; it is not merely a label-code typo, although later audits also found routing and feature-extraction defects. A threshold grid moved posterior-lock precision only from 44% to 51%, at lower coverage, and a stricter setting did not keep improving it. Thresholds can abstain but cannot reorder families or create a missing discriminating feature. This is an empirical ceiling of the current bank–feature–router pipeline, not an irreducible clinical limit or a calibration problem that threshold search alone can solve.

D Integrity Audits and External Evaluations

D.1 Source Governance and Disclosure Controls

Some training-profile seeds and the 201-case development benchmark were abstracted from publicly accessible, patient-authored Reddit posts. The retained project record does not establish the acquisition dates, acquisition mechanism, or platform-approved research access. Evaluation files omit account handles and source URLs, but no comprehensive audit of residual identifiers in free text was retained; we therefore do not characterize the source material as formally anonymized. The project record likewise contains no institutional ethics approval or exemption, and we make no such claim.

Some post-derived text and complete profiles were processed by remote model endpoints. Endpoint contracts and retention settings were not preserved. To limit disclosure through peer-review materials, the released artifact excludes source narratives, derived profiles and transcripts, original identifiers, and crosswalks. Its study records contain only independent random identifiers and categorical labels. These release controls reduce disclosure risk but do not remedy the missing source-governance record.

D.2 Training-Data Contamination Audit

The provenance audit recovered four benchmark-related paraphrase clusters in the safety-balanced source: hemorrhage-description (12 rows), curtain/shadow (16), halos with nausea (19; two near-duplicates), and a postoperative-laser mechanism (12). Detected source-row totals were 57 for the reference corpus, 59 for coverage-expanded, 59 for safety-balanced (118/4,517 exposures, 2.61%), and 0 for rule-stratified. This targeted recovery was not an exhaustive corpus-wide embedding scan; semantic paraphrases escaped the original 8-gram screen.

The contamination arose because failure-focused generation briefs retained benchmark-specific concepts, while a later export discarded the provenance fields needed to trace them. The revised control requires regeneration from only the rule identifier, symptom family, disposition, and contrast requirements. It also retains provenance throughout processing and combines normalized exact matching, 8-gram matching, and semantic-neighbor review (embedding cosine similarity above 0.85). These controls characterize the revised construction procedure; they do not retroactively remove exposure from models trained on the affected historical corpora.

We therefore performed a post hoc decontamination control. Removing the 59 flagged source rows from the safety-balanced corpus reduced the exposure stream from 4,517 to 4,399 rows while leaving the training recipe fixed. At the comparison epoch, the cleaned condition obtained policy-reference agreement of 0.707, $\kappa = 0.467$, E_{rec} 0.619, and U_{rec} 0.319, versus 0.746, 0.574, 0.810, and 0.511 for the original condition. The removal simultaneously eliminated 118 Emergent exposure rows and no Routine rows, thereby confounding contamination removal with Emergent-class exposure. At the fixed first epoch, the cleaned condition has lower κ and E_{rec} for two seeds and higher values for

one; U_{rec} is lower for two seeds and higher for one. Consequently, this control neither identifies a stable contamination effect nor establishes that the flagged rows were harmless.

D.3 Reference-Label Audit

The retained audit generator programmatically marks 25 original labels for review and assigns `AGREE` by default to the remaining 176; it is a candidate checklist, not evidence that 201 manual decisions were completed. The separate 25-row physician-facing worksheet has blank confirmation, final-label, and comment fields in the retained copy; the physician's final tiers were retained separately as a 201-case label file. It agrees with 9 of the 12 author revisions, keeps the original emergent label in two of the other three and assigns routine in the third, and keeps the original label in the other 13 flagged cases and all 176 unflagged ones; we report it as the clinician-adjudicated labeling of Section 4.4. Two rule-conditioned LLM runs omitted the original labels from their prompts and proposed the same alternative for eight of the 25 marked cases. They nevertheless shared the same author table and are not independent clinical annotators. A retained author-adjudication file changed 12 labels. The principal patterns were (i) over-escalation of flashes/floaters to Emergent, (ii) over-grading chronic conditions after recent surgery, (iii) under-triage through omission, and (iv) two pairs with identical case text but opposite original labels. Nine of the 12 changes were downgrades. One Urgent→Emergent change was retained with medium author confidence although neither blinded model supported Emergent; it remains in the reported reference set. Beyond this clinician review of the flagged cases, the available artifacts do not establish independent clinical review. These labels are therefore operational references rather than clinical gold standards.

D.4 Structured-Record and Reason-Grounding Audits

An earlier structured-record corpus was superseded and contributes to none of the reported results. Its row-level audit found duplicated rows, cross-split clones, and unmarked label changes, which the successor builds removed without a measured capability gain.

For the 5,177-row reason-grounded corpus, the rewrite was restricted to `Reason:` lines: no non-reason dialogue turn or final disposition changed. Post-rewrite checks found zero positive-hallucination rows, zero unasked-negative rows, zero contradictions, zero missing Emergent-decisive red flags, and 0/3,266 hallucinated red flags, with 398 unique reason templates. The rewrite normalized every reason line, which accounts for 2,451 changes despite an initial unsupported-reason estimate of approximately 19%. The subsequent dialogue revision contained no strict repeated-question events and preserved the final and system spans exactly.

D.5 Static Overlap Audit for the 5,177-Dialogue Lineage

The dialogue-revised corpus used in the three-view accounting analysis has no row-level provenance fields, so its relationship to the reused benchmark was audited directly from retained text. We compared patient-side utterances in all 5,177 rows with the title and body fields of all 201 benchmark cases after lowercasing and alphanumeric tokenization. Contiguous spans were counted only within an utterance or benchmark field; they were not allowed to cross turn or field boundaries.

At length eight, 112 training rows (2.16%) share a span with five benchmark cases, yielding 112 row–case pairs and seven distinct shared spans. At length ten, four rows share one span with one case. No row shares a contiguous span of 12, 15, or 20 words. A secondary sparse TF–IDF diagnostic over patient-side text gives maximum cosine 0.238 for word unigram–bigram features and 0.208 for character 5-grams; no row reaches a 0.30 similarity threshold. This threshold was not calibrated on independently labeled contamination examples, and TF–IDF is not a semantic-paraphrase classifier. The exact-span results therefore establish limited but nonzero lexical overlap, while neither screen rules out paraphrastic contamination. This is a materially weaker screen than the targeted semantic-neighbor recovery used for four known clusters in the earlier lineage, which found paraphrases that its lexical screen had missed. That earlier recovery was not an exhaustive corpus-wide embedding scan, and no corresponding frozen embedding audit is available for this 5,177-row corpus. The aggregate audit script and a text-free result file accompany the analysis source.

D.6 Extractor Limitations in Three-View Accounting

The frozen transcript-to-field extractor exhibits two relevant limitations. First, a negator in one clause can be assigned to an unrelated positive mention in the same sentence, and a later confirmation cannot overwrite an already confirmed field. Second, a compound question followed by a compound answer can be attributed to only one of the two requested fields. The frozen implementation was retained for every reported three-view measurement to avoid analysis-dependent modification.

On complete narratives, $\hat{\Phi}$-view disagreements with the reference occur in both directions: 32 over-triage and 40 under-triage cases, of which 33 fall in the taxonomy's two extraction-related categories (14 recall misses, 19 false positives). Mean dialogue coverage is 29.4% of 22 fields reachable through the archived mapping. The denominator excludes `severe_photophobia`, `neuro_signs`, and `pain_severity`: none is set by that mapping, and pain presence already serves as its available severity proxy. The released matcher nevertheless accesses all three among its 25 named keys. These limitations motivate interpreting the three-view quantities as descriptive contrasts rather than isolated causal components.

D.7 Family-Level Completeness Audit

Method. Two observed omissions—failure to ask about curtain or shadow in a flashes/floaters dialogue and failure to ask about swelling or eye opening in an eyelid dialogue—motivated an audit across all eight symptom families. This audit's eighth family, trauma/chemical/postoperative, is a construction grouping rather than the rule table's `other` family: the table's injury or chemical-burn row sits in `other`, and its postoperative rows sit in the vision, pain, flashes/floaters, and redness families. The phrasing-diversity increment (Appendix C.14) logged trauma and postoperative separately, giving nine categories. A small, manually constructed knowledge base enumerated canonical symptoms, patient-language variants, discriminative questions, and red-flag combinations. Its inherited literature-source tags were not independently re-verified in this audit and are not used as evidence of clinical validity. Keyword rules proposed candidates; an LLM at temperature 0 assigned each to one of four actions: retain unchanged, revise only the final disposition, regenerate the question path while preserving case facts and label, or place the label in a manual-review queue. Candidate generation by regular expressions was not itself treated as adjudication: false-positive rates against the LLM review ranged from 46% to 71% across families.

Results. All 4,328 candidates (Table 25) were reviewed: 1,913 were retained, 1,497 received a disposition-only revision, 913 were assigned to dialogue regeneration, and 5 to the manual-review queue. The retained record does not establish queue completion. The pupil family had the largest regeneration fraction, 124/166 (74.7%), driven primarily by missing neurological-red-flag questions. Flashes/floaters had the smallest fraction, 12/795 (1.5%), after an earlier targeted repair. Of the regeneration candidates, 217 failed a second audit and were removed together with the 5 manual-review rows. Family- and label-matched generation replaced 219/222 removals: 93.7% on the first pass and 98.6% after one retry. The remaining three-row deficit is entirely in eye-pain/Urgent. Review also found 22/4,328 candidates that the adjudicator flagged as potentially mislabeled but did not route to the review queue as the protocol specified. No family share changed by more than approximately five percentage points, although exact historical per-family proportions could not be independently reconstructed.

The disposition counts were nearly preserved: Urgent differed by 3 rows out of 1,909, while Emergent and Routine were unchanged. Exact duplicates fell from 1,228 to 796.

Post-construction validation. The successor pipeline corrected four defects: unasked fields treated as negative; dangling references in about half of a 12-row continuity sample; a fixed window that omitted the terminal disposition from about one quarter of long dialogues; and position-based rather than content-based row matching. Of 1,615 field corrections, 459 (28.4%) initially had the wrong polarity; 440 were corrected, including 169 denied interventions encoded as attempted. Family-specific prompts recovered most affected rows, but 806 still use the generic prompt, including all audited trauma/chemical/postoperative rows. These facts limit this sensitivity corpus and say nothing about defects in earlier corpora.

Table 25: Family-level completeness audit. The first 80 candidates in the final five families were reviewed before the remainder. Exposure counts include class-weighted replication and precede epoch repetition; deletion removed 217 failed regenerations and five queued rows.

(a) Candidate counts

Symptom family	Candidates
Eye pain	944
Vision loss or change	514
Eyelid symptoms	134
Redness or discharge	1,125
Photophobia	260
Pupil changes	166
Flashes or floaters	795
Trauma/chemical/postoperative	390
Total	4,328

(b) Exposure-example flow

Condition	Examples	Change
Pre-audit baseline	6,703	Baseline
Targeted repair	6,683	−20
Post deletion	6,461	−222
Completeness audited	6,680	+219 replacements

In a separate 30-row reason-grounding sample, 5 rows (16.7%) contained a reason assertion unsupported by the dialogue. All five defects predated this completeness audit. Because the sample is small, the proportion is not interpreted as a corpus-wide estimate.

Training on the completeness-audited corpus yielded loss 0.224 after two epochs. Its best recorded exact agreement was 0.687 at epoch 2 under the historical post-interaction heuristic; no operating checkpoint was designated. This successor condition postdates the principal comparisons and is not a controlled pair, so the value is a procedure-level sensitivity result, not an identified repair effect or pre-heuristic model metric.

D.8 External Full-Information Classification on Oxford-40

Benchmark and design. Oxford-40 is released under CC-BY 4.0 and contains 40 externally authored ophthalmic-emergency vignettes covering cornea, retina, glaucoma, orbit, and neuro-ophthalmology (8 per subspecialty) (Mittal & Aggarwal, 2026). Two ophthalmologists reviewed and approved each vignette, although the source reports neither independent blinded labeling nor inter-rater agreement. We therefore describe it as an expert-reviewed benchmark. Its reference distribution is 23 Emergent, 14 Urgent, and 3 Routine.

Five configurations receive the same vignette text and produce one terminal classification without follow-up questions: deterministic rule execution, a trained 9B precursor, the same 9B backbone before triage adaptation, and zero-shot and guideline-in-prompt GPT-5.5. The trained precursor is evaluated with the generic prompt indicated by the available evaluation code, at temperature 0 with thinking disabled. The pretrained backbone is not instruction-tuned and uses a two-shot plain-completion prompt. Consequently, their comparison holds the backbone fixed but not the prompt format and is a whole-configuration contrast rather than an isolated SFT effect. The base control was added after inspection of the four initially planned configurations; all base-related analyses are post hoc. The three preplanned comparisons were trained precursor versus rule execution, zero-shot GPT-5.5, and guideline-in-prompt GPT-5.5.

Overall results. Exact match is 9/40 for rule execution (22.5%, Wilson 95% CI [12.3, 37.5]), 20/40 for the pretrained 9B configuration (50.0%, [35.2, 64.8]), 25/40 for the trained precursor (62.5%, [47.0, 75.8]), 33/40 for zero-shot GPT-5.5 (82.5%, [68.1, 91.3]), and 25/40 for guideline-in-prompt GPT-5.5 (62.5%, [47.0, 75.8]). Because the reference distribution is Emergent-heavy (23/14/3), a constant-Emergent baseline attains 23/40 (57.5%), two cases below the trained precursor. The zero-shot count is consistent with the 33/40 implied by the source paper's aggregate ChatGPT-5 urgency score of 1.83 ± 0.38 on a 0–2 scale with no zero-score vignette; this is an aggregate consistency check, not a case-level replication.

The trained precursor has 40.0 percentage points higher exact-match agreement than rule execution (20 versus 4 discordant cases; exact McNemar $p = 0.0015$, Holm-adjusted $p = 0.0046$), is 20.0 points lower than zero-shot GPT-5.5 (2 versus 10; $p = 0.0386$, adjusted $p = 0.077$), and has the same exact-match agreement as guideline-in-prompt GPT-5.5 (5 versus 5; $p = 1.0$; Tango-score

paired-difference 95% CI $[-16.6, 16.6]$ points). The latter two conditions nevertheless have different prediction distributions: 34/4/2 for the trained precursor and 26/8/6 for guideline-in-prompt GPT-5.5 over Emergent/Urgent/Routine. The interval permits meaningful differences in either direction and does not support equivalence.

Post hoc same-backbone analysis. The pretrained 9B configuration has 27.5 percentage points higher exact-match agreement than deterministic rule execution despite using only two author-written demonstrations (one Emergent and one Routine). The trained precursor has 12.5-percentage-point higher exact-match agreement, or five additional cases, than the pretrained configuration, but the paired evidence is inconclusive (13 versus 8 discordant cases; exact McNemar $p = 0.383$). Macro recall is 47.96% for the pretrained configuration (Emergent 39.1%, Urgent 71.4%, Routine 33.3%) and 47.76% for the trained precursor (95.7%, 14.3%, and 33.3%). The exact- match difference is therefore concentrated in the majority Emergent class rather than accompanied by higher macro recall.

Class-specific tests were conducted after observing the overall result. On 23 reference-Emergent cases, the trained precursor alone agrees on 13 cases, with no pretrained-only agreements (exact McNemar $p = 0.000244$; Bonferroni-adjusted across three classes, $p = 0.00073$). On 14 reference-Urgent cases, the pretrained configuration alone agrees on eight, with no trained-only agreements ($p = 0.0078$; adjusted $p = 0.023$). Both configurations agree on 1/3 reference-Routine cases, with no discordance. Of the trained precursor's 12 Urgent errors, 11 are predicted Emergent and one Routine, indicating predominantly over-triage but leaving one potentially safety-relevant under-triage.

An additional comparison before multiplicity correction finds 20.0 percentage points higher exact-match agreement for zero-shot GPT-5.5 than for guideline-in-prompt GPT-5.5 (82.5% versus 62.5%; 10 versus 2 discordant cases; exact McNemar $p = 0.0386$). Supplying the operational rule table as unstructured text is therefore associated with lower exact-match agreement for this specific prompt pair, not with a general effect of guideline information.

Same-protocol comparison. The comparisons above hold the backbone fixed but not the prompt. Appendix I evaluates the third GAO-Triage run and the base model under one protocol, with each vignette given as the caller's first message: they agree with 25 and 26 of the 40 labels (p=1.00). In the same environment, the base model reproduces its two-shot result of 20/40 verdict for verdict, so the difference between 20 and 26 reflects the prompt. The precursor's advantage over the two-shot base configuration should therefore not be read as an effect of training.

Construct validity. The vignettes include pupillary examination, slit-lamp findings, and fundoscopy results that a telephone caller generally cannot provide. This is therefore a full-information classification test, not an evaluation of interactive elicitation. The paired estimates characterize the specified configurations; they neither isolate SFT from prompting nor establish how the ranking would change in caller-only encounters.

E Additional Negative and Diagnostic Results

Table 26 collects exploratory and null results under their original protocols; rows are not cross-comparable. Protocols A, A′, B, and B′ include the historical post-interaction heuristic. "Accuracy" means agreement with the operational reference, not clinical correctness.

Table 26: Exploratory and null results omitted from the main text. Protocols and denominators are row-specific; values are not direct comparisons.

Area	Intervention	Protocol	Observation	Interpretation
Data/SFT	Training-set size	Heterogeneous historical protocol; distribution-shifted benchmark	With a fixed recipe, 800, 1,200, 1,500, and 2,000 rows yield 47%, 57%, 43%, and 47% accuracy, respectively, against a 57% baseline.	Performance did not improve monotonically with training-set size in this earlier diagnostic.
Data/SFT	Score-filtered scaling	Heterogeneous historical protocol; score-filtered evaluation	Increasing the filtered pool from 5,000 to 6,000 rows changes accuracy from 61.4% to 58.4%; a 5,500-row variant attains 83% Urgent recall and 33% Emergent recall.	Aggregate scorer values do not preserve the class-specific operating point.
Data/SFT	Provider concentration	Protocol A, provider-concentrated corpus	One generator supplies 86.5% of unfiltered rows; the resulting configuration has $\kappa = 0.394$.	Provider identity and admission filtering are confounded in this experiment.
Data/SFT	Routine-rationale revisions	Protocol A, recitative/ discriminative corpora	The recitative-Routine condition has 30% over-triage. The discriminative-rationale condition has 10% over-triage, 90% Routine recall, 34% Urgent recall, and 19% under-triage.	The conditions occupy different class-specific operating points rather than uniformly improving it.
Data/SFT	Single-turn verdict upsampling	Protocol B′, two single-turn variants	The read-only proxy reaches 69.7%, whereas interactive accuracy is 43.3% (34.8% for the second variant).	The non-interactive proxy does not predict the sequential endpoint.
Data/SFT	Every-turn structured checklist	Protocol B′	The best structured-output arm reaches 67.7% accuracy versus 74.1% for the prose control; a 24.7% general-instruction mixture reaches 35.0%.	The observed values are configuration-level differences and do not isolate a checklist effect.
Data/SFT	Soft interaction penalties	Historical RL training diagnostic; step-50 evaluation	A step-50 evaluation of an RL run with reward-side penalties for repeated questions and missing final verdicts records 130 repeat-loop flags and 44 dialogues without a final verdict, at 48.8% reported accuracy (consistent with 98/201).	The later implementation used deterministic environment-side constraints instead (Appendix C.13); this single diagnostic does not isolate the penalty scheme as the cause.
Evaluation	Reference-label visibility to the judge	Evaluation diagnostic	The judge's agreement is 81.2% when the reference is included in its prompt and 51–55% when blinded.	Reference labels are excluded from all reported judge prompts.
Evaluation	Patient-simulator selection	Pre-evaluation audit	A 397B candidate produces systematic dramatization and excess escalation and is excluded.	Simulator fidelity is treated as an empirical selection criterion.
Evaluation	Zero-shot generator as nurse	Protocol B	Under the historical stack with the heuristic enabled, a Qwen3-235B zero-shot nurse run on 201-case development benchmark scores 101/201 (50.2%; $\kappa = 0.288$), Emergent recall 19/42, Urgent recall 40/47, Routine recall 42/112, averaging 3.5 turns per call. Re-running the same arm under the clean protocol raises it to 125/201 (62.2%; $\kappa = 0.410$), Emergent recall 15/42, Urgent recall 39/47, Routine recall 71/112; the main text reports the clean figure.	Both runs show the same short-interaction, Urgent-defaulting pattern; neither identifies effects of scale or fine-tuning, and the gap between them illustrates how much the heuristic distorts stored totals.

F Optimization Configurations

F.1 SFT configurations by experiment family

Table 27 lists the SFT configuration of each experiment family.

F.2 Post-Training Configurations

One optional post-training route starts from the boundary-revised SFT substrate. We report it under the clean protocol (`calibration_mode=none`, GPT-5.5 patient at temperature 0), and give the corresponding heuristic-enabled historical total for contrast, because the two disagree qualitatively.

Table 27: SFT configurations (LLaMA-Factory (Zheng et al., 2024)). Shared settings are listed once; context and microbatch settings differed across experiment families.

Base model	Qwen3.5-9B-Base (Qwen Team, 2026), full-parameter fine-tune
Distributed training	DeepSpeed ZeRO-3, bf16, FlashAttention-2
Learning rate/schedule	5×10^{-6}, cosine, warmup ratio 0.03
Early SFT	early conditions preceding the reference corpus (Table 3), including the rule-stratified and class-ratio-pilot variants: maximum sequence length 32,766; per-device batch size 4 $\times$ gradient-accumulation factor 2 $\times$ 4 GPUs
Primary controlled SFT	reference corpus through boundary-revised conditions: maximum sequence length 4,096; per-device batch size 2 $\times$ gradient-accumulation factor 4 $\times$ 4 GPUs
Dialogue-revised SFT	the flexibility-revised, reason-grounded, and dialogue-revised conditions: maximum sequence length 6,144; per-device batch size 2 $\times$ gradient-accumulation factor 4 $\times$ 4 GPUs
Effective batch size	32 for all three families
Epochs	3, with one model state retained after each epoch
Loss masking	supervision on all nurse tokens; caller tokens condition prediction but are masked from the loss
Template	Qwen3 chat template with reasoning disabled
Class weighting	deterministic weighted upsampling (safety-balanced condition: E:2.0, U:1.5, R:1.0)
Training time	$\approx$74 min for 5,177 rows, one epoch, on 4 GPUs (dialogue-revised family only; not separately timed for the other two families)

Outcome-level RL uses GRPO (verl; group size 8, KL coefficient 0.08, learning rate 2×10^{-7}, at most 25 steps) with outcome, rule-side symptom, format, and LLM-judged process rewards; training uses a Qwen3-235B patient and reports step 15. Random seeds were not retained.

Under the clean protocol, SFT and RL agreement is 149/201 and 91/201; E/U/R correct counts are 29/29/91 and 18/25/48, with 1 and 80 missing verdicts. Under the historical heuristic-enabled stack, the corresponding agreement totals are 151/201 and 151/201, both with zero missing verdicts, so neither pattern reproduces cleanly. For RL, 77 of 80 missing-verdict dialogues stop before the turn budget and 72 after one exchange; conditional agreement 91/121 (75.2%) is selected on that failure. The clean substrate is a same-recipe retrain because the original weights were deleted, whereas the post-training arm retains the original lineage, so these are checkpoint rather than matched-initialization comparisons.

F.3 Retained Reproducibility Boundary

Table 27 reports settings recoverable from retained configurations and logs. Two further trained conditions, safety-operating-point SFT and moderate-U SFT (Table 4), are class-reweightings of the

boundary-revised source corpus and are not separately listed; retained records indicate they reuse the primary controlled SFT row's settings, but this is not independently confirmed by a dedicated log entry. The record is otherwise incomplete for several consequential details, including random seeds and some optimizer defaults across SFT families and the hardware model. We therefore do not characterize it as a complete training specification.

Appendix B states the review-artifact boundary. In addition to its rule and accounting fixtures, the package regenerates the three-view and matched-control summaries and the Oxford-40 comparisons from public per-case labels. It excludes source narratives, weights, raw overwritten verdicts, a redistributable extractor revision, exact historical PMC prompts and sample order, and the embedding configuration for the retrospective neighbor audit; the PMC results are therefore recorded analyses, not independently regenerable evaluations.

G Question Audit and Emergent-Miss Attribution

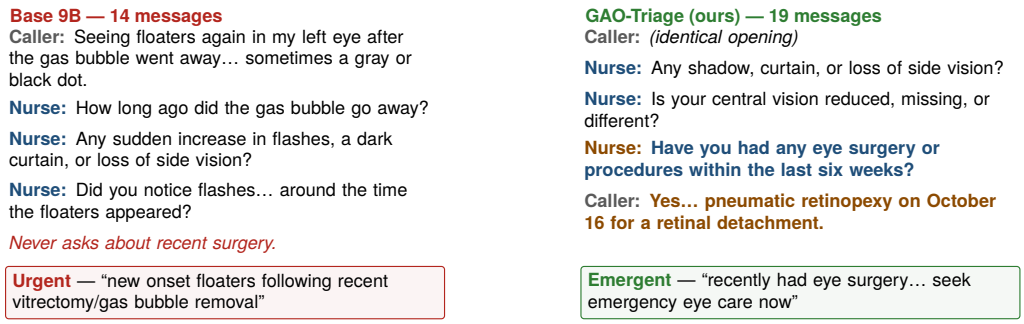


Figure 8: One paired emergent case under the overwrite-free protocol; excerpts are verbatim, shortened by ellipses. Both callers open by mentioning a gas bubble. The base model infers a recent vitrectomy in its reason yet calls the case urgent; GAO-Triage asks about procedures in the last six weeks, obtains the date of the retinal procedure, and escalates. One case illustrates the aggregate analysis of Section 4.2 but is not evidence for it.

Keyword coding. The audit in Section 4.2 reads the retained overwrite-free transcripts of four arms (201 calls each) and codes, for every call, whether any nurse turn that does not match the verdict pattern ("Triage Result") mentions each of five items. The rules are case-insensitive regular expressions: *surgery* matches "surg", "procedure", "operation"/"operated", "laser", or "injection"; *injury or chemical* matches "injur", "trauma", "chemical", "splash", "accident", "hit" in phrases such as "hit your eye" or "been hit", "struck", "poked", "foreign body/object", or "something got in"; *contact lens* matches "contact lens" or "contacts"; *which eye* matches "which eye", "right or left", "one eye or both", or "both eyes"; *onset* matches "when did", "how long", "start(ed)", or "sudden". The script `iclr_additions/scripts/elicitation_audit.py` of the supplementary material reproduces all numbers from the evaluation records, which are not released; it calls no model, and `iclr_additions/scripts/elicitation_tests.py` recomputes the tests from the released per-call flags.

Checks on the coding. For surgery and for injury, we read eight randomly drawn matching sentences in each of the base and GAO-Triage arms (32 in total); 31 of the 32 were questions about the coded item; the exception is a base-model turn that mentions trauma inside an explanation. Several base-model matches ask about lifetime history ("any history of eye injuries, surgeries") rather than recent events, which counts in the base model's favor. To check for misses, we searched calls without a match for near synonyms. In the base model this found questions such as "Is there any chance of recent trauma" and "getting hit", which led us to add the "been hit" and "getting hit" patterns; the remaining near-synonym hits were not questions about the item (for example, "When did you last see an eye doctor?"). This spot check found one false positive among 32 matches; it does not estimate the miss rate, and the "hit" patterns were added after inspecting base-model misses, so the base-model rates may still be slightly understated.

Table 28: Mean messages and non-verdict nurse turns per call, and the share of the 201 calls (%) in which a nurse turn other than the verdict mentions each item. Items are coded by keyword rules, so a mention is not necessarily a well-formed question (Appendix G); p is the exact McNemar test between GAO-Triage and its base model on the same calls (which eye: p=0.27; onset: p=0.080).

	Base 9B	GAO-Triage	Second run	GPT-5.5	p (GAO vs. base)
Messages per call	20.2	20.2	20.6	12.3	
Non-verdict nurse turns per call	9.8	9.5	9.7	5.6	
Eye surgery or procedure	47.8	92.5	94.5	87.6	5.7×10^{-22}
Injury or chemical exposure	66.7	89.1	96.5	87.1	3.6×10^{-7}
Contact-lens use	84.6	95.5	99.0	54.7	4.7×10^{-4}
Which eye	47.3	52.7	54.7	46.8	
Onset or timing	87.6	93.0	95.0	72.1	

Table 29: Topic-mention rates on the 42 emergent-reference calls (%), with the exact McNemar test between GAO-Triage and its base model on the same calls.

	Base 9B	GAO-Triage	Second run	GPT-5.5	p (GAO vs. base)
Eye surgery or procedure	45.2	88.1	88.1	64.3	2.8×10^{-4}
Injury or chemical exposure	59.5	90.5	97.6	61.9	4.4×10^{-3}
Contact-lens use	81.0	90.5	97.6	26.2	0.39
Which eye	47.6	52.4	54.8	31.0	
Onset or timing	81.0	88.1	95.2	54.8	

Emergent-miss attribution. GAO-Triage misses 13 of the 42 emergent-reference cases (8 called urgent, 5 routine). For each, we read the transcript and looked for an emergent row of the released table whose conditions are all stated in the call. Table 30 lists the result without case text; a reason is marked unsupported when it attributes to the caller a finding the caller did not report.

Table 30: The 13 emergent-reference cases GAO-Triage missed (first run). "Row" is the emergent row of the table whose conditions are stated in the call, or "none" if no emergent row applies. All conditions of the listed row are present in the transcript in cases 1–10.

#	Called	Emergent row satisfied by the call	Stated facts	Reason
1	U	Flashes/floaters after surgery or procedure within 6 weeks	floaters; retinal laser days earlier	supported
2	U	same	dark spot with flashes; retinal laser days earlier	supported
3	R	same	floaters; retinal laser days earlier	supported
4	U	Vision changes after surgery or procedure	distortion after retinal surgery 8 weeks earlier	unsupported
5–8	U	Redness or discharge in a contact lens wearer	red eye; contact-lens wearer	supported
9	R	Total darkness or part of vision missing, new onset	new fixed dark spot in central vision	supported
10	R	Shadow, curtain, or enlarging visual field defect	dark area in peripheral vision	supported
11	U	none (only the urgent flashes row applies)	new flashes; known retinal tear; no curtain	supported
12	R	none identified	sharp pain and redness after eye rubbing	supported
13	R	none identified	blood from the eye margin; eye otherwise normal	unsupported

H REFERENCE ROBUSTNESS, TRIAGE SAFETY, AND FACT POSITIONS

This appendix details the analyses behind Section 4. Every statistic in the tables below is recomputed by `iclr_additions/scripts/robustness_tests.py` from `iclr_additions/data/per_case_verdicts.json` in the supplementary material; the case selection of Appendix H.4 was done by reading the posts and is described there. For each of the 201 cases, under a fresh random identifier, that file holds the five labelings, the verdict of every arm with retained per-case output, the symptom family of the rule the auditors cited, and the verified-fact annotations described below; it contains no text.

H.1 ALTERNATIVE LABELINGS

The clinician-adjudicated labels are the final tiers a clinician assigned to the 25 flagged cases, with the original tiers elsewhere (44 emergent, 44 urgent, 113 routine). The clinician agreed with 9 of our 12 revisions, kept the original emergent label in 2 of the other 3, and assigned routine where we assigned urgent in the third, and kept the original label in all 13 flagged cases we left unchanged; the resulting labels differ from the main reference in three cases (Section 3). The original labels are the tiers the dataset carried before the 12 revisions of Section 3 (46 emergent, 46 urgent, 109 routine). Auditors A and B are the two rule-conditioned LLMs of the same section; each read the post and the rule table, not the original label, and assigned one tier to every case without abstaining (auditor A, DeepSeek-V3: 32/29/140 emergent/urgent/routine; auditor B, Qwen3-235B: 26/49/126). The two auditors agreed on the same alternative for 8 of the 25 flagged cases. The consensus subset keeps the 124 cases on which all five labelings agree (17/13/94). Table 32 extends Table 31 to the second training run and to GPT-5.5. The second run's agreement gain over the base model is significant under the main reference and the clinician-adjudicated labels, while its emergent-recall gain is significant under every labeling ($p \leq 0.001$). GAO-Triage and GPT-5.5 differ by at most 1.5 points under any labeling.

Table 31: GAO-Triage (G) versus its base model (B) under the main reference and the five reference-robustness views of Section 3. n counts cases (emergent in parentheses); agreement and emergent hits are counts, Δ is the paired agreement difference in points with its 95% interval, and p values are exact McNemar tests, uncorrected.

Labeling	n (E)	Agr. G / B	Δ	95% CI	p	E hits G / B	p
Main reference	201 (42)	149 / 124	+12.4	[4.3, 20.6]	0.0046	29 / 4	4.2×10^{-7}
Clinician-adjudicated	201 (44)	146 / 123	+11.4	[3.4, 19.5]	0.0086	29 / 4	4.2×10^{-7}
Original labels	201 (46)	142 / 122	+10.0	[1.9, 18.0]	0.022	27 / 3	8.0×10^{-7}
Auditor A (DeepSeek-V3)	201 (32)	146 / 120	+12.9	[4.7, 21.1]	0.0034	23 / 5	4.0×10^{-5}
Auditor B (Qwen3-235B)	201 (26)	134 / 128	+3.0	[−5.2, 11.1]	0.55	20 / 5	2.7×10^{-4}
Consensus of all five	124 (17)	104 / 90	+11.3	[1.2, 21.3]	0.044	13 / 2	0.0034

Table 32: Agreement and emergent hits under each labeling for GAO-Triage (G), its second training run (G2), the base model (B), and GPT-5.5. The last two columns give exact McNemar p-values for G2 versus B and for G versus GPT-5.5 agreement.

		Agreement				Emergent hits				p	
Labeling	n (E)	G	G2	B	GPT-5.5	G	G2	B	GPT-5.5	G2 vs. B	G vs. GPT-5.5
Main reference	201 (42)	149	147	124	151	29	33	4	27	0.013	0.90
Clinician-adjudicated	201 (44)	146	145	123	148	29	33	4	27	0.017	0.90
Original labels	201 (46)	142	137	122	143	27	30	3	24	0.11	1.00
Auditor A (DeepSeek-V3)	201 (32)	146	133	120	143	23	25	5	24	0.17	0.79
Auditor B (Qwen3-235B)	201 (26)	134	134	128	136	20	21	5	19	0.57	0.90
Consensus of all five	124 (17)	104	97	90	105	13	13	2	15	0.36	1.00

H.2 TRIAGE SAFETY FOR EVERY ARM

Table 33 counts under-triage (a disposition below the reference tier), over-triage (above it), emergent cases called routine, and calls without a disposition, which are excluded from the first three

counts. Per case, GAO-Triage under-triages 8 calls the base model does not, and the base model 34 calls GAO-Triage does not (paired difference -12.9 points, 95% CI -19.0 to -6.9, p=6.9×10^{-5}); over-triage discordance is 17 versus 17 (p=1.00). For the second run the corresponding counts are 8 versus 36 for under-triage (p=2.5×10^{-5}) and 20 versus 15 for over-triage (p=0.50). Among general-purpose models the balance varies widely: DeepSeek-V4-Pro under-triages 42 calls and over-triages 15, whereas Qwen3.5-397B under-triages 23 and over-triages 41.

Table 33: Triage errors against the main reference (201 calls per arm). Under and Over count dispositions below or above the reference tier; E→R counts emergent cases called routine; No v. counts calls without a disposition.

Nurse	Under	Over	E→R	No v.
GAO-Triage	25	26	5	1
GAO-Triage (second run)	23	31	1	0
Base 9B	51	26	11	0
Outcome GRPO	14	16	0	80
Boundary removed	23	31	3	0
Boundary swapped	19	32	2	0
Unmatched permutation	89	0	42	0
GPT-5.5	23	27	3	0
GPT-5.5 given the rule table	25	40	4	0
GPT-5.4	25	34	1	0
DeepSeek-V4-Pro	42	15	4	0
GLM-5.2	29	33	1	0
Qwen3.5-397B	23	41	0	0
Kimi-K2.6	19	37	2	2
Qwen3-235B	34	42	1	0

H.3 Symptom families

Cases carry no family label of their own, so we use the family of the table row that each auditor cited and keep it when both auditors cite rows of the same family; the remaining 49 cases are grouped as "auditors disagree". Table 34 reports agreement by family. Most families contain fewer than 25 cases, so the table is descriptive; the largest differences between GAO-Triage and its base model appear in vision loss or change, eye pain, and the `other` family, which holds the injury and chemical rows.

Table 34: Agreement with the main reference by symptom family (family of the auditors' cited rows). E/U/R is the reference distribution within the family.

Family	n	E/U/R	GAO-Triage	Second run	Base 9B	GPT-5.5
Vision loss or change	49	8/13/28	42	39	34	40
Flashes and floaters	23	7/8/8	12	18	11	16
Eye pain	9	5/0/4	8	6	3	4
Redness or discharge	15	5/2/8	10	11	9	11
Eyelid	16	0/5/11	13	14	13	14
Pupil	4	0/1/3	4	4	4	2
Photophobia	1	0/0/1	1	1	0	1
Other (incl. injury)	35	8/1/26	28	24	20	27
Auditors disagree	49	9/17/23	31	30	30	36

H.4 Verified facts and where they surface

Case selection. We searched the post text of all 201 cases for procedure terms (surgery, operation, laser, LASIK, PRK, vitrectomy, intravitreal, injection, scleral buckle, IOL, lens replacement or implant, cataract removal) and for injury terms (injury, trauma, chemical, bleach, splash, hit, struck, poked, punch, foreign body or object, welding, "stepped on", something getting in the eye), and read the context of every hit. Of 40 procedure hits we kept 31 cases that report a past eye surgery

or procedure, excluding six that mention only a planned or hypothetical operation, one occupational mention, one that mentions surgeons without a procedure, and one ambiguous injection. Of 22 injury hits we kept 11 cases that report an eye injury, a foreign body, or a chemical or welding-arc exposure, excluding five negations, five non-ocular or non-injury uses (for example emotional trauma or a laser "hitting" the retina), and one ambiguous mention. One case appears in both sets, so the 42 facts cover 41 cases, 22 of them emergent under the main reference.

Positions. For each fact and arm we locate the first caller message that matches the fact terms and the first nurse turn (other than the verdict) that matches the question keywords of Table 28. A fact surfaces in the *opening* if the first matching caller message is the caller's first message, *after a question* if a matching nurse turn precedes it, *later unprompted* otherwise, and is *absent* if no caller message matches. Keyword matching can miss paraphrases and can count a negated mention ("no recent surgery"), and "later unprompted" may include questions phrased without the matched keywords; the positions are therefore approximate, but they are computed identically for every arm.

Table 35: Where the 42 verified facts first surface in each arm's transcripts (Opening: the caller's first message; Asked: after a matching nurse question; Unprompted: later, without one), and emergent dispositions on the 22 emergent cases among them, overall and when the fact is in the opening.

	First surfaces (42 facts)				Called emergent	
Nurse	Opening	Asked	Unprompted	Absent	All	Opening
Base 9B	24	4	5	9	3/22	3/16
GAO-Triage	24	4	12	2	18/22	12/16
GAO-Triage (second run)	26	8	6	2	19/22	14/16
GPT-5.5	24	9	4	5	15/22	10/16

H.5 Run-to-run stability and complementarity

The two GAO-Triage training runs use the same configuration file apart from the output directory, so they also share the default random seed; they differ only through run-to-run nondeterminism. They give the same verdict in 151 of 201 calls; both are correct on 127, only the first run on 22, and only the second on 20. Aggregate agreement is therefore more stable than individual verdicts, and a per-case analysis of one run should be read with this variability in mind. GAO-Triage and GPT-5.5 give the same verdict in 134 calls and are correct on different cases: 30 only for GAO-Triage and 32 only for GPT-5.5, so at least one of the two is correct on 181 cases. Against its base model, GAO-Triage is correct alone on 49 cases and the base model alone on 24.

I Controlled Retraining in a Second Environment

We trained six further models on a separate cluster: three runs of GAO-Triage (the third, fourth, and fifth runs), a label-permutation control on the same corpus, and two runs of a second, disposition-permutation control. We evaluated the first four, together with the base model, in the interactive protocol, in a full-information condition, and on Oxford-40; the disposition-permutation control was evaluated in the interactive protocol (first run) and in the full-information condition (both runs). We also give the cost-sensitive comparison behind Figure 3.

Training. All six models start from the same Qwen3.5-9B-Base weights and use the archived GAO-Triage configuration (Table 27; three epochs, effective batch size 32, maximum length 4,096). The third run and the matched control use the default random seed (42), as did the first two runs; the fourth and fifth runs use seeds 43 and 44, and nothing else differs among the three runs. The two disposition-permutation runs use seeds 42 and 43. Each model trained on four A100-80GB GPUs in about 75 minutes, two models at a time. Relative to the first two runs, the cluster and software installation differ (Transformers 5.6, DeepSpeed 0.18), and attention uses PyTorch SDPA instead of FlashAttention-2; both change the order of floating-point accumulation, not the model or the objective. The three runs train on the SHA-256-verified corpus of Section 2.3. Final training losses are 0.189, 0.188, and 0.189 for the three runs, 0.192 for the matched control, and 0.179 and 0.180 for the disposition-permutation runs, whose final messages are fixed templates.

Matched label permutation. The control's corpus is the same 3,998 training exposures with the tier word of every final nurse message permuted across the corpus. The class counts are unchanged (1,250 emergent, 1,119 urgent, and 1,629 routine exposures), 2,664 exposures change tier, and all other text is identical. Because the permutation acts on exposures, 642 of the 3,000 dialogues, which appear more than once after class-weighted resampling, carry more than one tier. The corpus uses one fixed next-step sentence per tier (for emergent, "Please seek emergency eye care now or go to the emergency department immediately"), and the permutation leaves these sentences in place. In the control, the next-step sentence therefore still states the rule-derived disposition while the tier word does not. Unlike the earlier control of Appendix C.15, this one matches GAO-Triage in corpus, size, class weighting, and configuration.

Disposition permutation. The matched control leaves the rule-derived tier in the rest of each final message: in its next-step sentence, in its safety-net sentence, which differs between emergent and the other tiers, and in 716 reasons that name a tier. A second control removes all of it. Every exposure keeps its permuted tier from the matched control, so the class counts are again 1,250, 1,119, and 1,629, and its final nurse message becomes the fixed template of that tier: the tier word, the tier's next-step and safety-net sentences, and one neutral reason shared by all exposures ("Based on the information the caller provided."). Every exposure is identical to the original corpus up to its final message, which takes three distinct forms, one per tier. No text of the final message therefore depends on the rule-derived tier, although rule information remains in the earlier turns, generated under that tier, and in the family-specific training prompts, which state decision cues. We trained this control twice, with seeds 42 and 43 and otherwise the configuration of the matched control, and evaluated the first run in both conditions and the second with full information only.

Evaluation. Checkpoints are converted to text-only causal language models as in the original evaluation code and scored with the archived evaluation runner of the Table 1 period, unmodified: generic nurse prompt, greedy decoding with repetition penalty 1.15, a 20-turn limit, and no over-write heuristic. The caller is GPT-5.5 at temperature 0, reached through a different API gateway, which adds about 300 prompt tokens of its own to every request. Because the environment differs, every comparison below stays within it: the three runs and the controls are compared with the base model evaluated here, not with Table 1.

Full-information condition. The runner's complete-case mode gives the nurse the entire post as the caller's first message. Any further question receives the fixed reply "I don't have anything else to add beyond what I already described," no caller simulator is involved, and the nurse is instructed to give a disposition in its last two turns (one call still ended without one). The nurse prompt is the generic prompt of the interactive condition. The available information is thus fixed at the complete post, so errors in this condition cannot come from facts the model failed to elicit.

Trained runs against the base model. Table 36 gives the results. All three runs reproduce the emergent gain of the first two: emergent recall is 32, 30, and 23 of 42 against 3 for the base model (all $p \leq 1.9 \times 10^{-6}$), and under-triage falls from 55 to 22, 21, and 28 calls (all $p \leq 7.4 \times 10^{-6}$). None reproduces the overall gain: agreement is 138, 134, and 119 of 201 against 130 ($+4.0$, $+2.0$, and -5.5 points; p=0.46, 0.74, and 0.27), because every run also over-triages more (41, 46, and 54 calls against 16; all $p \leq 1.1 \times 10^{-4}$) and recalls fewer routine cases (82, 72, and 63 of 112 against 97). The same holds under every view (Table 37): the emergent gain is significant for every run under all six views, the overall gain for none, and the fifth run falls significantly below the base model under auditor B and on the consensus subset.

Environment and run-to-run variation. The same base weights agree on 130/201 here and on 124/201 in the original environment (p=0.47), giving the same verdict in 149 of 201 calls. The third run agrees on 138/201 against 149 and 147 for the first two runs (p=0.12 and 0.25) and shares 149 and 146 of their verdicts. Within this environment, runs that differ only in their random seed share 133 to 149 verdicts pairwise and differ in agreement by up to 19 cases: the fifth run agrees on fewer cases than the third (119 against 138, p=0.018) and the fourth (134, p=0.053), while the third and fourth do not differ (p=0.66). Numerical and gateway differences, for their part, change about a quarter of per-case verdicts even at temperature 0 with greedy decoding. We therefore compare arms only within one environment and report every run.

Table 36: Results in the second environment on the 201 reference cases (42 emergent, 47 urgent, 112 routine). Agreement and class recalls are counts; Under and Over count calls whose disposition is below or above the reference tier; No v. counts calls without a disposition; Msgs. is the mean number of messages per call. The full-information condition gives the complete post as the caller's first message.

Nurse	Condition	Agr.	E	U	R	Under	Over	No v.	Msgs.
Base 9B (re-evaluated)	interactive	130	3/42	30/47	97/112	55	16	0	21.9
GAO-Triage, third run	interactive	138	32/42	24/47	82/112	22	41	0	22.2
GAO-Triage, fourth run	interactive	134	30/42	32/47	72/112	21	46	0	21.6
GAO-Triage, fifth run	interactive	119	23/42	33/47	63/112	28	54	0	21.4
Matched label permutation	interactive	100	0/42	0/47	100/112	83	18	0	21.5
Disposition permutation	interactive	107	2/42	0/47	105/112	83	11	0	21.0
Base 9B (re-evaluated)	full information	131	7/42	18/47	106/112	62	8	0	3.8
GAO-Triage, third run	full information	135	33/42	12/47	90/112	28	37	1	20.2
GAO-Triage, fourth run	full information	123	28/42	26/47	69/112	25	53	0	20.7
GAO-Triage, fifth run	full information	118	28/42	29/47	61/112	25	58	0	20.4
Matched label permutation	full information	110	0/42	0/47	110/112	88	3	0	19.9
Disposition permutation	full information	108	1/42	0/47	107/112	85	8	0	19.6
Disposition permutation, run 2	full information	112	0/42	0/47	112/112	89	0	0	20.6

Table 37: The three second-environment runs against the base model evaluated in the same environment (B), under the main reference and the five reference-robustness views of Section 3. Agr. B and E B are the base model's agreement and emergent hits; for each run, Δ is the paired agreement difference from B in points with its exact McNemar p, and E its emergent hits. Every emergent difference has $p<0.001$.

				Third run		Fourth run		Fifth run	
Labeling	n (E)	Agr. B	E B	Δ (p)	E	Δ (p)	E	Δ (p)	E
Main reference	201 (42)	130	3	+4.0 (0.46)	32	+2.0 (0.74)	30	−5.5 (0.27)	23
Clinician-adjudicated	201 (44)	127	3	+4.0 (0.46)	32	+3.0 (0.59)	31	−5.5 (0.27)	23
Original labels	201 (46)	126	3	+2.5 (0.67)	30	+2.0 (0.74)	29	−6.0 (0.22)	21
Auditor A (DeepSeek-V3)	201 (32)	131	4	+1.0 (0.91)	26	−1.5 (0.83)	27	−7.5 (0.13)	22
Auditor B (Qwen3-235B)	201 (26)	134	3	−3.0 (0.58)	22	−3.5 (0.51)	22	−10.9 (0.021)	17
Consensus of all five	124 (17)	96	2	−0.8 (1.00)	15	−3.2 (0.66)	15	−12.9 (0.029)	13

Matched label permutation. The matched control answers routine in 183 of 201 calls and emergent in the other 18; it never answers urgent. It agrees on 100/201 with emergent recall 0/42, against 138/201 and 32/42 for the third run (+18.9 points, 95% CI 9.8 to 28.0, $p=1.1\times10^{-4}$) and 134 and 119 of 201 for the fourth and fifth runs, and falls below the base model (−14.9 points, $p=1.0\times10^{-4}$). Its topic mentions, by contrast, are about as frequent as those of the trained runs or more frequent: under the keyword coding of Table 28, it raises eye surgery or procedures in 190 calls (three runs 180 to 195, all $p\geq0.06$; base model 79), injury or chemical exposure in 187 (174 to 183), and contact-lens use in 197 (194 to 197). Raising the same topics therefore does not produce the tier-word gain once the tier words are permuted.

A post hoc reading of the same transcripts, added after we saw the control's outputs, shows where the tier went. Parsing the next-step sentence of each final message, which the permutation left tied to the rule-derived tier, the matched control states a disposition that agrees with the reference in 126 of 201 calls (emergent 25/42, urgent 18/47, routine 83/112), not significantly fewer than the third run's tier word (138/201; $p=0.21$). Its tier word and its own next-step sentence agree in only 97 calls. The model thus learned the rule-derived disposition from the text that kept it, while the tier word, whose supervision was permuted, fell back to the class prior. In the third run, tier word and next-step sentence agree in 200 of 201 calls.

Disposition permutation. The disposition-permutation model answers routine in 188 of 201 interactive calls and emergent in the other 13; it never answers urgent. It agrees on 107/201 with

emergent recall 2/42, where verdicts independent of the reference, in the same mix, would give 107.5 and 2.7. The third run agrees on 138 and calls 32 emergent cases emergent ($+15.4$ points, 95% CI 6.3 to 24.6, p=0.0017; emergent p=1.5×10^{-8}), and the model does not differ from the matched control (100/201 and 0/42; p=0.12 and 0.50). Its next-step sentence now follows its tier word in all 201 calls. Its questions are those of the trained runs: it raises eye surgery or procedures in 190 calls, injury or chemical exposure in 170, and contact-lens use in 196, not significantly different from the third run on any topic (smallest p=0.27) and more often than the base model on all three (all $p\leq6.6\times10^{-5}$), although less often than the matched control for injury or chemical exposure (187; p=0.0095). With full information, its two runs call 1 and 0 of the 42 emergent cases emergent, against 7 for the base model and 28 to 33 for the three runs of GAO-Triage.

Full information. Given the complete post, the base model decides within its first two turns in 194 of 201 calls and still calls only 7 of the 42 emergent cases emergent, against 33, 28, and 28 for the three runs (all $p\leq5.7\times10^{-6}$); under-triage is 62 against 28, 25, and 25 calls (all $p\leq5.4\times10^{-9}$). The emergent gain therefore does not depend on eliciting facts. Overall agreement again does not rise (135, 123, and 118 against 131; p=0.71, 0.45, and 0.20): the runs ask their usual questions (10.1 to 10.4 nurse turns on average), receive no new information, and over-triage 37, 53, and 58 calls against 8. The matched permutation control answers routine in 198 of 201 calls, and the two disposition-permutation runs in 192 and 201; read post hoc, the matched control's next-step sentence still calls 17 of the 42 emergent cases emergent. Within each model, full information changes agreement little (the three runs 135, 123, and 118 against 138, 134, and 119 interactive, all $p\geq0.24$; base model 131 against 130, p=1.00).

Oxford-40. We ran the same complete-case mode on the 40 vignettes of Appendix D.8, each given as the caller's first message (Table 38). The three runs agree with 25, 26, and 24 labels; the third shares 38 of 40 verdicts with the precursor of Appendix D.8, which also agreed with 25. The base model agrees with 26 in the same mode (all $p\geq0.77$ against it). They err differently: every run calls all 23 emergent vignettes emergent, against 18 for the base model, but also 12 or 13 of the 14 urgent ones, against 6. The matched permutation control agrees with 4, answering routine to 38 vignettes. With the two-shot completion prompt of Appendix D.8, the base model reproduces its earlier 20/40 verdict for verdict in this environment, so the gap between 20 and 26 comes from the prompt. Asked for a verdict in a single turn, as the precursor was, the third run instead asks a question in 17 of 40 vignettes; this is why we use the complete-case mode.

Table 38: Oxford-40 exact match and per-class recall (23 emergent, 14 urgent, 3 routine vignettes); no call ended without a disposition. Rows marked * are from Appendix D.8; the others were run in the second environment.

Nurse and prompt	Exact	E	U	R
Base 9B, two-shot prompt (rerun)	20/40	9/23	10/14	1/3
Base 9B, complete-case mode	26/40	18/23	7/14	1/3
GAO-Triage third run, complete-case mode	25/40	23/23	1/14	1/3
GAO-Triage fourth run, complete-case mode	26/40	23/23	2/14	1/3
GAO-Triage fifth run, complete-case mode	24/40	23/23	0/14	1/3
Matched label permutation, complete-case mode	4/40	1/23	0/14	3/3
Trained precursor, single turn*	25/40	22/23	2/14	1/3
Zero-shot GPT-5.5*	33/40	22/23	10/14	1/3

Cost-sensitive comparison. Table 39 gives the paired cost differences behind Figure 3b as totals over the 201 calls. A call costs k if under-triaged and 1 if over-triaged, a call without a disposition is charged as routine, and intervals come from 10,000 paired bootstrap resamples of cases. In the interactive protocol, the intervals of four of the five runs lie below zero at k=2 and those of all five at k=3; with full information, two of three at k=2 and all three at k=3. Charging each tier of error instead (an emergent case called routine costs $2k$, a routine case called emergent 2) leaves every trained-versus-base k^* at most 1.12 and moves every trained-versus-base interval at k=2 below zero. The paired cost intervals of GPT-5.5 and the first run include zero at every k shown. The matched permutation control costs more than the third run at every k above 0.38, because it under-triages 61 more calls.

Table 39: Paired cost differences (run minus comparator, totals over 201 calls) with paired bootstrap 95% intervals. ΔU and ΔO are the differences in under- and over-triaged calls; k^* is the break-even ratio above which the run costs less ("$<x$": less only below x; "—": never less). The last column gives k^* when each tier of error is charged.

Comparison	ΔU	ΔO	k^*	k=1	k=2	k=3	k^*, tiers
First environment, interactive							
Run 1 vs base	-26	$+0$	0.00	-26 [-43, -10]	-52 [-79, -26]	-78 [-117, -40]	0.19
Run 2 vs base	-28	$+5$	0.18	-23 [-40, -6]	-51 [-78, -24]	-79 [-118, -40]	0.21
Run 1 vs GPT-5.5	$+2$	-1	<0.50	$+1$ [-15, 17]	$+3$ [-20, 25]	$+5$ [-26, 35]	—
Second environment, interactive							
Run 3 vs base	-33	$+25$	0.76	-8 [-26, 10]	-41 [-70, -11]	-74 [-116, -32]	0.76
Run 4 vs base	-34	$+30$	0.88	-4 [-22, 14]	-38 [-66, -11]	-72 [-111, -35]	0.75
Run 5 vs base	-27	$+38$	1.41	$+11$ [-7, 29]	-16 [-44, 11]	-43 [-81, -5]	1.00
Run 3 vs permutation	-61	$+23$	0.38	-38 [-56, -20]	-99 [-128, -70]	-160 [-202, -119]	0.18
Second environment, full information							
Run 3 vs base	-34	$+29$	0.85	-5 [-21, 11]	-39 [-64, -15]	-73 [-109, -39]	0.93
Run 4 vs base	-37	$+45$	1.22	$+8$ [-10, 26]	-29 [-58, -2]	-66 [-106, -29]	1.12
Run 5 vs base	-37	$+50$	1.35	$+13$ [-6, 31]	-24 [-52, 4]	-61 [-100, -23]	1.11

J Extended Related Work

This appendix extends the related work of Section 5.

Rule-derived supervision. Data programming replaces per-instance annotation with labeling functions whose noisy votes a label model aggregates (Ratner et al., 2016; 2017), and rule distillation moves logic rules into network weights so that the rules are not needed at test time (Hu et al., 2016). Other work derives training data from written principles or from model outputs (Bai et al., 2022; Wang et al., 2023; Taori et al., 2023), including medical agent trajectories (Yu et al., 2025). GAO assigns each synthetic dialogue its tier from the rows its profile cites, and the learner is an interactive policy that must elicit the facts the label depends on; we ask what such labels teach it.

Guidelines, triage, and clinical dialogue agents. DialGuide conditions a dialogue model on developer guidelines at inference (Gupta et al., 2023), and TriageAgent retrieves the Emergency Severity Index handbook to triage static emergency-department cases (Lu et al., 2024); both keep the guideline in the loop, whereas GAO-Triage uses no guideline or retrieval at inference. Most medical LLM evaluations use question answering or full-information vignettes (Nori et al., 2023; Singhal et al., 2023; Thirunavukarasu et al., 2023), including emergency and ophthalmic triage (Lee et al., 2025; Mittal & Aggarwal, 2026). Interactive settings withhold findings until asked (Schmidgall et al., 2024; Nori et al., 2025; Tu et al., 2024), and a deployed telephone agent limits its LLM to scripted preparation under nurse review (Kini et al., 2026).

Learning what to ask. Dialogue policies have long been trained against simulated users (Schatzmann et al., 2007; Kreyssig et al., 2018), diagnosis dialogue systems learn which symptoms to acquire (Wei et al., 2018; Xu et al., 2019), and recent work trains and evaluates clarifying questions, including in clinical reasoning (Li et al., 2024; Huang et al., 2025; Li et al., 2025). We separate what an agent asks from how it decides, and find that the rule-derived labels shape the second. Preference and policy-gradient methods optimize outcome signals directly (Ouyang et al., 2022; Rafailov et al., 2023; Shao et al., 2024); our outcome-rewarded GRPO run ends calls early (Section 6).

Evaluating decisions with asymmetric errors. When errors differ in cost, accuracy is the wrong summary: cost-sensitive learning prices each error type (Elkan, 2001), and cost curves compare classifiers over the range of cost ratios (Drummond & Holte, 2006). We apply this view to triage, where under-triage is the costlier error. Synthetic pipelines can also carry benchmark content into training (Xu et al., 2024; Zhou et al., 2023), LLM judges have known biases (Zheng et al., 2023), and a reference informed by the method under test is biased toward it (Whiting et al., 2003; 2011). Our reference has this dependence, so we rescore the main comparisons under alternative labelings,

and for historical runs whose verdicts were overwritten we report bounds (Heitjan & Rubin, 1991; Horowitz & Manski, 1995; Manski, 2003).

Reinforcement learning for clinical dialogue agents. Recent agents learn clinical inquiry and decision making with reinforcement learning: Doctor-R1 trains a doctor agent for strategic multi-turn inquiry with a two-tiered reward and an experience repository (Lai et al., 2026); ATPO casts multi-turn medical dialogue as a hierarchical MDP and allocates rollouts by turn-level uncertainty (Cao et al., 2026); a hypothesis-driven agent learns to request and interpret tests with supervised and reinforcement learning (Bani-Harouni et al., 2026); and MMedAgent-RL trains a triage doctor that routes cases to specialist agents (Xia et al., 2026). KnowGuard instead abstains until graph-retrieved evidence suffices (Dang et al., 2026). The reinforcement-learning methods optimize against case-level answers or rewards; GAO obtains every training label from a compiled guideline, and our analysis asks what such labels teach an interactive agent.