CLR-voyance: Reinforcing Open-Ended Reasoning for Inpatient Clinical Decision Support with Outcome-Aware Rubrics

Aishik Nagar[2*] **Arun-Kumar Kaliya-Perumal**[3] **Yu-Hsuan Han**[4]
Andrew Sheng-Han Huang[5,6†] **Kristen Kee**[7] **Yushi Cao**[1]
Yiming Chen[1] **Hongchao Jiang**[1]

[1]ASUS Intelligent Cloud Services (AICS) [2]Carnegie Mellon University
[3]Rehabilitation Research Institute of Singapore, Nanyang Technological University
[4]Department of Family Medicine, Taipei Veterans General Hospital
[5]School of Medicine, National Yang Ming Chiao Tung University [6]National Taiwan University
[7]Yong Loo Lin School of Medicine, National University of Singapore

Abstract

Inpatient clinical reasoning is a sequential decision under partial observability: the clinician sees the admission so far and must choose the next action whose downstream consequences are not yet visible. Existing clinical LLM evaluations and reinforcement learning (RL) rewards discard this structure in one of three ways. They reduce reasoning to closed-form retrieval, they expose the full clinical journey to the model, or they rely on LLM judges with no grounding in patient outcomes. Here, we introduce *CLR-voyance*, an end-to-end framework that reformulates inpatient reasoning as a Partially Observable Markov Decision Process (POMDP) and supervises it with rewards that are simultaneously *grounded in successful care trajectories* and *clinician-validated*. We instantiate the formulation as *CLR-POMDP* (Clinical Reasoning POMDP), which partitions successful patient journeys into a policy-visible *past* and an oracle-only *future*. Using the withheld future, an oracle LLM writes a case-specific query, a reference answer, and a **per-case rubric** that is verifiable from the past at grading time. To our knowledge, this is the first rubric conditioned on a real patient's downstream trajectory and used as both reward and metric for inpatient reasoning. These rubrics are used for both RL post-training and evaluation of models for inpatient clinical reasoning. *CLR-voyance* post-trains Qwen3-8B and MedGemma-4B with GRPO followed by model merging, yielding state-of-the-art inpatient clinical reasoning while retaining generalist capabilities. *CLR-voyance*-8B achieves **84.95% on *CLR-POMDP***, ahead of frontier medical reasoning models like GPT-5 (77.87%) and MedGemma-27B (66.70%) and has comparable or better performance on existing medical benchmarks. To ensure that our setting is clinically meaningful, we conduct a large-scale clinician alignment study, where physicians curate per-case rubrics, grade candidate responses against them, and provide blinded pairwise preferences of model reasoning. This study provides insights on clinical LLM-as-a-judge and clinical preference-model selection, which can inform the community at large. *CLR-voyance* has been deployed for **6+ months at a partner public hospital**, drafting thousands of reasoning-heavy inpatient notes and 1.03T+ tokens processed to date.

*Work done as part of ASUS AICS.
†Work conducted at National Yang Ming Chiao Tung University.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

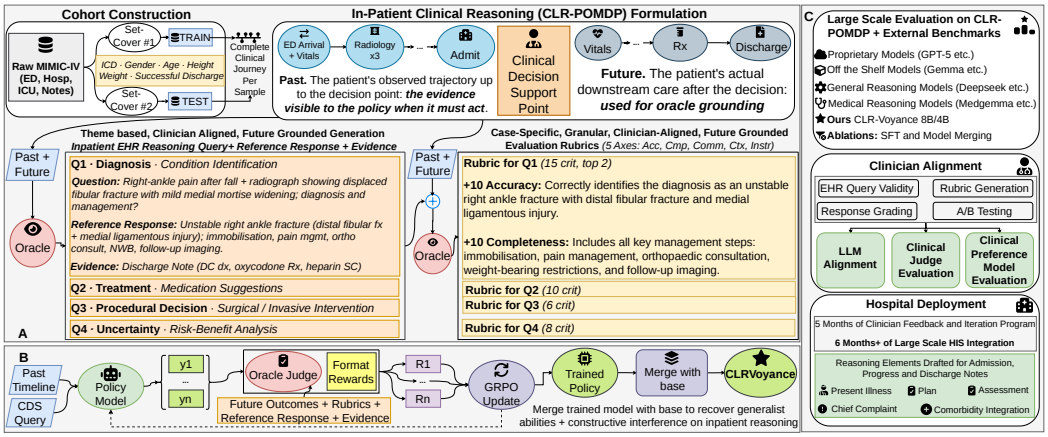

Figure 1: Overview of the *CLR-voyance* framework: **(A)** *CLR-POMDP* construction, **(B)** GRPO post-training and weight-space merging on the rubric reward, and **(C)** evaluation, two-cohort clinician validation, and hospital deployment.

1 Introduction

At the heart of inpatient medicine is a sequential decision under partial observability [Mehandru et al., 2025]. Standing at the bedside, a clinician sees the admission so far (triage, vitals, prior labs, imaging, the medications already given, the consults already placed) and must choose the next action whose downstream consequences they have not yet observed: which drug to start or hold, when to transfer. The information that would resolve the question is, by construction, in the future. The clinician acts under uncertainty, and the patient's trajectory afterwards is what tells us whether they acted well. The closer a large language model (LLM) gets to that bedside, the more its evaluations and reward signals must align with this process [Bedi et al., 2025]. Many everyday inpatient decisions cannot be posed without a past/future split. In a dosage adjustment ("does this patient's vancomycin need a renal-function-based dose change?") or a transfer-readiness decision, the past makes the question well-posed and the future supplies the verifiable answer, so benchmarks built on a single time-flattened document cannot supervise them without leaking the answer into the question.

Yet, existing clinical-LLM works usually do not align with this process. Exam-style benchmarks evaluate reasoning in closed-form formats that frontier models have largely saturated, and their scores remain only loosely predictive of real-world clinical performance [Zuo et al., 2025, Kim et al., 2025, Bedi et al., 2025, Wang et al., 2025b]. Longitudinal-EHR benchmarks have begun to push past the exam format, but they typically expose the entire admission and recover ground truth from the document that already records the answer [Kweon et al., 2024, Gao et al., 2023, Yang et al., 2026, Fleming et al., 2024, Liao et al., 2025]. Clinical reward signals present similar problems. Binary verifiers inspired by math and code work on MCQ style "verifiable medical problems" [Chen et al., 2024] but cannot grade open-ended management, while multi-objective rule-based clinical-RL rewards [Gu et al., 2025, Tziakouri and Menolascina, 2025, Zhang et al., 2025, Xu et al., 2025, Liao et al., 2025] ground in static references rather than in what a real care team did. General-domain rubric-as-reward work [Arora et al., 2025, Gunjal et al., 2025, Chen et al., 2025, Xu et al., 2026, Mu et al., 2024] replaces those rules with rubrics that are static, conversation-specific, or per-instance for general text, but never conditioned on a patient's downstream trajectory. Finally, most clinical-LLM studies stop at benchmark numbers [Chen et al., 2024, Gu et al., 2025, Zhang et al., 2025, Liao et al., 2025], clinician annotations are usually reported once as a grader–clinician correlation [Tu et al., 2024, Vedadi et al., 2025, Arora et al., 2025, Bedi et al., 2025], and the few hospital deployments, a pilot on eight patients [Masayoshi et al., 2025] and a retrieval and summarisation assistant [Griot et al., 2025], do not train on inpatient clinical reasoning. **Thus, no existing study jointly supports outcome-grounded inpatient reasoning at scale, with clinician-alignment at every step of the process, and real-world large scale evidence verified outside its training distribution**.

We introduce ***CLR-voyance***[3], an end-to-end framework that closes these gaps as outlined below. CLR stands for clinical reasoning, and the name plays on clairvoyance, since the oracle that writes each rubric sees the patient's future.

- **A real-EHR formulation for partially observable inpatient reasoning.** We introduce *CLR-POMDP*, a POMDP-style formulation built from MIMIC-IV admissions in which models observe only the patient event history before a sampled decision point, while evaluation is grounded in the successful future care actions taken for the patient. The construction preserves raw longitudinal EHR event streams at production-realistic context lengths ($\geq$60K tokens).
- **Trajectory-grounded judging with per-case rubrics.** We propose the *Grounded Judge*: an oracle LLM with access to the full admission trajectory which generates a clinically meaningful query, an event-cited reference answer, and a per-case multi-axis rubric conditioned on that patient's actual future. This rubric is then used to grade clinical reasoning responses into dense rewards, enabling open-ended clinical reasoning evaluation beyond static or outcome-agnostic rubrics. Rubric-as-reward methods already exist for general and health text [Gunjal et al., 2025, Arora et al., 2025]. Our contribution is a novel rubric formulation conditioned on the withheld future of a real patient trajectory, verifiable from the past at grading time, and used as both the RL reward and the evaluation metric for inpatient EHR reasoning.
- **Post-training small models for frontier-level inpatient reasoning.** We train Qwen3-8B and MedGemma-4B with GRPO on rubric-derived rewards, with weight-space merging to recover generalist capabilities. GRPO alone brings Qwen3-8B to parity with GPT-5 (77.99% vs. 77.87%), and merging the GRPO task vector back into the base raises it to 84.95%, ahead of GPT-5, MedGemma-27B, and open clinical-reasoning baselines. Controls show that only the GRPO task vector produces this gain (§4.2).
- **Clinician validation and large scale hospital integration.** Four board-certified physicians in spine/orthopaedic and general-medicine/obesity cohorts validate the pipeline, and their annotations support rubric and judge alignment and clinical LLM-as-a-judge and preference-model evaluations. *CLR-voyance* has been **deployed for over six months at a partner public hospital**, drafting the integrative reasoning elements of inpatient notes, with over 1.03 trillion tokens processed and **thousands of patient notes generated daily**.

2 Related Work

Clinical benchmark Longitudinal-EHR benchmarks cover discharge-summary QA, progress notes, structured snapshots, full-record instruction following, and code prediction [Kweon et al., 2024, Gao et al., 2023, Yang et al., 2026, Fleming et al., 2024, Ben Shoham and Rappoport, 2024, Wornow et al., 2023, Liao et al., 2025]. The closest formulation to our setting, ER-Reason [Mehandru et al., 2025], splits ER cases past/future across five decision points but exposes the full stay on two, and reports zero-shot frontier-model scores rather than training a model on its own task setting. Dialogue and clinical-pathway benchmarks [Nori et al., 2025, Tu et al., 2024, Vedadi et al., 2025, Li et al., 2025] take partial observability seriously but operate on patient-actor or online-dialogue data, not raw inpatient EHR. This line is overwhelmingly evaluation-only, and no partially observable inpatient-reasoning setting doubles as a training distribution.

Clinical reward-shaping HealthBench [Arora et al., 2025] uses rubric-based evaluation but targets free-form health *conversations* with little inpatient-EHR coverage. The "verifiability bottleneck" for clinical RL [Chen et al., 2024, Liu et al., 2025, Huang et al., 2025, Xu et al., 2026, Wang et al., 2025a] therefore remains: in clinical care almost every plausible action looks reasonable, and the one ground truth available for a single patient (the future care actions taken on that patient) has not, until now, been used as the reward signal. Outside medicine, rubrics now serve as per-instance contextual verifiers for software-engineering agents, grounded in repository state [Raghavendra et al., 2026], and, in contemporaneous work, as rewards for agent safety [Loye et al., 2026]. Our rubric shares this per-instance verification but grounds it in a patient's realised downstream trajectory and grounds with expert clinician validation.

[3]Appendix A contrasts *CLR-voyance* against 60+ prior works across eleven factors as a complete reference.

Practical evidence Clinician annotations in prior work [Zhou et al., 2025, Wu et al., 2025a, Shan et al., 2025, Zöller et al., 2024] are reported once as a grader–clinician correlation rather than reused as a training resource, and none target inpatient-EHR reasoning. Deployed clinical LLMs have been retrieval and summarisation tools [Griot et al., 2025, Fleming et al., 2024, Nagar et al., 2025] that condense the timeline into the descriptive elements of a note. In the largest reported deployment, an open-ended chatbot, decision support is the smallest use-case bucket at 11% [Griot et al., 2025], and no system yet supports day-to-day decisions under partial observability.

3 Methodology

Figure 1 summarises our entire methodology and we discuss it in detail below.

3.1 From EHRs to a Clinical POMDP: the *CLR-POMDP* formulation

Source and POMDP framing. We treat each admission as a POMDP $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{O}, \mathcal{T}, \mathcal{Z}, r)$ over latent patient states $\mathcal{S}$, clinical actions $\mathcal{A}$, EHR observations $\mathcal{O}$, transition kernel $\mathcal{T}$, observation-emission kernel $\mathcal{Z}$, and reward r (r_{rub} of §3.2). The clinician observes an event stream $o_{1:t} \in \mathcal{O}^t$ from a latent patient state $s_t \in \mathcal{S}$ and chooses the next action $a_t \in \mathcal{A}$ whose downstream consequences are not yet visible. The *CLR-POMDP* dataset is derived from MIMIC-IV [Johnson et al., 2023c, 2024, 2023a,b, Goldberger et al., 2000].We extract a unified, chronologically ordered timeline spanning the four MIMIC-IV modules (ED, hosp, ICU, notes) and present it to the policy verbatim in JSON-shaped event-level form (no pre-summarisation, no template), at a 65,536-token context window. The policy acts once per decision point. It reads $o_{1:t}$ and returns one next action with its reasoning, which the reward then scores against the withheld remainder. Partial observability is therefore a property of the task and of the reward, and the optimiser is single-turn response-level RL (§3.3) that learns neither a belief state nor a multi-turn plan.

Cohort and temporal split. We use a greedy set-cover cohort over (ICD code, sex, binned age, height, weight) to select training admissions, with a disjoint 20% random sample of the remainder forming the test split. The RL training set contains 5,099 admissions with four queries each (about 20.4K items). Each admission's chronologically sorted event sequence $e_{1:n}$ is split at $k \sim \mathcal{N}(n/2,\, n/6)$ clamped to $[1, n-1]$, drawing many splits per admission so the dataset stratifies across the stay (triage through discharge) without a hand-coded stage taxonomy. Two constraints support outcome-grounded supervision: (a) ICD-source events are removed from the past (the diagnostic label is what a reasoner should *recover*, not read off a row), and (b) admissions with re-admission or death within two years are excluded, so the future is a high-fidelity care trajectory rather than a record of downstream harm. This bounds every future to care whose end state was acceptable. It does not make each logged decision optimal, which is why the reward grades the reasoning rather than agreement with the logged action (§3.2).

Action spaces and training datum. Each split instantiates four reasoning action spaces that span the inpatient decision support under partial observability (Diagnosis, Treatment, Procedural, Uncertainty), one query per space. The resulting datum is $d = (\text{past},\, \text{meta},\, c,\, q,\, y^\star,\, \rho^\star,\, \text{src},\, R)$, where c is the action-space category, q the query, $y^\star$ and $\rho^\star$ the reference answer and reasoning, src the future events evidencing $y^\star$, and R the rubric (§3.2). Train and test splits are built admission-wise.

3.2 The Grounded Judge: outcome-aware queries, references, and per-case rubrics

The Grounded Judge is one oracle LLM, Qwen3-32B [Yang et al., 2025], used in three separate calls. Two authoring calls see the withheld future and write a query, a reference, and the rubric, and a third call grades responses. *Stage 1* (prompted with $(\text{past}, \text{future}, c)$) emits $(q, y^\star, \rho^\star, \text{src})$ under two hard constraints: q must be answerable from the past alone and $y^\star$ must be evidenced by at least one future event in src (source). *Stage 2* synthesises the rubric R conditioned on past, future, query, and reference. Each criterion is a self-contained pass/fail statement tagged with one of five clinician-centred axes (Accuracy, Completeness, ContextAwareness, CommunicationQuality, InstructionFollowing) and weighted in $[-10, +10]$, with positive points rewarding helpful behaviour and negative points penalising unsafe behaviour. These axes are designed to capture two complementary aspects of what matters to clinicians: *Accuracy*, *Completeness*, and *ContextAwareness* grade

Table 1: The four labelling schemes used in the paper, which are easy to conflate. Only the grading axes and action spaces appear as reported scores.

Scheme	Values	Used for
Grading axes (5)	Accuracy, Completeness, ContextAwareness, CommunicationQuality, InstructionFollowing	tags + per-axis scores
Action spaces (4)	Diagnosis, Treatment, Procedural, Uncertainty	one query each per split
Rubric themes (7)	HealthBench-style themes, e.g. Context Seeking (App. B.5)	rubric-authoring exemplars
Critical accuracy	Accuracy criteria with absolute weight ≥ 8 (App. B.10)	safety-weighted subset

what the candidate proposes, grounded in the patient's downstream events and to the actions a real care team actually took next, while *CommunicationQuality* and *InstructionFollowing* grade *how* that information is presented, reflecting both how clinicians consume model output and how they prefer the model's decision support to be integrated into hospital workflows. We forbid the rubric generation stage from generating criterion text that refer to future events. Thus, the future information is only used to ground the grading process with successful care actions, while each criterion itself can be verifiable from the past information and the generated response. Each criterion is a pass/fail statement about the reasoning (differential coverage, risk assessment, citing the patient's actual labs and medications, safety), so a clinically sound alternative to the logged action still satisfies most criteria. In clinician curation (§5) physicians marked 10.9% of oracle criteria not suitable, an upper bound on every rubric failure mode including future leakage, which Appendix E.4 documents as rare. We call R the *per-case rubric* throughout, and Table 1 separates the four labelling schemes the paper uses. Verbatim oracle prompts and the JSON schema are provided in Appendix B.4.

Rubric-based grading. A Grounded Judge call grades the policy's candidate $\hat{y}$ against R, seeing past, reference, source list, query, candidate, and the rubric with compact criterion identifiers, and returning per-criterion booleans. The scalar reward is

$$r_{\text{rub}}(\hat{y} \mid d) \;=\; \text{clip}_{[0,1]}\Big(\frac{\sum_{i:\, j_i=\text{true}} p_i}{\sum_{i:\, p_i>0} p_i}\Big),$$

where p_i is the signed weight of criterion i and j_i the grader's verdict. Normalising by positive weights only lets negative criteria drive the score downward without inflating the ceiling for compliant responses; the same rule is used both as the GRPO training signal and as the evaluation aggregator. The policy never receives the future, the reference, or the rubric. It sees only the scalar r_{rub}, a past-safe projection of the oracle's outcome-aware judgement during the update step.

3.3 Post-training and evaluation of *CLR-voyance*

Base Models. We post-train Qwen3-8B [Yang et al., 2025] (reasoning based language-only model) and MedGemma-4B [Sellergren et al., 2025] (domain-adapted vision-language model). The training loop is implemented in PyTorch [Paszke et al., 2019] with DeepSpeed ZeRO-3 [Rasley et al., 2020], vLLM rollouts [Kwon et al., 2023], and HuggingFace TRL [von Werra et al., 2020].

SFT and GRPO. For SFT we use *CLR-POMDP* with target $<\texttt{think}>\rho^\star<\texttt{/think}>\; y^\star$. GRPO [Shao et al., 2024] samples $G \in \{4, 8\}$ completions per prompt, scores each via the Grounded Judge, and uses the group-relative advantage as a clipped policy-gradient update with a KL penalty to a frozen reference. The canonical reward stack is $r_{\text{total}} = r_{\text{rub}} + r_{\text{format}} + r_{\text{tag}}$, where r_{rub} is the grounded rubric reward, r_{format} enforces the <think>/answer structure, and r_{tag} awards partial credit for well-formed tags.

Model merging. Domain-specialised RL is known to erode generalist capability via sparse, outlier-laden RL task vectors [Davari and Belilovsky, 2024, Deep et al., 2024, Yu et al., 2024], and weight-space merging has been used as a remedy for clinical SLMs to consolidate specialised expertise without losing generalist behaviour [Corbeil et al., 2025]. We merge the GRPO task vector back into the base, comparing two methods (DELLA [Deep et al., 2024] and Breadcrumbs [Davari and Belilovsky, 2024]), each optionally composed with Activation-Informed Merging (AIM) [Nobari et al., 2025].

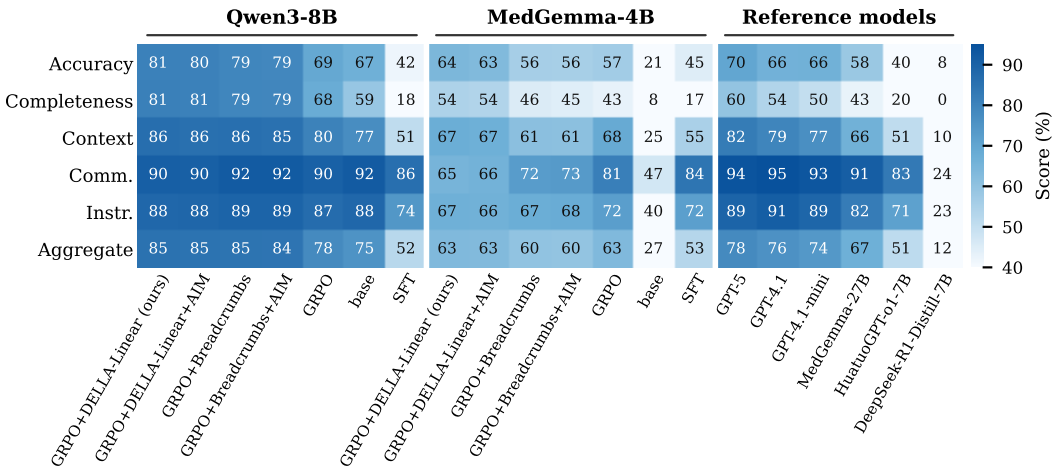

Figure 2: Per-axis *CLR-POMDP* scores (%), grouped by base family. Columns share their family's prefix, so "GRPO+DELLA-Linear (ours)" under Qwen3-8B is Qwen3-8B + GRPO + DELLA-Linear, i.e. *CLR-voyance*-8B.

Evaluation. Every model is evaluated on *CLR-POMDP* with the same response-generation prompt (Appendix B.4), graded by the Grounded Judge using r_{rub}, with results reported per-axis, per-action-space, and as an aggregate, and on external benchmarks (§4.1).

4 Results

We evaluate *CLR-voyance*-8B on the held-out *CLR-POMDP* test split, comprising 7,927 (admission $\times$ query) instances from about 2,000 held-out admissions, disjoint from training by admission. Aggregate rubric scores for the full panel are summarised in Figure 2. *CLR-voyance*-8B (Qwen3-GRPO with DELLA-Linear merging) attains an aggregate score of **84.95%**. For reference, GPT-5 [Singh et al., 2025] reaches 77.87%, Qwen3-8B GRPO 77.99%, GPT-4.1 75.77% [OpenAI, 2025], the Qwen3-8B base 75.32% [Yang et al., 2025], GPT-4.1-mini 73.81%, MedGemma-27B 66.70% [Sellergren et al., 2025], and HuatuoGPT-o1-7B [Chen et al., 2024] 51.27%. The four *CLR-voyance* merge variants (DELLA-Linear, DELLA-Linear+AIM, Breadcrumbs, Breadcrumbs+AIM) score within 0.7 percentage points (pp) of one another (84.29–84.95%), and each is at least 6.3 pp above the strongest non-merge baseline. A paired Wilcoxon signed-rank test[4] on matched per-admission scores yields $p < 10^{-300}$ for *CLR-voyance*-8B exceeding GPT-5.

Per-axis decomposition. The improvement over GPT-5 is concentrated on the three clinical-content axes, Completeness ($+21.37$ pp, 80.98 vs. 59.61), Accuracy ($+11.08$ pp), and ContextAwareness ($+4.05$ pp), while GPT-5 keeps a small advantage on CommunicationQuality (-3.38 pp) and InstructionFollowing (-0.98 pp). This is the asymmetry the per-case rubric design predicts (§3.2), since the content axes are graded against the patient's downstream events and the conversational axes grade style on which frontier models are already strong [Wang et al., 2025b, Kim et al., 2025], the same axis-level pattern Arora et al. [2025] report for HealthBench.

Action-space decomposition. Against GPT-5, *CLR-voyance*-8B is higher by 8.0 pp on Procedural, 7.6 pp on Uncertainty, 6.5 pp on Treatment, and 2.7 pp on Diagnosis. The smallest gain falls on Diagnosis, where frontier models have broad exposure to labelled diagnostic tasks [Wang et al., 2025b], and the largest on the two buckets whose correct answer depends most directly on what the care team did next [Mehandru et al., 2025].

Per-instance analysis. Against GPT-5 the two models tie on 47–57% of instances per action-space bucket, and on the rest *CLR-voyance* wins 3 to 5 times more often, up to $4.5\times$ on Responding under Uncertainty (668 wins to 148). *CLR-voyance* satisfies 73.6% of positive rubric criteria against

[4] A non-parametric paired test whose null is a zero median of per-admission score differences.

68.2% for GPT-5 and triggers the rubric's negative-point safety flags at 1.30% vs. 1.79%, a 27% relative reduction matching the role of rule-based safety rewards in RLHF [Mu et al., 2024]. The Completeness gain coincides with longer outputs ($\sim$8.3K vs. $\sim$3.0K characters), but the advantage holds when length is held fixed (§4.2). Four representative cases with full rubrics and per-criterion breakdowns are in Appendix G.

Effect of weight-space merging. On Qwen3-8B, GRPO alone reaches 77.99% aggregate (Qwen3-8B + GRPO, level with GPT-5), and DELLA-Linear merging [Deep et al., 2024] with the base then adds a further 6.96 pp to take *CLR-voyance*-8B (Qwen3-8B + GRPO + DELLA-Linear) to 84.95%, broadly distributed across the three clinical-content axes. The merge is a stage of the recipe, not a separate source of performance, since merging the base with itself returns the base and §4.2 shows that only the GRPO task vector improves it. It is also needed in practice. Over six months of deployment the GRPO-only checkpoint terminated cleanly on 98.89% of completions, clinicians rejected every one of the remaining degenerate over-long outputs, and the merged model terminated cleanly on 100%. The four merge variants (DELLA-Linear, DELLA-Linear+AIM [Nobari et al., 2025], Breadcrumbs [Davari and Belilovsky, 2024], Breadcrumbs+AIM) lie within 0.7 pp of one another, and their per-axis numbers and the 4B analysis are in Appendix B.12.

GRPO versus SFT on the same data. SFT on the oracle reference, run as an independent ablation from the same base, regresses Qwen3-8B from 75.32% to 52.36% by contracting the policy onto a single short answer, and raises the scaffolding-free MedGemma-4B from 27.01% to 52.89%, still well below GRPO from the same base (63.40%). Appendix B.12 gives the per-axis breakdown and failure modes.

The 4B family. MedGemma-4B rises from 27.01% to 63.40% under GRPO (MedGemma-4B + GRPO). The DELLA-Linear merge (MedGemma-4B + GRPO + DELLA-Linear) leaves the aggregate flat at 63.35% but moves points from the conversational axes to the content axes (Accuracy $+7.48$, Completeness $+11.32$, CommunicationQuality -16.22, InstructionFollowing -5.60 pp), which Appendix C.1 relates to this instruction-tuned base starting without reasoning training.

Reasoning-distilled baselines. Two reasoning-distilled checkpoints of the same size and family transfer poorly. DeepSeek-R1-Distill-Qwen-7B reaches 12.00%, emitting long traces that fail the positive Accuracy and Completeness criteria, and DeepSeek-R1-0528-Qwen3-8B produces corrupted outputs on most cases. This suggests that post-training for verifiable mathematics and code [Shao et al., 2024, Gunjal et al., 2025] does not transfer to open-ended inpatient reasoning, which motivates a reward designed for this setting. Since these checkpoints never saw our data or reward, we isolate the reward's contribution separately in §4.2.

4.1 External medical benchmarks

We evaluate *CLR-voyance* on four external benchmarks spanning clinical-knowledge MCQ (**MedMCQA**), open-set diagnostic naming (**DDXPlus**), numeric calculator answers (**MedCalc-Bench**) and MIMIC-IV instruction-following (**Mimic-Instr** [Wu et al., 2024]). We additionally evaluate on **HealthBench** [Arora et al., 2025], on which the *CLR-voyance*-8B improves the Qwen3-8B base from 0.364 to 0.415 ($+0.051$) and the *CLR-voyance*-4B merge improves the MedGemma-4B from 0.256 to 0.328 ($+0.072$).

CLR-voyance-8B improves on its base model on every benchmark in Table 2. *CLR-voyance*-4B improves on three of four and loses only on DDXPlus. The largest gain is on DDXPlus, the benchmark closest to open-set clinical reasoning. There *CLR-voyance*-8B reaches 48.5%, a $+5.0$ pp gain over its base. It is the only model in our 8B class to outperform GPT-5 (46.2%) and HuatuoGPT-o1 (46.0%).

The GRPO-only checkpoint shows what each stage contributes. GRPO alone produces the full DDXPlus gain, reaching 48.5%. On MedMCQA, which tests knowledge recall, GRPO alone stays at the base score of 65.5%. Merging the task vector back into the base model keeps the DDXPlus gain and adds $+1.0$ pp on MedMCQA. GRPO therefore drives the reasoning gains, and the merge keeps them while also improving recall. This is why we evaluate both stages of post-training. On the closed-form benchmarks, the frontier models still score highest.

Table 2: External benchmarks: accuracy %, n=200. **Bold** columns are the *CLR-voyance* headline checkpoints. Column groups: *F*=frontier API, *M/O*=medical / open-source, *8B*=Qwen3-8B family, *4B*=MedGemma-4B-IT family (Δ=DELLA-Linear, Δ +A=DELLA-Linear+AIM, B=Breadcrumbs, B+A=Breadcrumbs+AIM).

	Frontier API			**Medical / open-source**					**8B family**		**4B family** (MedGemma-4B-IT base + merges)				
Benchmark	***GPT-5***	***GPT-4.1***	***GPT-4.1-mini***	***MedGemma-27B***	***DS-R1-0528***	***DS-R1-Distill***	***Gemma-3-4B***	***HuatuoGPT-o1***	***Qwen3-8B base***	***Qwen3-Δ (ours)***	***MG-4B base***	***MG-Δ (ours)***	***MG-Δ+A***	***MG-B***	***MG-B+A***
MedCalc	81.4	72.0	67.5	51.9	31.6	8.6	7.3	27.0	41.5	**46.0**	10.7	**15.0**	15.5	21.0	18.5
MedMCQA	88.9	83.9	78.4	33.0	36.0	33.5	44.5	64.5	66.0	**67.0**	54.5	**58.5**	58.5	61.0	58.5
DDXPlus	46.2	43.0	42.0	8.0	18.5	15.0	36.0	46.0	43.5	**48.5**	37.5	**27.5**	26.0	25.5	24.5
Mimic-Instr	84.5	65.5	63.7	62.3	60.4	16.8	42.7	55.1	54.5	**58.9**	19.3	19.5	23.2	19.2	19.5

Table 3: Ablations on Qwen3-8B unless noted. **(a)** Merge isolation, varying only the task vector merged into the base at a fixed DELLA-Linear configuration (358 paired items). **(b)** Reward components on MedGemma-4B (mean per-item score on the 2,000-admission test sample). **(c)** Rubric construction on a fixed 250-admission slice, scored by agreement with blinded physician preferences (422 decided dyads) and by our margin over GPT-5 on 974 items. **(d)** Length-matched comparison with GPT-5 and a cross-family re-grade of the full test set with MedGemma-27B.

(a) Task vector merged into the base

Task vector	Score	Δ vs. base
GRPO (ours)	0.763	$+0.081$ [$+0.047$, $+0.115$]
SFT	0.108	-0.574 [-0.616, -0.531]
Random, norm-matched	0.232	-0.450 [-0.500, -0.400]
None (base)	0.682	–

(b) GRPO reward components (MedGemma-4B)

Configuration	Score
Base, no GRPO	22.9
GRPO, format + tag only (rubric removed)	27.5
GRPO, full reward with per-case rubric	53.6

(c) Rubric construction

Rubric	Physician agr.	Ours − GPT-5
Future-cond. (ours)	**69.4%**	$+0.044$ [$+0.025$, $+0.063$]
Past-only	66.9%	$+0.153$ [$+0.129$, $+0.177$]
Static generic	63.8%	-0.059 [-0.088, -0.031]

(d) Length and judge robustness

Setting	Ours	GPT-5	Base
Matched length (±25%, n=454)	0.775	0.752	–
MedGemma-27B judge (full test)	0.691	0.672	0.598

Together with the results on *CLR-POMDP* (§4), the HealthBench improvements, and clinician validation (§5), these results show that *CLR-voyance* improves performance across four distinct evaluation settings.

4.2 What drives the gain? Ablations

The gain could come from the merge, from the format and tag terms, from rubric choices other than future conditioning, from length, or from a Qwen-family judge. Table 3 tests each, with full protocols in Appendix C. Panel (d) checks length and judge effects (Appendices C.4–C.5).

Is the merge a free-standing booster? No. Only the GRPO task vector improves the base (Table 3a), while a random vector with the same magnitudes and the SFT vector fall far below it. The merge therefore transmits the structure the RL update wrote into a small, high-magnitude subset of weights [Yuan et al., 2026], which filtered merges keep [Deep et al., 2024, Davari and Belilovsky, 2024, Yu et al., 2024]. A merged model exceeding both parents on some axes, as on Qwen3-8B Accuracy, is a documented property of merging [Lu et al., 2025, Akiba et al., 2025].

Does the per-case rubric drive GRPO, or the format and tag terms? The rubric does. On MedGemma-4B, the full reward gains 30.7 points over the base model, but removing the rubric reward recovers only 4.6 of those points (Table 3b). The training dynamics explain why. GRPO learns only from terms that differ across a prompt's rollouts, since a term that is identical across the group gives no group-relative advantage. On Qwen3-8B, the format term never varies within a group, and the tag term saturates by step 148 of 3,218. The rubric is therefore the only source of advantage for over 95% of training.

Does conditioning on the future improve the rubric? We compare our future-conditioned rubric against two alternatives. The first is a past-only rubric, generated the same way but with the future withheld. The second is a fixed rubric that is the same for every patient. Each construction is lenient to a different degree, so raw scores cannot be compared directly. Instead, we follow the standard practice for validating reward models and judges [Lambert et al., 2024, Tan et al., 2024] and measure agreement with physicians. Physicians performed blinded comparison of pairs of model responses, and we measure how often each grading approach prefers the same response as physicians. The future-conditioned rubric agrees with physicians most often, both overall and for every model pair. The past-only rubric fails because its criteria are generic coverage checks. They reward a plausible wrong answer almost as much as a correct one, and this inflates our model's margin over GPT-5 to three times its size under the future-conditioned rubric. The fixed rubric does worse still and ranks the untrained base model first.

5 Clinician Validation

We performed a clinician validation study with **four board-certified physicians across two specialty cohorts**, comprising a *orthopaedic* cohort (two orthopaedic surgeons, senior+junior, using the keyword *spine* to filter cases) and a *general-medicine* cohort (two family-medicine clinicians, senior+junior, using keyword *obesity*) in three rounds, each with a distinct goal:

- **R1. Blinded A/B preference.** Each cohort compared *CLR-voyance*-8B side by side with GPT-5, HuatuoGPT-o1-7B, and the Qwen3-8B base, which also familiarised raters with the task before rubric work.
- **R2. Per-case rubric curation.** Raters marked each oracle criterion *Suitable* or *Not Suitable*, edited, added, removed, and reordered criteria, aligning the reward and metric with the signals clinicians use.
- **R3. Rubric-based grading.** Raters graded the *CLR-voyance*-8B response against the curated rubric, which aligns the oracle's grading with clinicians.

  Query validation. In every round raters could mark an oracle query invalid, giving a reason.

Across all rounds, query validation flagged almost nothing. Raters marked none of the 230 rubric cases in R2 and R3 as clinically invalid. Among the $1{,}000$ R1 comparisons, they flagged only one as asking something irrelevant for its patient, about 0.1% of all annotations. This directly supports our task setting. Sampling a decision point from a patient's past reliably produces questions that clinicians recognise as well-posed. Beyond validation, the annotations serve three purposes (Appendix D.5). They provide in-context examples for the rubric-author and grading-judge prompts. They let us measure how well existing **clinical LLM judges** and **clinical preference models** agree with clinicians. They also form the blinded head-to-head reported here. Full details are in Appendix D.

We report three agreement statistics, to measure different behaviours. In R2, physicians agree with each other on rubric criteria at κ=0.702. In R3, our grounded judge agrees with physicians on criterion verdicts at κ=0.42. In R1, physicians agree with each other on pairwise preferences at only κ=0.034 and 0.051. This low value reflects how lopsided the preferences were, not disagreement between raters. When nearly every rater picks the same answer, the agreement expected by chance is already high, and kappa drops toward zero. Finally, Qwen3-32B serves as both our oracle and one of the judges we evaluate. It agrees with physicians about as well as the other judges do, so our grading signal is not tied to the oracle.

R1: Blinded preference and clinical preference-model alignment. *CLR-voyance*-8B is preferred over every baseline on aggregate, decisively in the spine cohort ($p < 10^{-15}$ on each pair) and clearly on the HuatuoGPT-o1-7B and Qwen3-8B-base obesity pairs ($p < 10^{-4}$). Per-pair win rates with Wilson 95% CIs are in Table 4 (left), and Bradley–Terry rank intervals plus the position-bias diagnostic are in Appendix D.7. Scoring the same R1 dyads with five off-the-shelf judges against the two-clinician unanimous consensus (Table 4, right) provides a noteworthy finding: no single judge is uniformly best across cohorts, and GPT-5 is the worst-aligned in both with a 100% self-preference rate on the GPT-5 vs. *CLR-voyance* pair. This aligns with the self-preference behaviour documented for general-domain LLM-as-a-judge Panickssery et al. [2024].

Table 4: **R1 blinded A/B preference** (*left*) and **R1 LLM-as-a-judge alignment** (*right*). Best per cohort in **bold**.

R1 *CLR-voyance*-8B vs.	Spine Win % [95% CI]	Obesity Win % [95% CI]
GPT-5	**85.6** [78–91]	48.0 [41–55]
HuatuoGPT-o1-7B	**94.2** [88–97]	**82.2** [76–87]
Qwen3-8B (base)	**89.5** [82–94]	**62.7** [56–69]

Preference model	Spine 3-way / κ	Obesity 3-way / κ
Qwen3-32B	64.9 / 0.05	**85.7** / **0.64**
MedGemma-27B-text-it	**78.4** / -0.03	85.0 / 0.53
GPT-4.1	59.5 / 0.04	84.2 / 0.62
GPT-4.1-mini	55.4 / 0.03	78.2 / 0.51
GPT-5	43.2 / 0.02	69.2 / 0.40

R2 + R3: rubric curation and judge alignment. Across n=230 cases (101 patients, 2,496 rater–criterion pairs), the spine cohort edits the AI-generated rubric aggressively while the obesity cohort accepts almost all criteria as-is (Table 5, left). Both cohorts pervasively reorder criteria (49 of 50 sampled cases lead with Accuracy + Completeness) but concentrate every edit on the clinical-content axes and never touch the conversational axes. The Grounded Judge (Qwen3-32B) tracks the pooled clinician verdict at moderate agreement (κ=0.42 on the Landis–Koch scale [Landis and Koch, 1977]). Scored against the same R3 verdicts (Table 5, right), the two open-weights judges match or exceed larger GPT models and 5-shot ICL produces only marginal lift. This aligns with the R1 finding that judge$\times$cohort setting, not the judge model size, is the load-bearing axis. Per-cohort and per-rater breakdowns, the comparison with the static-jury based on MedHELM design [Bedi et al., 2025], and the 387-rationale failure-mode analysis are in Appendices D.6, E, and E.6.

Table 5: **R2 rubric quality** (*left*), **R3 Grounded-Judge–clinician alignment** (*centre*), and **R3 off-the-shelf LLM-as-a-judge alignment** (*right*). GJ = Grounded Judge; IRR = inter-rater. Best per column in **bold**.

Metric	Sp	Ob	Ag
n criteria	1,136	1,360	2,496
relevance rate	83.1%	94.0%	89.1%
modif. rate	3.4%	0.0%	1.5%
addition rate	2.7%	0.0%	1.2%
IRR κ	0.749	0.658	0.702

Metric	Sp	Ob	Ag
GJ Acc.	77.6%	75.9%	76.7%
GJ κ	0.439	0.402	0.419
Score MAE	0.124	0.192	0.160
Score r	0.571	0.434	0.478

Judge	Mode	Acc.	κ
Qwen3-32B	zero	76.7	**0.419**
Qwen3-32B	ICL	76.7	0.421
MedGemma-27B	zero	77.3	0.421
MedGemma-27B	ICL	**77.7**	0.399
GPT-5	zero	74.0	0.420
GPT-4.1	zero	76.9	0.414

6 Hospital Deployment and Scaling

CLR-voyance reached production through a three-stage programme at a partner public hospital. Over five months and three feedback rounds, models trained with SFT+DPO Rafailov et al. [2023] on an internal note corpus captured drafting style but underperformed on reasoning-heavy note elements (*Impressions*, *Assessment*, *Review of Systems*, *Plan*). *CLR-voyance* was trained for that gap and deployed network-wide after a positive clinician pilot. Clinicians accept or edit each draft before it is saved. On-premise concurrency, latency, and reliability requirements make the 8B model **the largest the system can support**, which sets our model scale. *CLR-voyance* has been in production for **over six months, processing over 1.03 trillion tokens across thousands of patient notes** (Appendix F).

7 Conclusion

We model inpatient clinical reasoning from partially observable patient information using LLMs. This is made possible by a novel approach to split each admission at a decision point, which allows using the past trajectory of the patient to generate well-defined questions and using the care that followed to generate grounded evaluation and reward rubrics about what an ideal answer should contain. This allows using routine records for reinforcing open-ended decisions that have no single verifiable answer. We train models at different scales and architectures using GRPO in this setting, followed by model merging to retain generalist abilities from base models. The result is *CLR-voyance*-8B, a model that runs under the compute constraints of a real-world hospital, outperforms every frontier model we evaluated on *CLR-POMDP*, is preferred by blinded physicians, and has been in routine inpatient use for six months. Our approach paves the way for scalable reinforcement learning of expert-aligned clinical reasoning in LLMs.

References

Takuya Akiba, Makoto Shing, Yujin Tang, Qi Sun, and David Ha. Evolutionary optimization of model merging recipes. *Nature Machine Intelligence*, 7:195–204, 2025.

Rahul K Arora, Jason Wei, Rebecca Soskin Hicks, Preston Bowman, Joaquin Quiñonero-Candela, Foivos Tsimpourlas, Michael Sharman, Meghan Shah, Andrea Vallone, Alex Beutel, et al. Healthbench: Evaluating large language models towards improved human health. *arXiv preprint arXiv:2505.08775*, 2025.

Ananth Balashankar, Ziteng Sun, Jonathan Berant, Jacob Eisenstein, Michael Collins, Adrian Hutter, Jong Lee, Chirag Nagpal, Flavien Prost, Aradhana Sinha, et al. Infalign: Inference-aware language model alignment. *arXiv preprint arXiv:2412.19792*, 2024.

Suhana Bedi, Hejie Cui, Miguel Fuentes, Alyssa Unell, Michael Wornow, Juan M Banda, Nikesh Kotecha, Timothy Keyes, Yifan Mai, Mert Oez, et al. Medhelm: Holistic evaluation of large language models for medical tasks. *arXiv preprint arXiv:2505.23802*, 2025.

Ofir Ben Shoham and Nadav Rappoport. Cpllm: Clinical prediction with large language models. *PLOS Digital Health*, 3(12):e0000680, 2024.

Junying Chen, Zhenyang Cai, Ke Ji, Xidong Wang, Wanlong Liu, Rongsheng Wang, Jianye Hou, and Benyou Wang. Huatuogpt-o1, towards medical complex reasoning with llms. *arXiv preprint arXiv:2412.18925*, 2024.

Xiusi Chen, Gaotang Li, Ziqi Wang, Bowen Jin, Cheng Qian, Yu Wang, Hongru Wang, Yu Zhang, Denghui Zhang, Tong Zhang, et al. Rm-r1: Reward modeling as reasoning. *arXiv preprint arXiv:2505.02387*, 2025.

Jean-Philippe Corbeil, Amin Dada, Jean-Michel Attendu, Asma Ben Abacha, Alessandro Sordoni, Lucas Caccia, Francois Beaulieu, Thomas Lin, Jens Kleesiek, and Paul Vozila. A modular approach for clinical SLMs driven by synthetic data with pre-instruction tuning, model merging, and clinical-tasks alignment. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 19352–19374, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.950. URL `https://aclanthology.org/2025.acl-long.950/`.

MohammadReza Davari and Eugene Belilovsky. Model breadcrumbs: Scaling multi-task model merging with sparse masks. In *European Conference on Computer Vision (ECCV)*, 2024. arXiv:2312.06795.

Pala Tej Deep, Rishabh Bhardwaj, and Soujanya Poria. DELLA-Merging: Reducing interference in model merging through magnitude-based sampling. *arXiv preprint arXiv:2406.11617*, 2024.

Chao Ding, Mouxiao Bian, Pengcheng Chen, Hongliang Zhang, Tianbin Li, Lihao Liu, Jiayuan Chen, Zhuoran Li, Yabei Zhong, Yongqi Liu, et al. Building a human-verified clinical reasoning dataset via a human llm hybrid pipeline for trustworthy medical ai. *arXiv preprint arXiv:2505.06912*, 2025.

Radhika Dua, Young Joon, Siddhant Dogra, Daniel Freedman, Diana Ruan, Motaz Nashawaty, Danielle Rigau, Daniel Alexander Alber, Kang Zhang, Kyunghyun Cho, et al. Clinically grounded agent-based report evaluation: An interpretable metric for radiology report generation. *arXiv preprint arXiv:2508.02808*, 2025.

Scott L Fleming, Alejandro Lozano, William J Haberkorn, Jenelle A Jindal, Eduardo Reis, Rahul Thapa, Louis Blankemeier, Julian Z Genkins, Ethan Steinberg, Ashwin Nayak, et al. Medalign: A clinician-generated dataset for instruction following with electronic medical records. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 22021–22030, 2024.

Yanjun Gao, Dmitriy Dligach, Timothy Miller, John Caskey, Brihat Sharma, Matthew M Churpek, and Majid Afshar. Dr. bench: Diagnostic reasoning benchmark for clinical natural language processing. *Journal of biomedical informatics*, 138:104286, 2023.

A. L. Goldberger, L. A. N. Amaral, L. Glass, J. M. Hausdorff, P. Ch. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, and H. E. Stanley. PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. *Circulation*, 101(23): e215–e220, 2000. doi: 10.1161/01.CIR.101.23.e215.

Maxime Griot, Jean Vanderdonckt, and Demet Yuksel. Implementation of large language models in electronic health records. *PLOS Digital Health*, 4(12):e0001141, 2025. doi: 10.1371/journal.pdig.0001141.

Boyang Gu, Hongjian Zhou, Bradley Max Segal, Jinge Wu, Zeyu Cao, Hantao Zhong, Lei Clifton, Fenglin Liu, and David A Clifton. Clinical-r1: Empowering large language models for faithful and comprehensive reasoning with clinical objective relative policy optimization. *arXiv preprint arXiv:2512.00601*, 2025.

Anisha Gunjal, Anthony Wang, Elaine Lau, Vaskar Nath, Yunzhong He, Bing Liu, and Sean Hendryx. Rubrics as rewards: Reinforcement learning beyond verifiable domains. *arXiv preprint arXiv:2507.17746*, 2025.

Yun He, Wenzhe Li, Hejia Zhang, Songlin Li, Karishma Mandyam, Sopan Khosla, Yuanhao Xiong, Nanshu Wang, Xiaoliang Peng, Beibin Li, et al. Advancedif: Rubric-based benchmarking and reinforcement learning for advancing llm instruction following. *arXiv preprint arXiv:2511.10507*, 2025.

Zenan Huang, Yihong Zhuang, Guoshan Lu, Zeyu Qin, Haokai Xu, Tianyu Zhao, Ru Peng, Jiaqi Hu, Zhanming Shen, Xiaomeng Hu, et al. Reinforcement learning with rubric anchors. *arXiv preprint arXiv:2508.12790*, 2025.

Pengcheng Jiang, Cao Xiao, Adam Cross, and Jimeng Sun. Graphcare: Enhancing healthcare predictions with personalized knowledge graphs. *arXiv preprint arXiv:2305.12788*, 2023.

Alistair Johnson, Lucas Bulgarelli, Tom Pollard, Leo Anthony Celi, Roger Mark, and Steven Horng. MIMIC-IV-ED (version 2.2). PhysioNet, 2023a.

Alistair Johnson, Tom Pollard, Steven Horng, Leo Anthony Celi, and Roger Mark. MIMIC-IV-Note: Deidentified free-text clinical notes (version 2.2). PhysioNet, 2023b.

Alistair Johnson, Lucas Bulgarelli, Tom Pollard, Steven Horng, Leo Anthony Celi, and Roger Mark. MIMIC-IV (version 3.0). PhysioNet, 2024.

Alistair E. W. Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J. Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, Li-wei H. Lehman, Leo Anthony Celi, and Roger G. Mark. MIMIC-IV, a freely accessible electronic health record dataset. *Scientific Data*, 10(1):1, 2023c. doi: 10.1038/s41597-022-01899-x.

Jonathan Kim, Anna Podlasek, Kie Shidara, Feng Liu, Ahmed Alaa, and Danilo Bernardo. Limitations of large language models in clinical problem-solving arising from inflexible reasoning. *Scientific reports*, 15(1):39426, 2025.

Sunjun Kweon, Jiyoun Kim, Heeyoung Kwak, Dongchul Cha, Hangyul Yoon, Kwanghyun Kim, Jeewon Yang, Seunghyun Won, and Edward Choi. Ehrnoteqa: An llm benchmark for real-world clinical practice using discharge summaries. *Advances in Neural Information Processing Systems*, 37:124575–124611, 2024.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with PagedAttention. In *Proceedings of the 29th Symposium on Operating Systems Principles (SOSP)*, pages 611–626, 2023. doi: 10.1145/3600006.3613165.

Nathan Lambert, Valentina Pyatkin, Jacob Morrison, LJ Miranda, Bill Yuchen Lin, Khyathi Chandu, Nouha Dziri, Sachin Kumar, Tom Zick, Yejin Choi, Noah A. Smith, and Hannaneh Hajishirzi. RewardBench: Evaluating reward models for language modeling. *arXiv preprint arXiv:2403.13787*, 2024.

J. Richard Landis and Gary G. Koch. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174, 1977.

Shuyue Stella Li, Jimin Mun, Faeze Brahman, Jonathan S Ilgen, Yulia Tsvetkov, and Maarten Sap. Aligning llms to ask good questions a case study in clinical reasoning. *arXiv e-prints*, pages arXiv–2502, 2025.

Yusheng Liao, Chaoyi Wu, Junwei Liu, Shuyang Jiang, Pengcheng Qiu, Haowen Wang, Yun Yue, Shuai Zhen, Jian Wang, Qianrui Fan, et al. Ehr-r1: A reasoning-enhanced foundational language model for electronic health record analysis. *arXiv preprint arXiv:2510.25628*, 2025.

Che Liu, Haozhe Wang, Jiazhen Pan, Zhongwei Wan, Yong Dai, Fangzhen Lin, Wenjia Bai, Daniel Rueckert, and Rossella Arcucci. Beyond distillation: Pushing the limits of medical llm reasoning with minimalist rule-based rl. *arXiv preprint arXiv:2505.17952*, 2025.

Xian Qi Loye, Qinglin Su, Zhexin Zhang, Shiyao Cui, Qi Zhu, Fei Mi, Hongning Wang, and Minlie Huang. RUBAS: Rubric-based reinforcement learning for agent safety. *arXiv preprint arXiv:2606.04051*, 2026.

Wei Lu, Rachel K. Luu, and Markus J. Buehler. Fine-tuning large language models for domain adaptation: Exploration of training strategies, scaling, model merging and synergistic capabilities. *npj Computational Materials*, 2025. arXiv:2409.03444.

Neil Mallinar, A Ali Heydari, Xin Liu, Anthony Z Faranesh, Brent Winslow, Nova Hammerquist, Benjamin Graef, Cathy Speed, Mark Malhotra, Shwetak Patel, et al. A scalable framework for evaluating health language models. *npj Digital Medicine*, 2026.

Kanato Masayoshi, Masahiro Hashimoto, Ryoichi Yokoyama, Naoki Toda, Yoshifumi Uwamino, Shogo Fukuda, Ho Namkoong, and Masahiro Jinzaki. Ehr-mcp: Real-world evaluation of clinical information retrieval by large language models via model context protocol. *arXiv preprint arXiv:2509.15957*, 2025.

Nikita Mehandru, Niloufar Golchini, David Bamman, Travis Zack, Melanie F Molina, and Ahmed Alaa. Er-reason: A benchmark dataset for llm-based clinical reasoning in the emergency room. *arXiv preprint arXiv:2505.22919*, 2025.

Tong Mu, Alec Helyar, Johannes Heidecke, Joshua Achiam, Andrea Vallone, Ian Kivlichan, Molly Lin, Alex Beutel, John Schulman, and Lilian Weng. Rule based rewards for language model safety. *Advances in Neural Information Processing Systems*, 37:108877–108901, 2024.

Aishik Nagar, Yutong Liu, Andy T Liu, Viktor Schlegel, Vijay Prakash Dwivedi, Arun-Kumar Kaliya-Perumal, Guna Pratheep Kalanchiam, Yili Tang, and Robby T Tan. umedsum: a unified framework for clinical abstractive summarization. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2653–2672, 2025.

Amin Heyrani Nobari, Kaveh Alimohammadi, Ali ArjomandBigdeli, Akash Srivastava, Faez Ahmed, and Navid Azizan. Activation-informed merging of large language models, 2025. URL `https://arxiv.org/abs/2502.02421`.

Harsha Nori, Mayank Daswani, Christopher Kelly, Scott Lundberg, Marco Tulio Ribeiro, Marc Wilson, Xiaoxuan Liu, Viknesh Sounderajah, Jonathan Carlson, Matthew P Lungren, et al. Sequential diagnosis with language models. *arXiv preprint arXiv:2506.22405*, 2025.

OpenAI. Introducing GPT-4.1 in the API, April 2025. URL `https://openai.com`. Accessed: 2026-05-05.

Arjun Panickssery, Samuel R Bowman, and Shi Feng. Llm evaluators recognize and favor their own generations. *Advances in Neural Information Processing Systems*, 37:68772–68802, 2024.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, 2019.

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. *Advances in neural information processing systems*, 36:53728–53741, 2023.

Mohit Raghavendra, Anisha Gunjal, Bing Liu, and Yunzhong He. Agentic rubrics as contextual verifiers for SWE agents. *arXiv preprint arXiv:2601.04171*, 2026.

Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. DeepSpeed: System optimizations enable training deep learning models with over 100 billion parameters. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD)*, pages 3505–3506, 2020. doi: 10.1145/3394486.3406703.

Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, et al. Medgemma technical report. *arXiv preprint arXiv:2507.05201*, 2025.

Guxue Shan, Xiaonan Chen, Chen Wang, Li Liu, Yuanjing Gu, Huiping Jiang, and Tingqi Shi. Comparing diagnostic accuracy of clinical professionals and large language models: systematic review and meta-analysis. *JMIR Medical Informatics*, 13(1):e64963, 2025.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, et al. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*, 2025.

Sijun Tan, Siyuan Zhuang, Kyle Montgomery, William Y. Tang, Alejandro Cuadron, Chenguang Wang, Raluca Ada Popa, and Ion Stoica. JudgeBench: A benchmark for evaluating LLM-based judges. *arXiv preprint arXiv:2410.12784*, 2024.

Tao Tu, Anil Palepu, Mike Schaekermann, Khaled Saab, Jan Freyberg, Ryutaro Tanno, Amy Wang, Brenna Li, Mohamed Amin, Nenad Tomasev, et al. Towards conversational diagnostic ai. *arXiv preprint arXiv:2401.05654*, 2024.

Anni Tziakouri and Filippo Menolascina. Reinforcement learning for clinical reasoning: Aligning llms with acr imaging appropriateness criteria. *arXiv preprint arXiv:2510.05194*, 2025.

Elahe Vedadi, David Barrett, Natalie Harris, Ellery Wulczyn, Shashir Reddy, Roma Ruparel, Mike Schaekermann, Tim Strother, Ryutaro Tanno, Yash Sharma, et al. Towards physician-centered oversight of conversational diagnostic ai. *arXiv preprint arXiv:2507.15743*, 2025.

Leandro von Werra, Younes Belkada, Lewis Tunstall, Edward Beeching, Tristan Thrush, Nathan Lambert, Shengyi Huang, Kashif Rasul, and Quentin Gallouedec. TRL: Transformer reinforcement learning. `https://github.com/huggingface/trl`, 2020.

Hanyang Wang, Lu Wang, Chaoyun Zhang, Tianjun Mao, Si Qin, Qingwei Lin, Saravan Rajmohan, and Dongmei Zhang. Text2grad: Reinforcement learning from natural language feedback. *arXiv preprint arXiv:2505.22338*, 2025a.

Haoxiang Wang, Wei Xiong, Tengyang Xie, Han Zhao, and Tong Zhang. Interpretable preferences via multi-objective reward modeling and mixture-of-experts. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 10582–10592, 2024.

Ling Wang, Jinglin Li, Boyang Zhuang, Shasha Huang, Meilin Fang, Cunze Wang, Wen Li, Mohan Zhang, and Shurong Gong. Accuracy of large language models when answering clinical research questions: systematic review and network meta-analysis. *Journal of Medical Internet Research*, 27:e64486, 2025b.

Michael Wornow, Rahul Thapa, Ethan Steinberg, Jason Fries, and Nigam Shah. Ehrshot: An ehr benchmark for few-shot evaluation of foundation models. *Advances in Neural Information Processing Systems*, 36:67125–67137, 2023.

David Wu, Fateme Nateghi Haredasht, Saloni Kumar Maharaj, Priyank Jain, Jessica Tran, Matthew Gwiazdon, Arjun Rustagi, Jenelle Jindal, Jacob M Koshy, Vinay Kadiyala, et al. First, do noharm: towards clinically safe large language models. *arXiv preprint arXiv:2512.01241*, 2025a.

Juncheng Wu, Wenlong Deng, Xingxuan Li, Sheng Liu, Taomian Mi, Yifan Peng, Ziyang Xu, Yi Liu, Hyunjin Cho, Chang-In Choi, et al. Medreason: Eliciting factual medical reasoning steps in llms via knowledge graphs. *arXiv preprint arXiv:2504.00993*, 2025b.

Zhenbang Wu, Anant Dadu, Mike Nalls, Faraz Faghri, and Jimeng Sun. Instruction tuning large language models to understand electronic health records. *Advances in Neural Information Processing Systems*, 37:54772–54786, 2024.

Huihui Xu, Yuanpeng Nie, Hualiang Wang, Ying Chen, Wei Li, Junzhi Ning, Lihao Liu, Hongqiu Wang, Lei Zhu, Jiyao Liu, et al. Medground-r1: Advancing medical image grounding via spatial-semantic rewarded group relative policy optimization. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 391–401. Springer, 2025.

Ran Xu, Tianci Liu, Zihan Dong, Tony You, Ilgee Hong, Carl Yang, Linjun Zhang, Tao Zhao, and Haoyu Wang. Alternating reinforcement learning for rubric-based reward modeling in non-verifiable llm post-training. *arXiv preprint arXiv:2602.01511*, 2026.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Xiao Yang, Xuejiao Zhao, and Zhiqi Shen. Ehrstruct: A comprehensive benchmark framework for evaluating large language models on structured electronic health record tasks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 34340–34348, 2026.

Le Yu, Bowen Yu, Haiyang Yu, Fei Huang, and Yongbin Li. Language models are super Mario: Absorbing abilities from homologous models as a free lunch. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2024. arXiv:2311.03099.

Xiangchi Yuan, Dachuan Shi, Chunhui Zhang, Zheyuan Liu, Shenglong Yao, Soroush Vosoughi, and Wenke Lee. Behavior knowledge merge in reinforced agentic models. *arXiv preprint arXiv:2601.13572*, 2026.

Kai Zhang, Christopher Malon, Lichao Sun, and Martin Renqiang Min. Editgrpo: Reinforcement learning with post-rollout edits for clinically accurate chest x-ray report generation. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 304–316, 2025.

Shuang Zhou, Wenya Xie, Jiaxi Li, Zaifu Zhan, Meijia Song, Han Yang, Cheyenna Espinoza, Lindsay Welton, Xinnie Mai, Yanwei Jin, et al. Automating expert-level medical reasoning evaluation of large language models. *npj Digital Medicine*, 2025.

Nikolas Zöller, Julian Berger, Irving Lin, Nathan Fu, Jayanth Komarneni, Gioele Barabucci, Kyle Laskowski, Victor Shia, Benjamin Harack, Eugene A Chu, et al. Human-ai collectives produce the most accurate differential diagnoses. *arXiv preprint arXiv:2406.14981*, 2024.

Yuxin Zuo, Shang Qu, Yifei Li, Zhangren Chen, Xuekai Zhu, Ermo Hua, Kaiyan Zhang, Ning Ding, and Bowen Zhou. Medxpertqa: Benchmarking expert-level medical reasoning and understanding. *arXiv preprint arXiv:2501.18362*, 2025.

A Contrast matrix against prior work, by design pillar

Table 6 positions every prior work cited in the introduction (and several closer analogues that the body of the paper does not have room to discuss in line) against the eleven design pillars of *CLR-voyance*. The table is intended as a navigational aid: the pillar columns map one-to-one to the framework's contributions in §3, and an entry of • indicates that the corresponding prior work substantively addresses the pillar, ● partial coverage (e.g., the right framing on a different modality), ○ no coverage, and "—" that the pillar is not applicable to that work's setting (e.g., rubric-design pillars for an EHR-prediction benchmark with no rubric).

Reading the columns. (1) POMDP framing of inpatient reasoning; (2) raw long-context EHR as input; (3) outcome-grounded query generation; (4) outcome-grounded answer / reference generation; (5) rubric-as-reward design; (6) per-case patient-trajectory-conditioned rubric synthesis; (7) outcome-aware grading judge; (8) RL post-training on this combined setting; (9) full-pipeline clinician alignment (query / rubric / judge / model); (10) production-scale physician-scoped deployment; (11a) frontier-model evaluation on a clinician-aligned inpatient-reasoning benchmark; (11b) reuse of clinician annotations to adapt the clinical judge and clinical preference model.

Reading the entries. An entry of ● on Pillar 9 means *all four* alignment levels are covered; ● on the same pillar means at least one (often A/B preference alone) is covered. Pillar 6 (per-case patient-trajectory-conditioned rubrics) is the column on which the literature is sparsest; the closest analogues (RM-R1, Rubric-ARM) are general-domain reward-model rubrics that are per-instance but not patient-trajectory-conditioned.

Table 6: Eleven-pillar coverage matrix for *CLR-voyance* versus prior work. ● addresses the pillar; ● partial / adjacent coverage; ○ does not address; — not applicable. Pillar 11 is split: 11a frontier-model eval, 11b judge / preference task adaptation from clinician annotations.

Paper	*1 POMDP*	*2 Raw EHR*	*3 Query gnd*	*4 Answer gnd*	*5 Rubric*	*6 Per-case rub*	*7 Outcome judge*	*8 RL on this*	*9 Full align*	*10 Deploy*	*11a Frontier eval*	*11b Judge/pref adapt*
HealthBench [Arora et al., 2025]	○	○	●	○	●	●	●	○	●	○	●	●
RBR [Mu et al., 2024]	—	—	—	○	●	○	●	●	●	○	○	●
RM-R1 [Chen et al., 2025]	—	—	—	○	●	●	●	●	○	○	○	●
Rubric-ARM [Xu et al., 2026]	—	—	—	○	●	●	●	●	○	○	○	●
RaR [Gunjal et al., 2025]	—	—	—	○	●	○	○	●	○	○	●	●
APB [Mallinar et al., 2026]	○	●	○	○	●	○	●	○	●	○	○	○
RLRA [Huang et al., 2025]	—	—	—	○	●	○	○	●	○	○	○	●
ArmoRM [Wang et al., 2024]	—	—	—	○	●	○	—	●	○	○	●	●
AdvancedIF / RIFL [He et al., 2025]	—	—	—	○	●	○	●	●	●	○	○	●
Text2Grad [Wang et al., 2025a]	—	—	—	○	●	○	●	●	○	○	○	●
InfAlign [Balashankar et al., 2024]	—	—	—	○	○	○	○	●	○	○	○	●
HuatuoGPT-o1 [Chen et al., 2024]	○	○	○	○	○	○	○	●	○	○	●	○
EditGRPO [Zhang et al., 2025]	○	○	○	●	●	○	○	●	○	○	●	○
MedGround-R1 [Xu et al., 2025]	○	○	○	●	●	○	○	●	○	○	○	○
CRPO [Gu et al., 2025]	○	○	○	○	●	○	○	●	○	○	●	○
ACR-RL [Tziakouri and Menolascina, 2025]	○	○	●	●	●	○	○	●	○	○	○	○
MedReason [Wu et al., 2025b]	○	○	●	○	○	○	○	●	○	○	●	○
EHR-R1 [Liao et al., 2025]	○	●	○	○	—	—	—	●	○	○	●	○
ER-Reason [Mehandru et al., 2025]	●	●	●	●	—	—	—	○	●	○	●	○
MedAlign [Fleming et al., 2024]	○	●	●	○	—	—	—	○	●	○	●	○
EHR-MCP [Masayoshi et al., 2025]	●	●	●	○	—	—	—	○	○	●	○	○
EHR-LLM-Impl [Griot et al., 2025]	○	●	○	○	—	—	—	○	●	●	○	○
Chinese clin. CoT [Ding et al., 2025]	○	○	●	●	○	○	○	○	●	○	○	●
EHRSHOT [Wornow et al., 2023]	○	●	—	—	—	—	—	○	○	○	○	○
CPLLM [Ben Shoham and Rappoport, 2024]	○	●	—	—	—	—	—	○	○	○	○	○
EHRStruct [Yang et al., 2026]	○	●	—	—	—	—	—	○	○	○	●	○
EHRNoteQA [Kweon et al., 2024]	○	●	●	○	—	—	●	○	●	○	●	○
DR.BENCH [Gao et al., 2023]	○	●	○	○	—	—	○	○	○	○	○	○
GraphCare [Jiang et al., 2023]	○	●	○	●	—	—	—	○	○	○	○	○
AMIE [Tu et al., 2024]	●	○	●	●	—	—	—	○	●	○	●	○
g-AMIE [Vedadi et al., 2025]	●	○	●	○	○	○	○	○	●	○	●	○
SDBench [Nori et al., 2025]	●	○	●	○	—	—	—	○	○	○	●	○
ALFA [Li et al., 2025]	●	○	●	●	○	○	○	●	●	●	○	●
MedThink-Bench [Zhou et al., 2025]	○	○	●	○	●	●	●	○	●	○	●	●
MedHELM [Bedi et al., 2025]	○	●	●	○	●	○	●	○	●	○	●	●
ICARE [Dua et al., 2025]	○	○	—	—	●	○	●	○	●	○	●	○
M-ARC [Kim et al., 2025]	○	○	○	○	—	—	—	○	●	○	●	○
Phys.-LLM meta-review [Shan et al., 2025]	○	○	—	—	—	—	—	○	●	○	●	○
LLM accuracy NMA [Wang et al., 2025b]	○	○	—	—	—	—	—	○	●	○	●	○
NOHARM [Wu et al., 2025a]	○	○	●	●	●	●	—	○	●	○	●	○
MedXpertQA [Zuo et al., 2025]	○	●	●	○	—	—	—	○	●	○	●	○
uMedSum [Nagar et al., 2025]	○	○	○	○	○	○	○	○	●	●	○	○
***CLR-voyance* (this paper)**	●	●	●	●	●	●	●	●	●	●	●	●

Where prior work concentrates, where it does not. The literature is dense on Pillar 5 (rubric design as a general construct) and Pillars 8 and 11a (RL post-training and frontier-model evaluation). It is sparser, in this order, on: Pillar 9 (full four-level clinician alignment, where the closest neighbour pairs are HealthBench [Arora et al., 2025] for rubric authorship plus AMIE [Tu et al., 2024] for A/B preference, but never both at once); Pillar 10 (production-scale physician-scoped deployment, on which only EHR-MCP [Masayoshi et al., 2025], the on-premises RAG retrieval/summarisation rollout of Griot et al. [2025] (1,028 users, 14,910 conversations across nine specialties), and summarisation work [Nagar et al., 2025] have any non-OSCE evidence; none of these trains an inpatient-reasoning model); Pillar 11b (reuse of clinician annotations as judge / preference task-adaptation data, on which the closest direct neighbours are general-domain: RM-R1 [Chen et al., 2025], Rubric-ARM [Xu et al., 2026], ArmoRM [Wang et al., 2024], InfAlign [Balashankar et al., 2024]); and finally Pillar 6 (per-case patient-trajectory-conditioned rubrics), which is empty in the medical literature and is approximated only by general-domain per-instance rubric methods. *CLR-voyance* is the first work we are aware of to address all eleven pillars at once on a single inpatient-reasoning task; the matrix above is the per-pillar bookkeeping.

B Reproducibility Details

This appendix provides the implementation-neutral specification of the *CLR-voyance*framework. It fully expands the prompts, hyper-parameters, scoring rules, and merge configurations referenced in the main text. An independent team should be able to reproduce every stage of the pipeline using only this appendix and the data-curation rules in Section 3.1.

Artefact release statement. We are unable to publicly release *CLR-voyance*-specific artefacts (model weights, generated rubrics, candidate responses, or curated patient timelines) because of the non-disclosure agreements and patient-privacy contracts we have signed with the partner public health networks under which the deployment runs. Every component of the pipeline is, however, specified in sufficient detail in this appendix and the main text to be reproduced from scratch given access to the publicly available MIMIC-IV source data. To stress-test that claim, an independent researcher who was not involved in the study in any capacity used commodity coding agents to replicate the entire pipeline (data extraction, query and rubric generation, GRPO post-training, model merging, and evaluation) end-to-end in under a day from the moment MIMIC-IV access was granted, working only from the paper and this appendix.

B.1 Data pipeline

Source modules. *CLR-POMDP* is constructed from MIMIC-IV v3.0, MIMIC-IV-ED v2.2, and MIMIC-IV-Note v2.2. For every admission we pull a unified event stream from four schema groups:

- *ED*: ED stays, diagnosis, medication reconciliation, pyxis, vital signs, triage.
- *Hospital*: admissions, transfers, lab events, EMAR, HCPCS events, prescriptions, ICD-coded procedures.
- *ICU*: chart events, datetime events, input events, output events, procedure events.
- *Notes*: discharge summaries, radiology reports.

Each event is enriched with the dictionary tables (lab items, ICD diagnoses, ICD procedures, HCPCS, ICU items) so free-text labels (*not* integer IDs) are exposed downstream.

Per-admission JSON structure. Every admission is serialised as

```
{
  "subject_id": int,
  "hadm_id": int,
  "demographics": {gender, anchor_age, anchor_year_group, ...},
  "timeline": [
    {"time": ISO-8601, "source": table-name,
     "table": schema.table, "data": {...},
     "descriptions": {dictionary enrichments}},
```

```
    ...
  ],
  "misc": {
    "ED_triage": [...],
    "Patient": {subject_id, gender, anchor_age, anchor_year_group, dod},
    "ICD_Diagnoses": [...], "ICD_Procedures": [...], "HCPCS": [...]
  }
}
```

The full `timeline` is sorted in descending time order; items with missing timestamps are dropped.

Cohort selection (greedy set cover). Let $\mathcal{H}$ be the set of admissions and $\mathcal{C}_h$ the multiset of categorical tokens of admission h, drawn from

$$\mathcal{C}_h \;=\; \big\{\text{ICD:code–ver}\big\}_h \;\cup\; \{\text{Gender}\}_h \;\cup\; \{\text{Age}\}_h \;\cup\; \{\text{Height_bin}\}_h \;\cup\; \{\text{Weight_bin}\}_h.$$

Starting from $U = \bigcup_h \mathcal{C}_h$ and $S = \emptyset$, repeatedly pick $h^\star = \arg\max_h |\mathcal{C}_h \cap U|$, add $h^\star$ to S, and remove $\mathcal{C}_{h^\star}$ from U until $U = \emptyset$. A disjoint random sample of 20% of $\mathcal{H} \setminus S$ forms the test cohort; seed $= 42$ for all samplers.

Timeline splitting. Sort a patient's timeline chronologically into $e_{1:n}$. Sample

$$k \sim \mathcal{N}(\mu = n/2,\, \sigma = n/6) \quad \text{clamped to } [1, n-1],$$

and let $\text{past} = e_{1:k}$, $\text{future} = e_{k+1:n}$, $\text{split_time} = \text{time}(e_{k+1})$. Events whose `source` string contains the substring `ICD` are removed from the past before it is presented to any policy. Only `Patient` and `ED_triage` entries are retained from `misc`; all other misc blocks are dropped.

B.2 Question–reference generation prompt

The *query and reference generation* call of the Grounded Judge uses the simulator LLM `Qwen3-32B` under structured JSON decoding. The schema enforced on the output is:

```
class SourceDetail(BaseModel):
    event: str
    time: str
    source: str

class ClinicalQAPair(BaseModel):
    question: str
    final_answer: str
    answer_reasoning: str
    action_space_category: str
    action_space_subcategory: Optional[str]
    source: List[SourceDetail]
```

The prompt template (verbatim, with placeholders in `{}`) is:

```
You are a clinical reasoning simulator LLM. Your goal is to generate a
question-answer pair for training a smaller LLM in the context of
reasoning about clinical decision support.
The clinical encounter has been split into two parts: past and future.
The past data includes the patient's history up to a certain point,
and the future data includes the patient's clinical outcomes and
events after that point.
--------------------
PAST DATA:
{past_data_json}
{misc_data_json}

Action Space Category: {action_space_category}
```

```
Action Space Description: {action_space_description}

Generate a clinical question that is answerable from the past data
only. The clinician-style answer and answer_reasoning must be written
as if the model had access only to the past, even though you (the
simulator) can see the future. Avoid information-extraction style
questions ("What was the last lactate?"); prefer synthesis questions
that require reasoning over multiple signals (labs, vitals, notes).

Use the FUTURE DATA below only to (a) pick clinically meaningful
questions whose gold answer actually occurs downstream, and (b)
populate the "source" field with the specific future events that
verify the answer. Do NOT mention the future in the question,
final_answer, or answer_reasoning text.

FUTURE DATA:
{future_data_json}

Return a single JSON object with fields
{question, final_answer, answer_reasoning, action_space_category,
 action_space_subcategory, source}.
```

Sampling parameters: temperature 1.0, top-p 0.95, top-k 64. Retry: up to three attempts with exponential backoff on transient API errors or schema-validation failures.

B.3 Action-space category descriptions

The four action-space descriptions injected into the QA generation prompt are reproduced verbatim below.

Diagnosis Assistance.

> This category focuses on clinical reasoning to determine the correct diagnosis. It includes the following subcategories: (i) Differential Diagnosis Generation & Diagnostic Hypothesis Ranking: the model generates a ranked list of potential diagnoses — covering common conditions as well as critical must-not-miss possibilities — by analysing abnormal vitals, lab trends, and clinical notes; (ii) Diagnostic Test Suggestions: in cases of uncertainty, the model may recommend additional tests (e.g., ordering a troponin test for suspected myocardial infarction) to refine the diagnosis. Additional contextual information (e.g., family history, vaccination status, travel history) should be considered when available.

Treatment Recommendations.

> This category is aimed at guiding therapeutic interventions and ensuring best practices in patient care. It is subdivided into: (i) Medication Management — the model recommends appropriate medications, specifying drug name, dosage, and duration, while taking into account patient-specific factors (e.g., allergies or prior adverse reactions); (ii) Supportive Care & Monitoring — the model suggests supportive measures such as IV fluids, oxygen therapy, or nursing care orders that complement primary treatments; (iii) Follow-up and Long-Term Planning — the model outlines future monitoring steps, reassessment timings, or specialist consultations. The answer should propose a future treatment event rather than merely extracting past actions.

Procedural Decision Making.

> This category addresses decisions related to both diagnostic and therapeutic procedures. It is organised into: (i) Diagnostic Procedures — recommendations for tests such as imaging, biopsies, or endoscopies, determined from evolving clinical data; (ii) Therapeutic Procedures — decisions about invasive interventions (e.g., surgery, catheterisation) when indicated; (iii) Timing, Approach, and Watchful Waiting — the model assesses the optimal timing for procedures, chooses between alternative approaches (e.g., minimally invasive versus open surgery), or recommends non-intervention when appropriate.

Responding under Uncertainty.

> This category focuses on how the model should handle situations where the clinical data is incomplete or ambiguous. The questions generated for this category should be intentionally ambiguous and must require the agent to reason about how to handle uncertainty. It includes: (1) Uncertainty Acknowledgment; (2) Risk-Benefit Analysis; (3) Alternative Recommendations; (4) Escalation Protocols; and (5) Context Seeking (asking for additional information that could help clarify the clinical picture).

B.4 Policy system prompt (training and evaluation)

The policy's system prompt is identical during SFT, GRPO, and inference, and is reproduced verbatim below. A second variant with an explicit `<answer>` wrapper is used for base models that are not reasoning-tuned; the only difference is the output structure line (`<think>...</think><answer>...</answer>` instead of `<think>...</think>Final Answer...`).

> You are a helpful AI Assistant in the clinical domain that provides well-reasoned and detailed responses. You first think about the reasoning process as an internal monologue and then provide the user with the answer. You Respond in the following format: <think>\n...\n</think>\nFinal Answer....
>
> You are a highly specialized AI Assistant operating in the clinical domain. Your primary purpose is to provide meticulously reasoned, detailed, and evidence-based responses to queries about patient cases. You must analyze all provided clinical data to generate comprehensive summaries, assessments, differential diagnoses, and potential management plans. Your communication must be clear, structured, and tailored for a clinical audience.
>
> For every single request, you MUST adhere to a strict two-step process. First, you will conduct an internal, step-by-step reasoning process. Second, you will formulate the final, polished answer for the user. Your entire output must be encapsulated within the following structure:
>
> <think> ... Your detailed chain-of-thought reasoning process, analysis, and step-by-step logic, phrased in first person, go here. ... </think> ... Your final, user-facing, and fully formatted answer goes here. ...
>
> # The <think> Block: Internal Monologue This section is your private workspace for reasoning and is not part of the final answer presented to the user. In this block, you must:
>
> [leftmargin=*,itemsep=0pt]
>
> - Deconstruct the user's query to identify the core clinical question(s).
> - Systematically review all provided patient information (e.g., demographics, history of present illness (HPI), past medical history, medications, lab results, imaging reports and any other presented information).
> - Synthesize the data, explicitly connecting relevant findings and noting significant positives and negatives.
> - Formulate a clinical assessment or differential diagnosis by weighing the evidence.
> - Outline the logical foundation for your final recommendations or plan.
> - Briefly reference relevant clinical guidelines or established medical knowledge that informs your reasoning.
> - Evaluate uncertainty based on available information, query clarity and ensure you have enough information to reply to the user query confidently. Request for further information and clarification if needed.
>
> Your think process should be an internal monologue and should reason about how to answer the user's question based on the conversation

```
history and any external context provided to you, it should not be a
direct answer to the user's question but rather a chain of thought
in the form of an internal monologue process that leads to the final
answer.
# The Final Response This section contains the polished, final
answer for the user.  It MUST strictly follow these rules:
    [leftmargin=*,itemsep=0pt]
  • Markdown Formatting:  The entire content must be formatted using
    Markdown for optimal readability.  Use headers (#, ##), lists
    (*, 1.), bold (**text**), and other elements to structure the
    information clearly.
  • Cite the Context:  You must support your statements by
    explicitly citing the patient data provided in the prompt.
    For example:  ``The patient's fever of $39.1^\circ C$ (from
    Vitals)...''.
  • Professional Tone:  Maintain a clinical, objective, and precise
    tone throughout.
  • LaTeX for Notation:  Use LaTeX for all mathematical and
    scientific notations.  Enclose all LaTeX code within $ or $$
    delimiters (e.g., $\text{Na}^+ > 145 \text{ mEq/L}$).
```

User-turn template. At both training and inference time the user turn is:

```
You are a clinical reasoning assistant. Using the past timeline data
provided below, answer the following question with your final answer
and step-by-step reasoning.

Patient Info:
{json patient_info}

Past Timeline:
{json past_timeline}

Question: {question}

{reasoning-format instruction selected per model family}
```

The reasoning-format instruction is one of: (a) a JSON schema with `answer_reasoning` and `final_answer` (for off-the-shelf non-reasoning models); (b) the `<think>...</think><answer>...</answer>` wrapper; (c) the `<think>...</think>` + free-text answer format (our Qwen3-family setting); (d) the `## Thinking / ## Final Response` headers (for comparison with distilled medical reasoners).

B.5 Rubric-generation system prompt

The Grounded Judge authors rubrics with the following system prompt, reproduced verbatim. The value of `{TASK}` is the string "Clinical Reasoning". The body of `{INPUT_DATA_DESCRIPTION}` is reproduced verbatim below.

```
You are an expert clinician who is tasked with generating a rubric
for evaluating the quality of health-related responses.
This rubric will judge the quality of responses generated by a
large language model (LLM) in the context of health-related tasks,
specifically for {TASK}.
{INPUT_DATA_DESCRIPTION}.
The target audience for reading the answers will always be
healthcare professionals such as clinicians, physicians and nurses.
Your generated rubric must ALWAYS contain at least one
criterion from each of the five axes:  Accuracy, Completeness,
CommunicationQuality, ContextAwareness, InstructionFollowing.
```

You must generate only a single rubric with multiple criteria. Do not generate more than one rubric or any other text!

NEVER mention the future timeline in your rubric, it is imperative that the rubric is based on the past timeline only. Only use the future timeline as your own reference to build the rubric criteria, but never include a direct reference to it in your criteria or as a task/requirement for the model to follow.

{INPUT_DATA_DESCRIPTION} expands to:

The input data to judge will contain a timeline of a patient record, which is split into the past and future. The user query will ask a question about the patient based on the past timeline which can be answered by some information in the future timeline of the patient.

The ideal answer contains both a reasoning and a final answer. The reasoning should explain the final answer and the reasoning should be based on the past timeline. Your generated rubrics should contain criteria which judge the quality of both the reasoning and the final answer.

For instruction following the LLM will only have the past timeline to answer from, so ensure that you don't put an instruction-following reward requiring access to the future timeline.

Having a reasoning and final answer are enforced on the model. It is not necessary to have a separate instruction following reward for this aspect of the output in the rubrics. If no instruction following reward is required, you may skip it; it is more for rewarding user requests such as formatting, information retrieval, etc.

Rubric rules (appended to the system prompt):

RUBRIC RULES
[leftmargin=*,itemsep=0pt]

- Every criterion is self-contained and graded pass / fail.
- Give each criterion a non-zero integer score from -10 to $+10$ -- positive values reward helpful behaviour, negative values penalise unsafe or distracting behaviour.
  [leftmargin=*,itemsep=0pt]
  - For negative values, the description should still be something which can be checked off as present or not. If the description is present in the response, the negative score is added to the model's score (e.g., "Mentions wrong medication" $\rightarrow -5$).
  - For positive values, the description should likewise be binary-checkable (e.g., "Mentions precaution in clinical care" $\rightarrow +7$).
- Penalise negative behaviour and reward positive behaviour; negative behaviour is particularly important in clinical settings where errors can be critical.
- Final score $=$ (sum of points for met items) $\div$ (sum of all positive point values), then clip to $[0, 1]$.
- Tag each criterion with exactly one axis: Accuracy · Completeness · CommunicationQuality · ContextAwareness · InstructionFollowing. The finished rubric must include at least one criterion per axis.
- Use any of the 34 consensus criteria (inspired by HealthBench and enumerated inline in the prompt) whenever they apply, or write new criteria.
- Keep wording neutral; do not award partial credit; never mention model internals.

Output schema (required JSON form):

```
{
  "meta": { "theme": "string" },
  "criteria": [
    {
      "axis": "Accuracy | Completeness | CommunicationQuality
             | ContextAwareness | InstructionFollowing",
      "description": "string",
      "points": 5            // -10..+10, non-zero
    }
    // ... one or more entries, every axis appears >= 1x
  ]
}
```

Themes and in-context exemplars. Each rubric is anchored to exactly one of seven themes: *Emergency Referrals, Responding under Uncertainty, Health Data Tasks, Global Health, Expertise-Specific Communication, Context Seeking, Response Depth*. For each theme, one full demonstration block (conversation $\rightarrow$ candidate answer $\rightarrow$ reference rubric) is appended to the system prompt as in-context learning. The demonstration blocks are inspired by the consensus-criteria structure in HealthBench but are rewritten for the inpatient-reasoning setting, and are iteratively updated with edits collected from licensed clinicians on a held-out calibration split so that the oracle internalises physician preferences.

B.6 Rubric-grading (judge) system prompt

At scoring time the Grounded Judge operates with the following system prompt, reproduced verbatim:

```
You are a medical-AI evaluator.
Given patient context, the full conversation, the assistant's answer,
and a rubric with numbered criteria, decide independently for each
criterion whether the answer MEETS it.
    [leftmargin=*,itemsep=0pt]
  • Return true if the answer satisfies the criterion, false
    otherwise.
  • Output JSON only, whose keys are the criterion IDs shown in the
    rubric.
  • If you find that the answer has repeating text or is nonsensical,
    you must mark all criteria as false.  This is a sign of
    degenerate outputs which are not acceptable under any
    circumstance, regardless of if certain parts of the answer may
    seem to meet some criteria.
  • If both the reasoning and answer to be judged are empty, mark
    all criteria as false.
  • Accept multilingual responses; do not judge against the language
    used as long as the output makes sense.
Example format.  If the rubric lists
c1:  "Provides clear emergency advice"
c2:  "Uses correct drug dose"
then your reply must be exactly
{ "c1":  true, "c2":  false }
No other keys, comments, or text.
```

User-turn template for the grader.

```
PATIENT CONTEXT
{context_str}

CONVERSATION
{conversation_history}
```

```
LAST USER MESSAGE
{user_query}

ASSISTANT RESPONSE TO BE JUDGED
{model_response}

RUBRIC  (each criterion has an ID)
c1: {description_1}
c2: {description_2}
...

INSTRUCTIONS
Return one JSON object. Every key must be exactly the criterion ID
(e.g. "c1") and the value must be ‘true‘ or ‘false‘.
```

where `context_str` is the concatenation of labelled blocks: `PATIENT PAST TIMELINE`, `PATIENT MISC INFO`, `REFERENCE ANSWER`, `REFERENCE ANSWER REASONING`, `SOURCE FOR REFERENCE ANSWER`. The reference and source blocks are present at grading time; they are what give the oracle its outcome awareness while keeping the rubric text itself past-verifiable.

Default fallback rubric. If rubric parsing fails the grader falls back to the following six-criterion default:

Axis	Pts	Criterion
Accuracy	+10	Correctly identifies the answer with clinical accuracy.
Completeness	+8	Mentions supporting evidence for the identified answer from the patient timeline.
CommunicationQuality	+7	Uses clear and concise language appropriate for healthcare professionals.
ContextAwareness	+9	References relevant clinical context such as the identified source from the EHR document.
Completeness	-6	References unnecessary or imprecise information in its reasoning or answer.
InstructionFollowing	+5	Provides a reasoning section and final answer as requested.

B.7 Reward functions used during GRPO

All rewards are computed on-policy per completion; the group mean and variance across the G rollouts form the baseline.

Grounded rubric reward r_{rub}. Computed via the judge pipeline above:

$$r_{\text{rub}}(\hat{y} \mid d) = \text{clip}_{[0,1]}\Big(\frac{\sum_{i:j_i=\text{true}} p_i}{\sum_{i:p_i>0} p_i}\Big).$$

Degenerate rollouts (empty reasoning/answer or trivial repetition) are mapped to $r_{\text{rub}} = 0$.

Format rewards. Two mutually-exclusive variants depending on whether the policy emits an `<answer>` block.

With explicit answer wrapper. $r_{\text{format}} = 1$ iff the completion matches the regex

```
^<think>.*?</think>.*?<answer>.*?</answer>$
```

evaluated with `DOTALL`. Otherwise $r_{\text{format}} = 0$.

Without answer wrapper (Qwen3-family). $r_{\text{format}} = 1$ iff the completion matches

```
^\s*<think>\s*.*?\s*</think>\s*(\S[\s\S]*)$
```

and the trailing content does not itself start with another `<think>`.

Tag rewards. Partial-credit structural rewards on the `<think>` scaffolding.

With answer wrapper. $r_{\text{tag}} = 0.25 \cdot [n(\texttt{<think>}) = 1] + 0.25 \cdot [n(\texttt{</think>}) = 1] + 0.25 \cdot [n(\texttt{<answer>}) = 1] + 0.25 \cdot [n(\texttt{</answer>}) = 1]$.

Without answer wrapper. $r_{\text{tag}} = 0.4 \cdot [n(\texttt{<think>}) = 1] + 0.4 \cdot [n(\texttt{</think>}) = 1] + 0.2 \cdot$ [non-empty content after `</think>` that does not start with `<think>`].

Reasoning-steps reward (optional). $r_{\text{steps}} = \min(1,\, k/3)$ where k counts regex matches of step markers ("Step N:", numbered lists, bullet points, transition words) inside the `<think>` block. We found this useful for the 1.5B scale but not for the 8B scale.

Repetition penalty (optional). $r_{\text{rep}} = -\lambda \cdot \big(1 - |\text{unique-}n\text{-grams}| \,/\, |\text{total-}n\text{-grams}|\big)$ with $n = 3$ and $\lambda = 1.0$.

Canonical reward stack. For Qwen3-8B and Qwen2.5-7B we use $r_{\text{total}} = r_{\text{rub}} + r_{\text{format}} + r_{\text{tag}}$ (unit weights). For the 1.5B anchor we additionally enable r_{steps} and r_{rep} for early stabilisation.

B.8 Training hyper-parameters

Hyper-parameter	Qwen2.5-1.5B	Qwen2.5-7B	Qwen3-8B	MedGemma-4B
SFT (ablation only; not in the production recipe)				
Learning rate	1×10^{-5}	1×10^{-5}	1×10^{-5}	1×10^{-5}
Per-device batch	8	2	2	2
Grad. accumulation	2	4	4	4
Epochs	3	1–3	1	3
GRPO (production recipe)				
Init. checkpoint	base	base	base	base
Learning rate	2×10^{-5}	1×10^{-6}	1×10^{-6}	1×10^{-6}
G (generations)	8	4	8	4
Per-device batch	8	1	4	1
Grad. accumulation	4	2	4	2
Max prompt length	65,536	65,536	65,536	65,536
Max completion len	2,048	2,048	4,096	2,048
Epochs	5	1	1	1
LR schedule	cosine	cosine	cosine	cosine
Warm-up ratio	0.1	0.1	0.1	0.1
Precision	bf16	bf16	bf16	bf16
Attention	FlashAttn-2	FlashAttn-2	FlashAttn-2	FlashAttn-2
Distributed	ZeRO-3	ZeRO-3	ZeRO-3	ZeRO-3

Production initialisation. Every production GRPO run is initialised directly from the base checkpoint listed at the top of its column; SFT is reported only as a stand-alone ablation in Appendix B.12 and is not chained into the GRPO step.

B.9 Merge configurations (weight-space)

All merges are between the base model θ_{base} and its GRPO-trained counterpart θ_{grpo}. For *CLR-voyance*-8B the base is `Qwen/Qwen3-8B` and the GRPO checkpoint is the final policy of Section B.8. The filter is applied to the task vector $\Delta = \theta_{\text{grpo}} - \theta_{\text{base}}$ before adding it back at unit weight.

DELLA-Linear (final recipe used for deployment). Row-wise adaptive pruning of Δ: within each row of each weight matrix, the largest-magnitude entries are retained at density ρ; the rest are zeroed. We use $\rho = 0.15$, `epsilon` $= 0.02$, and merge weight 1.0; dtype `bfloat16`. This is the configuration that produced our headline clinical-reasoning and generalist-preservation numbers.

Breadcrumbs (ablation). Two-sided magnitude pruning of Δ: the top γ quantile (destructive outlier spikes) and the bottom β quantile (RL noise) are pruned; the remaining middle band is retained at density ρ. We use $\rho = 0.15$, $\gamma = 0.02$, weight 1.0.

DARE-Linear (ablation). Random dropout of Δ at rate $1 - \rho$ with rescaling by $1/\rho$, followed by linear addition. We use $\rho = 0.5$, weight 1.0.

Activation-Informed Merging (negative-result ablation). A continual-learning post-filter applied after any of the above: a calibration batch is passed through the base model, per-weight activation magnitudes are recorded, and weights whose activation magnitude exceeds a threshold are locked to their base values before the filtered Δ is added. Across all three filters we found AIM neither improved clinical-reasoning scores nor measurably protected generalist behaviour once DELLA-Linear was in place; in a minority of settings it slightly regressed both. We therefore do not include AIM in the production recipe.

B.10 Evaluation protocol (reproducibility view)

***CLR-POMDP* evaluation.** For every (admission, action-space) item in the held-out split:

[leftmargin=*,itemsep=0pt]

1. Construct the policy prompt (Section B.4).
2. Sample one generation from the model under test with temperature 0.2 and a maximum of 4,096 new tokens.
3. Parse the generation into `answer_reasoning` and `final_answer` according to the model's reasoning-trace family.
4. Call the Grounded Judge with the per-item rubric (Section B.5) and compute r_{rub}.

Report: overall mean r_{rub}; per-axis aggregates (Accuracy, Completeness, Communication Quality, Context Awareness, Instruction Following); per-action-space aggregates (four categories); and a *critical-accuracy* score computed over criteria whose axis is ACCURACY and whose absolute weight ≥ 8. Rubrics and verdicts are persisted so that re-aggregation against alternative scoring rules is possible post-hoc.

OpenAI HealthBench. The same checkpoints are evaluated on the public HealthBench FULL, HARD, and CONSENSUS subsets using the official grader template and the same positive-point-normalised scoring rule. No fine-tuning on HealthBench is performed; this is a pure transfer test.

Clinician A/B preference. A blinded A/B study with licensed physicians on a held-out panel of inpatient cases compares pairwise responses from *CLR-voyance*, GPT-5, and HuatuoGPT-o1. Reviewers see only the patient past and the two anonymised responses; judgements are collected on clinical usefulness, safety, and communication clarity.

B.11 Determinism and seeding

All data-preparation samplers (cohort selection, timeline splitting, train/test split) use seed 42. RL rollouts are non-deterministic by design; we report bootstrapped confidence intervals over the test set rather than single-seed point estimates. Every stage of the pipeline writes a self-contained artefact (traces, QA pairs, rubrics, rollouts, verdicts) so that re-running a downstream stage against a frozen upstream is straightforward.

B.12 Training-stack ablation: SFT vs. GRPO and merge variants

This appendix expands the two-sentence ablation reference in §4. We report the full per-stage and per-merge-variant per-axis numbers on *CLR-POMDP*, and then walk through the failure modes that distinguish each stage by inspecting representative outputs and the per-criterion verdict patterns from the Grounded Judge.

Per-stage table. Table 7 reports the per-axis scores for the four configurations we evaluate on each base family: *Base*, *SFT* on the oracle reference $(\rho^\star, y^\star)$ as a stand-alone ablation, *GRPO* on the rubric reward r_{rub} initialised from the same base, and the two strongest weight-space merges of the GRPO task vector back into that base (DELLA-Linear and Breadcrumbs). SFT and GRPO are both initialised from *Base* and are not chained; the merge rows are initialised from the corresponding GRPO row. All numbers are micro-aggregated over per-criterion points across the held-out 7,927-admission test split (Section B.10) under the same Grounded Judge and sampling parameters.

Table 7: Per-axis *CLR-POMDP* scores (%) for each training configuration on the Qwen3-8B and MedGemma-4B families. **Bold** marks the best score per family per axis. SFT and GRPO are independent ablations off *Base*; the merge rows are initialised from the corresponding GRPO row.

Family	Configuration	Acc.	Comp.	Ctx.	Comm.	Instr.	Aggr.
Qwen3-8B	Base	66.57	58.72	76.97	**91.62**	88.24	75.32
Qwen3-8B	SFT (ablation)	42.10	17.83	51.08	85.98	74.17	52.36
Qwen3-8B	GRPO	69.21	67.92	80.29	90.15	86.68	77.99
Qwen3-8B	GRPO + Breadcrumbs	79.22	79.29	85.87	**91.67**	**89.05**	84.53
Qwen3-8B	GRPO + DELLA-Linear	**80.77**	**80.98**	**86.19**	90.50	88.14	**84.95**
MedGemma-4B	Base	20.57	8.11	24.79	47.19	40.06	27.01
MedGemma-4B	SFT (ablation)	44.78	16.74	55.19	**84.24**	71.73	52.89
MedGemma-4B	GRPO	56.63	43.17	**68.47**	81.20	**72.16**	**63.40**
MedGemma-4B	GRPO + Breadcrumbs	55.87	45.58	61.14	72.45	67.29	59.79
MedGemma-4B	GRPO + DELLA-Linear	**64.11**	**54.49**	67.29	64.98	66.56	63.35

Why SFT regresses on Qwen3-8B. The most striking row in Table 7 is the Qwen3-8B SFT ablation, which drops the aggregate by 22.95 pp relative to the base. Qwen3-8B already carries a reasoning prior, and the SFT target $< \texttt{think} > \rho^\star < /\texttt{think} > y^\star$ contracts the policy in two harmful ways simultaneously. First, $y^\star$ is a single concise oracle answer that mentions one downstream care action; cloning it freezes Completeness onto a one-action ceiling that the rubric's per-criterion enumeration explicitly rewards beyond (-40.89 pp Completeness vs. base). Second, $\rho^\star$ is written as a tidy oracle monologue rather than as the model's natural exploratory $< \texttt{think} >$ trace, and cloning it suppresses the model's pre-installed reasoning style on prompts that fall outside the SFT distribution (-25.89 pp ContextAwareness, -14.07 pp InstructionFollowing). On reading the SFT outputs we observe two characteristic failure modes. *Mode A: short verbatim mimicry.* The policy emits a one-paragraph $< \texttt{think} >$ followed by a one-line answer, hits the matched-criterion in the rubric, and misses every other positive criterion the case carries. *Mode B: format collapse on out-of-distribution prompts.* On prompts that drift from the SFT format the policy emits a $< \texttt{think} >$ block but no terminal answer, and the rubric scores all positive criteria as false. The conversational axes (CommunicationQuality, InstructionFollowing) survive SFT, while the content axes collapse.

Why SFT lifts MedGemma-4B but still trails GRPO. Run as the same stand-alone ablation, SFT on MedGemma-4B reaches 52.89%, 25.88 pp above its very low base (27.01%); the lift is concentrated on the conversational axes ($+37.05$ Communication, $+31.67$ Instruction) and is more modest on the content axes ($+24.21$ Accuracy, $+8.63$ Completeness). The reading is that the SFT target is essentially a faithful demonstration of the $< \texttt{think} >$/answer scaffolding plus a clinical answer, and that scaffolding is what MedGemma-4B was missing relative to a reasoning-tuned base. GRPO from the same base reaches 63.40% ($+36.39$ pp over base, $+10.51$ pp over the SFT ablation), and the lift sits on the content axes ($+36.06$ Accuracy, $+35.06$ Completeness, $+43.68$ ContextAwareness vs. base) where the rubric reward provides a direct gradient that the static SFT target does not. The content gains come at a small concession on the conversational axes (-3.04 Communication, -0.43 Instruction vs. SFT), which the merge step then has to weigh against the content lift it provides.

Merge-variant comparison on Qwen3-8B. Table 8 reports the four merge variants we evaluate on the Qwen3-8B GRPO task vector. All four variants score within 0.7 pp of one another at the aggregate level, and the per-axis ordering is also stable: every variant sits within ± 1 pp of the panel mean on Communication and Instruction, within ± 1 pp of one another on Context, and within ± 2 pp on the two content axes. This narrow band tells us that on a Qwen3-8B–scale base with our GRPO recipe the merge step delivers most of its lift through the structural act of restoring base-model weights to a high-density subset of the task-vector entries; the choice between density-controlled (DELLA-Linear), magnitude-banded (Breadcrumbs), and stochastic-thinning (DARE-Linear; not shown in the table because it underperformed by > 3 pp aggregate in initial runs and we did not scale it) filters matters less than the presence of a filter at all. Activation-Informed Merging (AIM) sits within noise of its base filter on every axis (-0.21 aggregate against DELLA-Linear, -0.24 aggregate against Breadcrumbs); we therefore drop it from the production recipe for both the cost-of-calibration and the negligible lift.

Table 8: Merge-variant per-axis comparison on Qwen3-8B GRPO (*CLR-POMDP*, %). All four variants are within 0.7 pp of one another at the aggregate; AIM is within noise of its base filter.

Variant	Acc.	Comp.	Ctx.	Comm.	Instr.	Aggr.
GRPO + DELLA-Linear	**80.77**	**80.98**	**86.19**	90.50	88.14	**84.95**
GRPO + DELLA-Linear + AIM	80.40	80.76	85.60	90.48	88.41	84.74
GRPO + Breadcrumbs	79.22	79.29	85.87	91.67	89.05	84.53
GRPO + Breadcrumbs + AIM	78.71	78.99	85.34	**91.91**	**89.08**	84.29

Merge-variant comparison on MedGemma-4B and the axis-rebalancing pattern. Table 9 reports the same comparison for MedGemma-4B GRPO. Here the merge step does not raise the aggregate (DELLA-Linear 63.35, Breadcrumbs 59.79, vs. pre-merge GRPO 63.40); it *re-balances* the per-axis allocation. DELLA-Linear lifts the two content axes by 7.48 and 11.32 pp and concedes 16.22 and 5.60 pp on the conversational axes. Breadcrumbs is more conservative on the conversational side (-8.75 Communication) but also less aggressive on content. Reading paired outputs on the same admissions, the conversational drop is mostly the merged 4B re-acquiring the base MedGemma-4B's tendency to drop into mid-answer markdown and to truncate before a closing summary; the content lift is the merged 4B preserving the GRPO-installed differential-and-evidence structure. We retain DELLA-Linear as the production 4B recipe because the per-axis content lifts are the ones that matter for the deployment surface (§6), and because the conversational drop is partially recovered by the deployed system prompt. The qualitative outputs that justify this trade are the four representative cases in Appendix G.

Table 9: Merge-variant per-axis comparison on MedGemma-4B GRPO (*CLR-POMDP*, %). The merge step on the 4B family is a per-axis rebalancing rather than an aggregate lift: DELLA-Linear gains content-axis points and concedes conversational-axis points relative to pre-merge GRPO.

Variant	Acc.	Comp.	Ctx.	Comm.	Instr.	Aggr.
GRPO (no merge)	56.63	43.17	**68.47**	**81.20**	**72.16**	**63.40**
GRPO + DELLA-Linear	**64.11**	**54.49**	67.29	64.98	66.56	63.35
GRPO + DELLA-Linear + AIM	63.05	53.59	66.77	65.96	66.25	62.94
GRPO + Breadcrumbs	55.87	45.58	61.14	72.45	67.29	59.79
GRPO + Breadcrumbs + AIM	55.61	44.61	61.02	73.27	67.80	59.74

Per-criterion verdict patterns. To complement the per-axis aggregates we also inspected the Grounded Judge's per-criterion verdict pattern on a fixed 200-item slice. Three patterns recurred. (i) Compared head-to-head against the SFT ablation on Qwen3-8B, GRPO converts on average 4.7 false-to-true verdicts per case on positive criteria of the Completeness axis, dominated by "mentions a differential beyond the most likely diagnosis" and "cites the relevant past lab/vital with a value". (ii) The GRPO $\rightarrow$ DELLA-Linear merge step adds an additional 1.8 false-to-true conversions per case on positive Accuracy criteria, almost exclusively those that require correct named entities (drug names, anatomic sites, ICD-style descriptors) where the merge restores base-model lexical breadth. (iii) The same merge step does not move negative-points criteria in either direction on the spine cohort, but it lifts the negative-trigger rate by 0.4 pp on the obesity cohort, consistent with the cohort-conditional comprehensiveness-vs-overhead split documented in Appendix E.2.

What this ablation does and does not show. This ablation supports the design choice of GRPO with a weight-space merge as the production recipe, and it supports DELLA-Linear as the chosen merge filter for both families. It does not show that merging is uniformly beneficial: on the 4B family it is a per-axis re-allocation rather than a clean win, and the choice to retain it is a deployment-surface decision rather than a benchmark-score decision. We also do not vary the GRPO group size G or the merge density ρ in this table; those variations are smaller in magnitude than the configuration-level effects shown here and do not change the ordering.

C Extended ablations

This appendix gives the protocols and full results behind Table 3. Unless stated otherwise, every comparison holds the benchmark items, the Grounded Judge (Qwen3-32B), the grading prompt, and

the scoring rule r_{rub} fixed, so that only the factor under study varies. Confidence intervals are paired bootstrap intervals over items.

C.1 Merge isolation

A merge could raise the score for two different reasons. It could restore generic base-model capability that any reasonable task vector would bring back, or it could transmit something specific that the GRPO update learned. To separate these we fixed the DELLA-Linear configuration of Appendix B.9 (density 0.15, $\epsilon = 0.02$, unit weight) and varied only the task vector merged into Qwen3-8B. The three task vectors are the GRPO update used by *CLR-voyance*-8B, the SFT update from the SFT ablation of Appendix B.12, and a random permutation of the GRPO task vector's entries, which keeps its magnitude distribution exactly but destroys its structure. All arms were graded on the same 358 paired items with identical judge and rubrics. Merging the base with itself returns the base by construction.

Only the GRPO task vector improves the base (0.763 vs. 0.682, $+0.081$ [$+0.047$, $+0.115$]). The SFT task vector (0.108) and the norm-matched random vector (0.232) both fall far below the base. If the merge gain were generic restoration, the random vector, which carries the same magnitudes, would recover a similar score. It does not, so the merge transmits the structure of the RL update. This agrees with the observation of Yuan et al. [2026] that an RL update concentrates in a small, high-magnitude subset of weights, which magnitude-filtered merges keep and rescale. Across filters the merged Qwen3-8B stays within 0.7 pp (Table 8), so the gain does not depend on one lucky configuration.

The two base families respond differently because they start from different places. Qwen3-8B is already reasoning-post-trained, so GRPO amplifies existing reasoning behaviour and the merge has a broad effect on content criteria that require the correct named entity, such as a drug or an anatomic site. MedGemma-4B is instruction-tuned without reasoning training and starts from a low base, and its merge moves the content axes ($+7.48$ Accuracy, $+11.32$ Completeness) while giving back conversational points (Table 9).

C.2 Reward components

The GRPO reward is $r_{\text{rub}} + r_{\text{format}} + r_{\text{tag}}$ at unit weights. GRPO normalises each rollout's reward against its group of G rollouts for the same prompt, so a term that takes the same value on every rollout contributes zero advantage and zero gradient. The training logs of the run that produced the reported Qwen3-8B GRPO checkpoint show that the format term has a within-group standard deviation of 0.000 from the first step, because the base already emits the required reasoning envelope, and that the tag term saturates by step 148 of 3,218, under 5% of training. For over 95% of training the rubric term is the only component with non-zero within-group variance, so it drives every update. The per-axis gains of GRPO over the base sit on the three content axes that only the rubric grades (Completeness $+9.20$, ContextAwareness $+3.32$, Accuracy $+2.64$, Table 7).

To test the rubric term directly, we re-ran MedGemma-4B GRPO with the rubric reward removed and everything else identical, and evaluated it on the same test sample as the base and the full-reward model. The scaffolding-only model scores 27.5, against 22.9 for the base and 53.6 for the full-reward model on this sample (mean per-item scores, a different aggregation from the pooled 27.01 base of Figure 2), so the per-case rubric term supplies nearly all of the GRPO gain.

C.3 Future-conditioned, past-only, and static rubrics

Constructions. On a fixed slice of 250 test admissions we compared three rubrics for the same items. The *future-conditioned* rubric is the paper's. The *past-only* rubric is regenerated by the same oracle with the identical prompt, exemplars, and schema, but with the future withheld. The *static generic* rubric is a fixed, patient-independent rubric of six criteria, including a negative criterion for imprecise information. The stored responses of *CLR-voyance*-8B, GPT-5, and the Qwen3-8B base were re-graded under each rubric.

Both per-case constructions use the same number of criteria at the median, so score differences are not a criteria-count artifact. Without the future, the oracle writes patient-specific safety criteria about

half as often (0.45% versus 0.8% of criteria) and invents 489 one-off themes across 977 rubrics instead of reusing the seven canonical ones.

Why agreement with physicians, not raw scores. Each rubric has its own leniency, so a construction that scores everyone higher is not thereby better. We follow the meta-evaluation used for reward models and LLM judges [Lambert et al., 2024, Tan et al., 2024] and measure how often a construction ranks the physician-preferred response higher. For each of the 422 decided R1 dyads, in which physicians gave blinded preferences with no rubric shown, we built all three rubrics for the case and graded both responses. This agreement is invariant to leniency. Table 10 reports it overall and per comparison pair, together with the per-construction scores. Cohen's κ is near zero for all three constructions for the same reason as the human–human A/B κ in §5, since near-unanimous preferences inflate chance agreement, so agreement accuracy is the informative statistic here.

Table 10: Agreement with the physician-preferred response in 422 decided R1 dyads, overall and on the GPT-5 vs. *CLR-voyance*-8B pairs, and mean scores with our margin over GPT-5 on the 974-item slice.

	Agreement with physicians		Scores on the slice			
Rubric	Overall	GPT-5 vs. ours	Ours	GPT-5	Base	Ours − GPT-5
Future-conditioned	**69.4**	**62.8**	0.724	0.680	0.660	$+0.044$ [$+0.025$, $+0.063$]
Past-only	66.9	57.3	0.883	0.730	0.755	$+0.153$ [$+0.129$, $+0.177$]
Static generic	63.8	58.5	0.744	0.803	0.809	-0.059 [-0.088, -0.031]

The future-conditioned rubric agrees with physicians most and is the best-agreeing construction in every comparison pair, including GPT-5 vs. ours. Both per-case constructions agree more than the patient-independent one, so patient-specific grounding drives the agreement and conditioning on the future is its strongest form.

What the past-only rubric gets wrong. Under the past-only rubric every model scores higher and our margin over GPT-5 more than triples. Without the future the oracle cannot tell which action was decisive, so it writes generic coverage criteria that a vague answer satisfies as readily as a correct one, and it can reward the wrong answer. On a patient with anaemia and coagulopathy after an attempted endoscopic cyst-gastrostomy, the reference answer is chronic pancreatitis with nutritional deficiencies. The future-conditioned rubric rewards exactly that diagnosis, while the past-only rubric rewards identifying "a hematologic disorder" and marks the correct answer wrong. Across the slice, the future-conditioned rubric names a specific clinical entity that the past-only rubric omits entirely on 18% of cases.

What the static rubric gets wrong. A patient-independent rubric cannot check whether a response gives the correct answer for the patient. On an anaemia case whose reference answer is upper and lower endoscopy for occult gastrointestinal bleeding, the static rubric gives full marks to a base response that names no procedure and near zero to our response that names the endoscopy. Its negative imprecise-information criterion penalises detail whether or not the detail is correct, firing on 23.8% of the more detailed responses, which is why it ranks the untrained base above both trained models.

C.4 Is the advantage a length artifact?

Our responses are longer than GPT-5's, about 8.3K against 3.0K characters on average (§4), and in the spine cohort the longer response won most decisive R1 verdicts (Appendix D.9). A judge that rewards coverage could therefore favour our model simply because it writes more. We examined this on the 7,929 items of the automated benchmark that both models answered.

The most direct check compares the two models only where they wrote about the same amount. On the 454 items where our response is within 25% of GPT-5's length, our model still scores higher (0.775 against 0.752) and wins 72.4% of the comparisons that are not ties.

If length were driving the score, our win rate should also rise as our responses grow longer relative to GPT-5's. It does not. Whether our response is slightly longer, up to twice as long, or more than twice as long, we win between 72% and 78% of decisive comparisons.

Finally, our longest answers are not our best-scoring ones. Our score has no correlation with our own response length ($r = 0.00$), whereas GPT-5's score rises slightly with its length ($r = 0.15$). The advantage is therefore present when length is held fixed and does not depend on verbosity.

C.5 Is the advantage a judge artifact?

The Grounded Judge and the base of *CLR-voyance*-8B both come from the Qwen family, so the judge might favour responses written in a style it finds familiar. To test this, we re-graded the stored responses of *CLR-voyance*-8B, GPT-5, and the Qwen3-8B base on the full test set with MedGemma-27B-text-it, a judge from a different model family, while keeping the rubrics, grading prompt, and scoring rule fixed.

The ranking does not change. Our model scores 0.691, GPT-5 0.672, and the base 0.598. Our paired margin over GPT-5 is 1.8 points (95% CI 1.3 to 2.4, 7,929 items), and our margin over the base is 9.2 points (95% CI 8.6 to 9.9, 7,946 items).

The two judges weigh the comparisons somewhat differently. MedGemma-27B gives a larger gain over the base than Qwen3-32B does and a smaller margin over GPT-5, so the statement that holds under both graders is that training clearly improves the 8B model over its base and keeps it ahead of GPT-5. This test changes only the grading step, since the rubrics were still written by Qwen3-32B. Replacing the rubric-authoring teacher would require regenerating the training rubrics and retraining, which we leave to future work. For the grading step itself, Table 5 (right) shows that judges from four model families agree with physicians at similar rates.

C.6 Deployment termination telemetry

Over six months of production use (§6), the GRPO-only checkpoint terminated cleanly on 98.89% of completions. The remaining 1.1% were degenerate over-long outputs, and clinicians rejected all of them. The merged model terminated cleanly on 100% of completions, which is the practical reason the merge is part of the deployed recipe.

D Clinician Validation: Extended Protocol and Results

This appendix provides the implementation-neutral specification of the user study summarised in §5. All formulas, annotation schemas, and analysis steps are reproduced here so that the protocol is reproducible without access to the underlying application.

D.1 Rater pool and cohort construction

Raters. Four board-certified physicians participated, organised into two specialty cohorts of two raters each. The *spine/orthopaedic* cohort comprised two orthopaedic surgeons, one senior (attending) and one junior (resident/fellow). The *general-medicine/obesity* cohort comprised two family-medicine clinicians, also one senior and one junior. All four raters were clinically active at the time of annotation and had no prior exposure to the model outputs under evaluation. Raters were instructed to skip any case they felt under-qualified to annotate. The study was conducted as part of an on-site research collaboration with the partner network.

Case cohorts. From the *CLR-POMDP*held-out split, two case cohorts were constructed by keyword filtering over the patient timeline text (diagnosis descriptions, procedure descriptions, and clinical note content). The *spine* cohort was filtered with the keyword set $\{\text{spine}, \text{orthopaedic}, \text{orthopedic}\}$. The *obesity* cohort was filtered with the keyword set $\{\text{obesity}, \text{obese}, \text{BMI}, \text{bariatric}, \text{weight loss}\}$. Each cohort was then capped to a comparable annotation budget per cohort. The same cohort cases are used for both R2–R3 and R1, so that every case produces paired rubric-alignment and pairwise-preference data joinable at the case level.

Coverage across action spaces. Because a single admission in *CLR-POMDP* contributes one question per action space (§3.1), each cohort's cases span *Diagnosis Assistance*, *Treatment Recommendations*, *Procedural Decision Making*, and *Responding under Uncertainty*, ensuring that the per-axis and per-action-space breakdowns reported in Tables 11–12 are populated for both cohorts.

D.2 Annotation interface

All three rounds are delivered through a purpose-built web application. Annotations are persisted to a central store so that drafts can be resumed, and every DOM-level edit event (toggles, drag-reorders, inline edits, rationale keystrokes) is accumulated as an `interaction_count` effort proxy per submission.

Shared elements. Every page presents the patient's past timeline as event cards in chronological order, an on-screen "Clinical Question" panel placed at the end of the past, a *Future Outcomes* panel clearly demarcated as *"not visible to models"*, and the oracle-authored reference response with an amber disclaimer stating it is AI-generated with full access to the future and is a *reference only, not a candidate to grade*. Any case can be marked invalid at any time via a modal that requires a free-text justification; invalid submissions are stored but excluded from aggregate quality metrics.

R2–R3 interface (rubrics). A three-pane layout: (i) patient context and timeline, (ii) reference response plus model response (the latter hidden during Step 1), (iii) a rubric console with per-criterion cards. Step 1 (Figure 3) presents each criterion as an editable card with inline actions: toggle SUITABLE / NOT SUITABLE, edit description, move up/down, delete (only for physician-added criteria), reset to oracle wording (only for modified criteria), and a bottom button to add a new criterion. A submit guard prevents advancing to Step 2 unless every criterion has an explicit SUITABLE/NOT SUITABLE verdict. Step 2 (Figure 4) reveals the candidate model response, locks the rubric, and presents each criterion with ACCURATE/NOT ACCURATE radios plus a short free-text rationale; raters may traverse back to Step 1 to refine the rubric before submitting. Exactly one candidate response is graded per case in R2–R3; model identity is not shown.

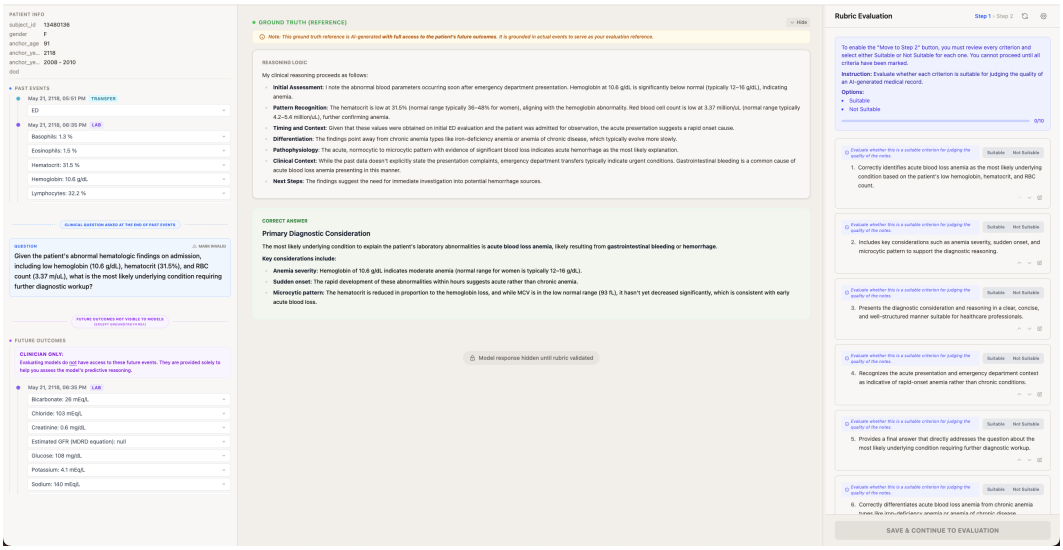

Figure 3: R2 (rubric editing) annotation interface. Left: patient timeline and clinical question. Right: per-criterion cards with SUITABLE/NOT SUITABLE toggles, in-place editing, drag-reorder, delete, and an "add criterion" affordance at the bottom. The model response is hidden in this step so that rubric edits cannot be motivated by the candidate output.

R1 interface (A/B). A four-pane layout (Figure 5): (i) patient context and timeline, (ii) reference response, (iii) response A, (iv) response B. A sticky decision bar offers {*A is Better*, *Tie*, *B is Better*} (keyboard shortcuts 1 and 2). For each case, three model pairs are shown sequentially via a pill-style pair selector; the rater must complete all three before advancing. For each pair the client randomises left/right placement with probability $1/2$ and persists the mapping {`actualModelA`, `actualModelB`, `displayedAsA`, `displayedAsB`}; a post-hoc position-bias check uses this mapping. Model names are removed from the UI at both the component and the serialisation layer to prevent leakage.

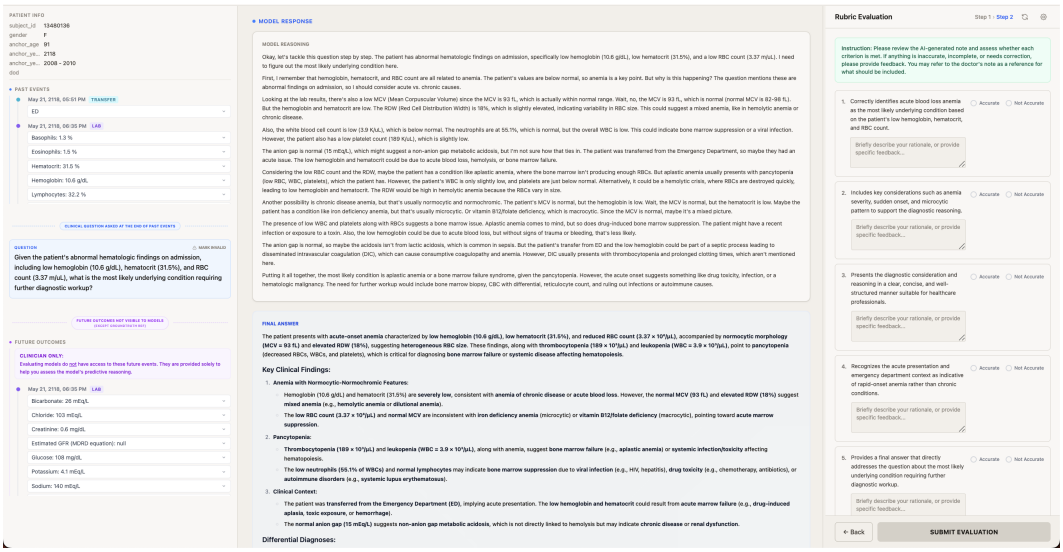

Figure 4: R3 (rubric grading) annotation interface. The candidate model response is now revealed and the rubric is locked; each criterion gets an ACCURATE/NOT ACCURATE radio plus a short free-text rationale. A back-arrow returns to Step 1 if the rater wants to refine the rubric in light of the response.

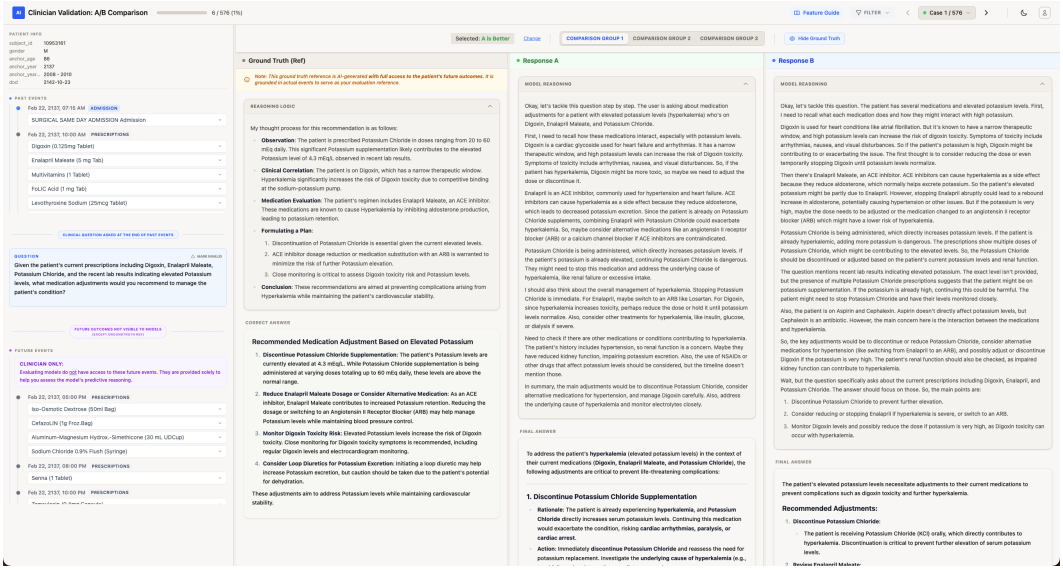

Figure 5: R1 (blinded A/B preference) annotation interface. Top: patient context and clinical question. Below: side-by-side anonymised candidate responses with the reference response visible above (clearly labelled as oracle output, not a candidate). The pair-selector pill row at the top of the response panel rotates between the three model pairs; left/right placement is randomised per (case, pair) and the mapping is persisted for the position-bias diagnostic.

Audit fields persisted per submission. For every case and every submission type the store captures the rater identifier, experiment identifier, patient identifier, sample identifier, submission type (`clinical_reasoning` for rubric judgement or `ab_clinical_reasoning` for A/B), the type-specific payload (criterion verdicts or pairwise choice), `is_invalid` and `invalid_reason`, `results_metadata.interaction_count`, `is_draft`, and server-side timestamps for each attempt. Drafts are upserted; a final submission promotes the draft in place.

D.3 Verbatim rater instructions

The following instructions are reproduced verbatim from the physician guide, lightly elided only to remove UI-navigation details that are obvious from Appendix D.2.

Study framing.

> *There are 2 key things we want clinician validation for our research and development: (1) The Judgement Criteria — This will be the first phase of the clinician validation process. Since our entire reward generation and judgement process is automated, we want to collect data on how aligned the judgement criteria is compared to actual clinicians evaluating the model response. (2) The Model Performance — Next, we want to supplement our automated testing process with a blind A/B test where we present our model response compared to other frontier model responses and let clinicians select the "better" response.*

Task setting.

> *We have taken completed patient journeys and split them into a past and a future. The goal is to answer a complex clinical reasoning query at the point when the past timeline ends. Practically, this aims to emulate an in-patient setting, with the future timeline serving as a ground truth of actions taken to ensure a successful discharge.*

R2–R3 rater prompt.

> *Think of this step as designing an exam for the model. You are the examiner and your goal is to do the following: (1) Set the correct question to ask the student (model) — evaluate whether the question is valid to ask given the past timeline. This means whether answering the question provides crucial insights into further actions, follow-ups or interventions for the patient. (2) Set the correct evaluation rubrics — if the question is valid, how would you evaluate a response for the question? Do the different criteria for evaluation align with how you would judge the response of a student responding to the question? (3) Grade the response — given a response by a model, use the rubrics you defined to grade the model.*

R2 operating rules (rubric edits).

> *Each card represents one specific fact or clinical reasoning step that must be present in the answer. You should mark it as suitable if the criterion is suitable for judging the quality of the output, else mark it as unsuitable. Note that this is not just the accuracy of the response, but also the presentation, formatting, context awareness and any other expectation you may have when reviewing a response from the model.*
>
> *Add missing Criteria in the evaluation rubrics. When to use: You need to evaluate a factor that isn't in the current rubric but important for judging a response to the question. Guidelines: (a) Atomic: Add separate criteria for different concepts (e.g., reasoning steps, formatting). (b) Yes/No: Criteria must be answerable by "Accurate" or "Not Accurate" based on the model's output. (c) Prioritize: Drag and reorder your added criteria to rank them by importance.*

R3 grading rule (including the negative-criterion rule).

> *In Step 2, review the AI's output: Reasoning: Internal thought process. Should be clinically sound. Final Answer: User-facing response. Should be safe and well-formatted. Evaluate both aspects. Do not grade the reference answer in green, only the model response in blue! For each criterion, mark if the model was Accurate or Not Accurate. Use the Reasoning Box to explain why the model failed (e.g., "Hallucinated lab value"). **Note: For negative criteria (e.g., "Model does NOT mention X"), mark Accurate if the statement is true (i.e., X is missing).***

R1 rater prompt.

> *Review the Model A and Model B response. Reasoning: Internal clinical thought process. Answer: Final user-facing response. Ensure the reasoning is sound and the answer is well-formatted. Consider Accuracy, Factuality, Faithfulness, Informativeness, Context Awareness, Formatting and Presentation of findings when selecting the best response. Also consider*

which reasoning is more sound, which final answer is generally better. Look for nuances like missing lab values, considerations for patient conditions or dangerous interactions. Choose the better response. If both are equally good (or equally bad), select "It's a Tie". Toggle between the 3 comparison groups to compare 3 different model pairs. Ensure that you compare all 3 before moving on to the next case.

D.4 Analysis-engine specification

Physician annotations are ingested by an analysis engine that produces a single machine-generated report per run. The engine consumes (i) the per-case rubric with its original oracle wording and point weights, (ii) the per-rater R2–R3 submissions, (iii) the per-rater R1 submissions, and (iv) the oracle's own stored verdict vector and scalar score for each case. It emits a markdown report and a machine-readable JSON artefact. All metrics below are computed on pooled annotations unless stated otherwise; senior and junior results are additionally reported separately.

Rubric-level quality metrics (R2). Let $\mathcal{K}$ be the full set of criteria after the rater has finished Step 1 on a case; let $\mathcal{K}^{\text{orig}} \subseteq \mathcal{K}$ be the oracle-authored subset, and let $\mathcal{K}^{\text{new}} = \mathcal{K} \setminus \mathcal{K}^{\text{orig}}$ be the rater-added criteria. Each criterion carries flags `is_new`, `is_modified`, `not_relevant` and metadata {`axis`, `description`, `points`, `order`}.

$$\begin{aligned}
\text{Query validity rate} &= 1 - \frac{|\{\text{cases marked invalid}\}|}{|\{\text{total cases}\}|} \\
\text{Criterion relevance rate} &= 1 - \frac{|\{k \in \mathcal{K}^{\text{orig}} : \texttt{not_relevant}\}|}{|\mathcal{K}^{\text{orig}}|} \\
\text{Modification rate} &= \frac{|\{k \in \mathcal{K}^{\text{orig}} : \texttt{is_modified}\}|}{|\mathcal{K}^{\text{orig}}|} \\
\text{Addition rate} &= \frac{|\mathcal{K}^{\text{new}}|}{|\mathcal{K}^{\text{orig}}|} \\
\text{Axis coverage}_a &= \mathbf{1}\big[\exists k \in \mathcal{K} : \texttt{axis}(k) = a\big], \quad a \in \text{axes}.
\end{aligned}$$

Per-axis versions of relevance, modification, and addition are reported alongside the pooled totals (Appendix D.6).

Inter-rater reliability (R3). Let $v_k^s \in \{0,1\}$ and $v_k^j \in \{0,1\}$ denote the binary ACCURATE/NOT ACCURATE verdicts of the senior and junior surgeon on criterion k, restricted to criteria retained as SUITABLE. Over all (case, criterion) pairs we report

[leftmargin=*,itemsep=1pt]

- **Percent agreement**: $\text{PA} = \frac{1}{N}\sum_k \mathbf{1}[v_k^s = v_k^j]$.
- **Cohen's κ**: $\kappa = (P_o - P_e)/(1 - P_e)$ with $P_o = \text{PA}$ and P_e the product of marginal rates for each value.
- **Fleiss' κ**: multi-rater generalisation of the above with equal-sized judge panels per item.
- **Krippendorff's α**: the nominal-metric variant of Krippendorff's α over the same binary verdicts, robust to missing raters per item.

Judge–clinician alignment (R3, oracle vs. clinician). Let v_k^o be the oracle's binary verdict on the clinician-aligned criterion k and v_k^c be the pooled clinician verdict (majority; ties broken conservatively as NOT ACCURATE). Define the confusion matrix

$$\text{TP} = \sum_{k: v_k^c=1, v_k^o=1} 1, \quad \text{FP} = \sum_{k: v_k^c=0, v_k^o=1} 1, \quad \text{FN} = \sum_{k: v_k^c=1, v_k^o=0} 1, \quad \text{TN} = \sum_{k: v_k^c=0, v_k^o=0} 1.$$

From this we compute accuracy, precision, recall, F_1, and balanced F_1 (mean of pass-F_1 and fail-F_1); and Cohen's κ between v^o and v^c. We additionally report the per-axis F_1 for each of the five axes.

Score-level alignment (R3). Per case d we compute the *normalised scalar score* both from the oracle and from the clinician using the scoring rule of §3.2:

$$r_{\text{rub}}^{(\cdot)}(d) = \text{clip}_{[0,1]}\Big(\frac{\sum_{k:\, v_k^{(\cdot)}=1} p_k}{\sum_{k:\, p_k>0} p_k}\Big),$$

and then

$$\text{MAE} = \tfrac{1}{N}\sum_{d}\left|r^{o}_{\text{rub}}(d) - r^{c}_{\text{rub}}(d)\right|, \qquad r_{\text{Pearson}}, \rho_{\text{Spearman}} \text{ over } \{r^{o}_{\text{rub}}(d)\}_d, \{r^{c}_{\text{rub}}(d)\}_d.$$

A/B preference metrics (R1). For each pair (m_1, m_2) let w_{m_1}, w_{m_2}, t be the pooled counts of "m_1 better", "m_2 better", and "tie" across raters and cases. We report:

$$\text{Win rate}(m_1 \text{ vs. } m_2) = \frac{w_{m_1}}{w_{m_1}+w_{m_2}+t}, \quad \text{Tie rate} = \frac{t}{w_{m_1}+w_{m_2}+t},$$
$$\text{Position-bias rate} = \frac{|\{\text{cases where left pane preferred}\}|}{|\{\text{non-tie cases}\}|},$$
$$\text{Senior-vs-junior concordance} = \frac{|\{\text{cases where the two raters agree}\}|}{|\{\text{cases both rated}\}|}.$$

Uncertainty quantification. All pooled rates are reported with a two-sided bootstrap 95% confidence interval computed with 1,000 resamples of cases (with replacement). For pairwise preference and judge–clinician accuracy, we additionally report a one-sided binomial test of the null "arm m_1 is not preferred over arm m_2" and "oracle and clinician are independent" respectively, with Bonferroni correction across the three A/B arms.

Analysis report contents. The auto-generated report contains, at minimum: (i) per-case diagnostic rows with raw annotations, (ii) pooled R2–R3 tables, (iii) the full confusion matrix (v^o, v^c) with per-axis breakdowns, (iv) the pooled A/B win/tie/loss matrix with CIs, (v) senior-vs-junior splits for every metric, (vi) disagreement case studies with model outputs, clinician rationales, and the specific criteria that disagreed. The JSON artefact contains the same data in a machine-consumable form.

D.5 Three reuses of the clinician-annotated set

The R1–R3 annotation pass is structurally one artefact (per-criterion verdicts $\times$ pairwise preferences across two cohorts of two clinicians each), reused for three structurally distinct downstream purposes that we deliberately keep separate.

Use 1: ICL task adaptation of the rubric-author and per-criterion judge LLMs. Clinician-edited rubrics from R2 (criteria added, deleted, reworded, reweighted) are inserted as in-context exemplars into the rubric-authoring oracle prompt; the clinician verdict vectors v^c from R3 are inserted as in-context exemplars into the grading-judge prompt. We do *not* fine-tune the judge weights (no SFT, no DPO): with on the order of $\sim$100 cases per cohort, the gradient step is too noisy to risk freezing one cohort's stylistic preferences into the judge weights, while the in-context exemplars remain re-applicable across model generations and across cohorts. The headline ICL alignment numbers are reported in Table 5 of the main paper; the full per-cohort breakdown including the negative ICL result on both cohorts is in Table 13.

Use 2: clinical LLM-as-a-judge and clinical preference-model alignment analysis. The R2–R3 per-criterion verdicts (paired with the case, the rubric, and the candidate response) provide a held-out alignment surface against which off-the-shelf clinical LLM-as-a-judge configurations are scored on accuracy, F_1, Cohen's κ against v^c, and on score-MAE of the per-case scalar score (Table 5). Symmetrically, the R1 pairwise preferences provide a held-out alignment surface for clinical preference-model configurations, scored on 3-way and A/B accuracy and rank-correlation against the unanimous-consensus clinician preference (Table 4). The resulting comparisons complement the static-jury approach of MedHELM [Bedi et al., 2025] by characterising how off-the-shelf judge configurations align with the clinician signal across cohorts.

Use 3: blinded A/B comparison of *CLR-voyance* against state-of-the-art baselines. R1 doubles as the head-to-head reported in Table 4: *CLR-voyance*-8B vs. GPT-5 (proprietary frontier), *CLR-voyance*-8B vs. HuatuoGPT-o1-7B (open-source medical-reasoning), and *CLR-voyance*-8B vs. Qwen3-8B base (within-family anchor). This use is the deployment-relevance signal and is the comparison reviewers and clinicians both want answered: when asked to choose, do clinicians prefer our model's response over the strongest proprietary and the strongest open-source-medical alternatives?

Table 11: Per-axis R2–R3 rubric quality + judge–clinician alignment, aggregated across both cohorts. *Relevance* is $1 - \frac{|\{\text{not_relevant}\}|}{|\text{AI-gen criteria}|}$. *Modif.* and *Added* are normalised by the same denominator. The right-hand columns report the headline judge (Qwen3-32B, zero-shot) per-axis alignment vs. the pooled clinician verdict on the (criterion, candidate-response) pairs of Step 2.

Axis	n criteria	Relevance	Modif.	Added	Judge F_1 (bal.)	Judge κ
Accuracy	561	84.3%	2.9%	5.3%	0.764	0.529
Completeness	493	88.8%	2.4%	0.0%	0.729	0.459
CommunicationQuality	523	93.3%	0.0%	0.0%	0.619	0.241
ContextAwareness	492	88.0%	2.0%	0.0%	0.686	0.373
InstructionFollowing	427	91.6%	0.0%	0.0%	0.694	0.388

D.6 Extended R2–R3 results

The per-axis pattern is informative. Clinicians edited only the two clinically actionable axes (Accuracy gets 2.9% modification + 5.3% addition; Completeness 2.4% modification) and did not touch the conversational axes (CommunicationQuality, InstructionFollowing — both 0.0% on edits, only marking individual criteria as not relevant). The judge–clinician alignment shows the symmetric pattern: highest agreement on Accuracy (κ=0.53, F_1=0.76) where the clinician verdict has the strongest factual anchor, lowest on CommunicationQuality (κ=0.24) where the clinician verdict is more stylistic and varies between cohorts.

Table 12: R2–R3 results stratified by cohort (the within-cohort rater pair is identified in Appendix D.1); "Aggregate" pools both cohorts. The judge–clinician block reports the headline judge (Qwen3-32B, zero-shot) on each cohort's evaluable rater–criterion pairs.

Metric	Spine	Obesity	Aggregate
n rater-annotated cases	103	127	230
n unique patients	37	69	101
n rater–criterion pairs	1,136	1,360	2,496
Rubric quality			
Criterion relevance rate	83.1%	94.0%	89.1%
Modification rate	3.4%	0.0%	1.5%
Addition rate	2.7%	0.0%	1.2%
Inter-rater reliability			
Cohen's κ (engine)	0.749	0.658	0.702
Judge–clinician alignment (Qwen3-32B, zero-shot)			
n_{eval}	1,138	1,359	2,497
Accuracy	77.6%	75.9%	76.7%
F_1 (balanced)	0.719	0.701	0.709
Cohen's κ	0.439	0.402	0.419
Score MAE	0.124	0.192	0.160
Score Pearson r	0.571	0.434	0.478

Table 13: Per-cohort clinical LLM-as-a-judge benchmark on R2–R3 verdicts. Same models / configurations as Table 5 of the main paper; the aggregate column there is the row-by-row pool of these two cohorts. The 5-shot ICL variant is only applicable to the two open-weights models. Best-per-column highlighted in **bold**.

		Spine cohort				Obesity cohort			
Judge	**Mode**	n	Acc	F_1	κ	n	Acc	F_1	κ
Qwen3-32B	zero-shot	1,138	77.6%	0.719	0.439	1,359	75.9%	0.701	0.402
Qwen3-32B	5-shot ICL	1,132	77.7%	0.723	0.447	1,353	75.8%	0.700	0.400
MedGemma-27B-text-it	zero-shot	1,138	77.7%	0.715	0.431	1,360	77.1%	0.707	0.414
MedGemma-27B-text-it	5-shot ICL	1,133	**78.8%**	0.713	0.431	1,355	76.8%	0.685	0.372
GPT-5	zero-shot	1,138	71.5%	0.676	0.357	1,360	76.1%	**0.731**	**0.472**
GPT-4.1	zero-shot	1,135	76.0%	0.694	0.389	1,360	**77.6%**	0.718	0.436

The same cohort-asymmetry that drives the R1 preference-model leaderboard inversion (Table 4) shows up here: *which* judge is best depends on the cohort. MedGemma-27B-text-it edges Qwen3-32B by 1.0 pp on the spine cohort with 5-shot ICL (78.8% vs. 77.7%), while GPT-5 jumps from worst-on-spine (71.5%, κ=0.36) to best-on-obesity (κ=0.47). The pooled column in the main-paper table averages these effects out, but the per-cohort leaderboard exposes that any single judge configuration carries cohort-specific blind spots; this asymmetry is one reason the deployed judge composes Qwen3-32B verdicts with the clinician-edited rubric rather than relying on a single monolithic judge.

ICL is not a reliable lever on this R2–R3 surface either. On the open-weights models, 5-shot ICL produces -0.2 to $+1.1$ pp on accuracy and small mixed effects on κ. The same pattern surfaces in the clinical preference-model alignment analysis (Appendix D.5, Use 2): with on the order of $\sim$100 clinician-consensus dyads per cohort, the few-shot demonstration set is too small to reliably steer a 27B/32B judge's verdict on a clinical-reasoning surface where the underlying clinical-judgement variance between cohorts is itself larger than the ICL signal. We therefore report only the zero-shot configurations as the headline (Table 5).

D.7 Extended R1 results

This subsection reproduces the per-pair, per-rater full report from the analysis engine for both cohorts. Raters within each cohort are anonymised as R_1, R_2 to protect identity; the labelling is alphabetical on the original anonymised username string, fixed across all subsections so that any cross-table comparison is consistent.

Per-rater preference breakdown (consensus subset). For each cohort and each rater we report the rater's decisive preference rate for *CLR-voyance*-8B against the column-arm baseline, computed on the consensus subset (dyads where both raters agreed; ties excluded from the denominator). The pooled column is the cohort win rate already reported in Table 4.

Table 14: R1 preference rate of *CLR-voyance*-8B over each column-arm baseline, stratified by rater within each cohort. Decisive verdicts only; consensus subset.

Cohort	Comparator	Rater A rate (n)	Rater B rate (n)	Pooled rate
Spine	GPT-5	100.0% (24)	100.0% (24)	100.0%
Spine	HuatuoGPT-o1-7B	100.0% (30)	100.0% (30)	100.0%
Spine	Qwen3-8B (base)	96.0% (25)	96.0% (25)	96.0%
Obesity	GPT-5	37.0% (27)	37.0% (27)	37.0%
Obesity	HuatuoGPT-o1-7B	98.4% (61)	98.4% (61)	98.4%
Obesity	Qwen3-8B (base)	72.9% (48)	72.9% (48)	72.9%

By construction of the unanimous-consensus rule, the per-rater decisive rate within each (cohort, comparator) row equals the pooled rate (both raters picked the same arm on every dyad in the subset). The numbers therefore differ from the all-decisive aggregation reported in Table 4 of the main paper, which is computed on the strictly larger superset of every decisive verdict (including the non-consensus dyads where the two raters disagreed). The divergence between raters that drives the cohort-level κ value in Table 16 is concentrated in those excluded buckets, summarised in Table 15 below.

Per-pair consensus bucket distribution. The unanimous-consensus rule used in Table 4 discards dyads on which the two raters disagreed on the 3-way {a, b, tie} decision. Table 15 reports the bucket distribution per pair so the consensus subset can be sized: *strict-consensus* = both raters agreed on a decisive verdict; *semi-consensus* = both rated, one decisive and one tie, agreeing on the dominant arm; *single-rater* = only one rater submitted a verdict (these are excluded from κ but retained for the pooled win rate); *non-consensus* = both rated and disagreed on the decisive arm.

Inter-rater agreement. Table 16 reports the cohort-level inter-rater statistics on the *full* A/B verdict set (i.e., before consensus filtering, on dyads rated by both clinicians). The 3-way percent agreement and Cohen's κ are the reference for the divergence in interpretive variance between cohorts that the main paper flags.

Table 15: R1 dyad bucket counts per (cohort, pair). The decisive-consensus column is the row used by Table 4.

Cohort	Pair	Total	Strict	Semi	Single	Non-cons.
Spine	GPT-5 vs. ours	86	24	1	53	8
Spine	HuatuoGPT vs. ours	95	30	0	60	5
Spine	Qwen3 vs. ours	87	25	1	55	6
Obesity	GPT-5 vs. ours	108	27	1	11	69
Obesity	HuatuoGPT vs. ours	109	61	1	9	38
Obesity	Qwen3 vs. ours	109	48	1	11	49

Table 16: Inter-rater agreement on the union set of dual-rated dyads. *3-way* treats $\{a, b, \text{tie}\}$ as the target; *A/B-only* drops tie rows from either side. Both metrics are computed cohort-wide; the per-pair Cohen's κ values are noisy at this sample size and are reported in the engine JSON for completeness.

Cohort	n dual-rated	PA 3-way	κ 3-way	κ A/B-only
Spine	101	78.2%	0.034	0.100
Obesity	97	47.4%	0.051	0.047

The spine-cohort row is the textbook Cohen's κ paradox: percent agreement is high (78.2%) but κ is near zero because both raters concentrate their decisive verdicts on the same arm of every pair (*CLR-voyance*-8B), so the chance-agreement baseline P_e is itself close to the observed agreement and the normalisation collapses. The obesity row is a different regime, with much lower percent agreement (47.4%); the verdict mass is genuinely split between arms on that cohort, and the low κ here reflects real disagreement rather than a marginal artefact. The R1 headline win-rates (Table 4) are computed on the unanimous-consensus subset and so are not affected by the disagreed-on dyads counted here.

Position-bias diagnostics. The client randomises left/right placement per (case, pair); under the null the left pane should be preferred on ∼50% of decisive verdicts. Per-rater rates and the cohort-pooled rate (with Wilson 95% CI and a one-sided exact-binomial p-value against 0.5) are reported in Table 17. Neither cohort nor any individual rater shows a significant deviation, confirming the randomisation.

Table 17: Display-A win rate on decisive verdicts. Significance is two-sided exact binomial against $H_0 = 0.5$; the position-bias diagnostic from Table 4 is the pooled row.

Cohort	Rater	Display-A rate	n decisive	p vs. 0.5
Spine	A	53.2%	79	0.65
Spine	B	51.9%	79	0.83
Spine	Pooled	52.5%	158	0.58
Obesity	A	53.7%	136	0.44
Obesity	B	55.9%	136	0.21
Obesity	Pooled	54.8%	272	0.13

Length-bias diagnostic. For each decisive consensus dyad we ask whether the rater preferred the longer of the two responses. Under the null this should also be ∼50%. The pooled and per-pair length-bias rates are reported in Table 18. The spine cohort shows a strong length-favouring effect (98.7% pooled, $p < 10^{-43}$), which we read as a confound between length and our model's identity (*CLR-voyance*-8B produces more exhaustive enumerations of clinical content; cf. the per-axis Completeness lift in §4). On the obesity cohort the length effect is much more pair-dependent, in particular the GPT-5 vs. *CLR-voyance* pair has a 37.0% longer-side rate (the shorter response wins more often), which is one mechanical reason that pair scores at parity in Table 4.

Rater-effort and engagement diagnostics. Decision-time medians and quartiles on the per-case timeline give a first-order proxy for annotation effort. Spine raters median 13–18 s per case (with a

Table 18: R1 length-bias diagnostic. Each row is the fraction of decisive verdicts on which the longer of the two responses was preferred. *Overall* pools all three pairs.

Cohort	Pair	Longer-side wins	n decisive
Spine	GPT-5 vs. ours	100.0%	48
Spine	HuatuoGPT vs. ours	100.0%	60
Spine	Qwen3 vs. ours	96.0%	50
Spine	*Overall*	98.7%	158
Obesity	GPT-5 vs. ours	37.0%	54
Obesity	HuatuoGPT vs. ours	98.4%	122
Obesity	Qwen3 vs. ours	75.0%	96
Obesity	*Overall*	77.9%	272

long tail to 90–300 s on harder cases); obesity raters median 35–55 s per case, consistent with the lower 3-way agreement and the more contested verdict mix in that cohort. *Self-consistency* is the rate at which a rater issued the same verdict on a re-shown dyad (only Rater B in each cohort was shown re-display; that rater issued one revision per cohort, both flips).

Table 19: R1 effort and engagement diagnostics, anonymised per cohort.

Diagnostic	Spine A	Spine B	Obesity A	Obesity B
n unique verdicts	79	79	138	138
Tie rate	0.0%	0.0%	1.4%	1.4%
Median decision time (s)	13.2	18.5	54.9	35.5
P75 decision time (s)	40.8	57.0	95.3	92.0
P90 decision time (s)	93.8	303.5	152.4	195.3
Revisions on re-displayed dyads	0	1	2	1

D.8 Clinical preference model benchmark: per-pair and position-bias diagnostics

This subsection extends the cohort-level results in Table 4 with the per-pair breakdown, the preference-model-side position-bias diagnostic, and the ICL negative result. All five reported preference models are graded against the two-clinician unanimous consensus on the R1 dyads.

Per-pair A/B-only accuracy. Table 20 reports each preference model's two-class (A/B) accuracy stratified by the three model pairs in the study. The most informative cell is the *GPT-5 vs. CLR-voyance-8B* column on the spine cohort: the clinician consensus picks *CLR-voyance*-8B on 85.6% of decisive verdicts, but every off-the-shelf preference model except MedGemma-27B (40.9%) scores below 15% on this pair. GPT-5 in particular picks GPT-5 on *every* decisive verdict (0.0% alignment), the same self-preference effect that has been documented for general-domain LLM-as-a-judge tasks [Panickssery et al., 2024]. On the obesity cohort the same column flattens (every preference model lands in $72{-}80\%$, including GPT-5 itself), reflecting the underlying clinician parity between *CLR-voyance* and GPT-5 on family-medicine cases.

Preference-model-side position bias. The benchmark randomises display A/B placement per (case, pair) under a deterministic seed; under the null each preference model should pick display-A on $\sim$50% of decisive verdicts. Four of the five preference models sit within a $\sim$8 pp band of 0.5 on both cohorts; the only outlier is GPT-4.1-mini on the obesity cohort (67.4% display-A win rate), which already exhibits a stable display-A prior in the zero-shot setting and explains its lower obesity ranking despite competitive content judgement.

Negative result on 5-shot ICL. We also evaluated 5-shot in-context-learning variants of the two open-weight preference models (Qwen3-32B-ICL, MedGemma-27B-ICL), where each demonstration was a held-out clinician-consensus dyad with the gold winner label. Demonstrations were balanced across the displayed A/B positions so the gold-letter distribution is equalised within the few-shot block. Neither variant produced a robust improvement over its zero-shot configuration on the cross-cohort aggregate. We read this as the few-shot demonstration set being too small (on the order of

Table 20: Per-pair A/B-only accuracy of preference models against clinician consensus. "*CLR-voyance*" is short for *CLR-voyance*-8B, the trained model in this paper.

Cohort	Preference model	GPT-5 vs. *CLR-voyance*	HuatuoGPT vs. *CLR-voyance*	Qwen3 vs. *CLR-voyance*
Spine	Qwen3-32B	9.1%	96.4%	79.2%
Spine	MedGemma-27B-text-it	40.9%	100.0%	87.5%
Spine	GPT-4.1-mini	13.6%	75.0%	70.8%
Spine	GPT-4.1	9.1%	88.9%	78.3%
Spine	GPT-5	0.0%	71.4%	50.0%
Obesity	Qwen3-32B	72.0%	96.6%	83.0%
Obesity	MedGemma-27B-text-it	76.0%	100.0%	74.5%
Obesity	GPT-4.1-mini	80.0%	82.8%	76.6%
Obesity	GPT-4.1	80.0%	92.9%	87.0%
Obesity	GPT-5	72.0%	67.2%	76.7%

Table 21: Position-bias diagnostic on the preference-model side. Display-A win rate is the fraction of decisive verdicts on which the preference model picked the response shown on display position A.

Cohort	Preference model	Display-A rate	n decisive	Tie rate
Spine	Qwen3-32B	48.6%	74	0.0%
Spine	MedGemma-27B-text-it	57.9%	76	0.0%
Spine	GPT-4.1-mini	63.5%	74	0.0%
Spine	GPT-4.1	52.8%	72	2.7%
Spine	GPT-5	54.1%	74	0.0%
Obesity	Qwen3-32B	55.6%	133	0.0%
Obesity	MedGemma-27B-text-it	54.9%	133	0.0%
Obesity	GPT-4.1-mini	67.4%	132	0.8%
Obesity	GPT-4.1	57.4%	129	3.0%
Obesity	GPT-5	53.9%	128	1.5%

a few dozen consensus dyads per cohort) to reliably steer a 27B/32B judge on a clinical-reasoning surface where the underlying clinical-judgement variance between cohorts is itself larger than the ICL signal. We therefore report only the zero-shot configurations in Table 4.

D.9 Limitations

Observed care is not guaranteed optimal. The future trajectories that condition each rubric record what a care team did, which can reflect local practice, patient preference, or a suboptimal decision. We bound this by keeping only admissions with no re-admission or death within two years, so every conditioning trajectory reached a successful discharge and sustained recovery, but this does not make each logged decision optimal. The reward is not behaviour cloning, since the policy never sees the logged action and is scored on past-verifiable reasoning criteria, and negative-point criteria penalise unsafe proposals regardless of the log. Physicians accepted 89.1% of criteria unchanged and marked no case invalid. Some noise from imperfect trajectories may remain, which is why the paper describes the rubrics as grounded in successful care trajectories rather than in optimal care.

Single-turn optimisation. The task is a POMDP, but the optimiser is single-turn response-level RL that neither maintains a belief state nor plans across turns. Multi-turn optimisation over the same past/future construction is the natural next step.

Clinician study scope. The study is scoped to two inpatient specialty cohorts (spine/orthopaedic and general-medicine/obesity); generalisation to other specialties should be retested with appropriately credentialed raters. In future iterations we plan to widen the cohort design, expand the rater pool, and decompose R1 ties via a follow-up question asked only when the rater selects *Tie*.

Length-vs-identity confound in R1. Longer responses won 88.9% of decisive R1 verdicts in the spine cohort and 64.3% in the obesity cohort. A fraction of the blinded preference signal is therefore recoverable to verbosity rather than to clinical content. Length-controlled win rates are reported in

Appendix D.7, the automated benchmark is length-stratified in Appendix C.4, and the merged model retains the directional preference under that control on every comparison except GPT-5 in obesity, where the parity finding is unchanged.

Parameter scale. Our 8B parameter scale is set by the on-premise GPU concurrency, latency, and reliability SLAs at the partner network (§6). Whether the per-case rubric pipeline yields proportional gains on substantially larger or smaller bases is unverified. A same-family GRPO-only ladder (for example Qwen3-4B, 8B, and 14B without merging) would test this directly, but one stable GRPO run with the judge in the loop takes on the order of 4,000 H200 GPU-hours, so we instead show generality across two very different bases and ablate the recipe on both (Appendix C).

Teacher sensitivity. The same model family (Qwen3-32B) writes the rubrics and grades the responses. A cross-family re-grade with MedGemma-27B preserves the ranking (Appendix C.5) and judge–physician agreement is indistinguishable across four judge families, but swapping the rubric-authoring teacher would require regenerating the training rubrics and retraining, which we leave to future work.

Merge-step behaviour across base families. The merge-step contribution on the open-ended *CLR-POMDP* surface is broad-spectrum on the Qwen3 family and primarily a re-balancing across axes on the MedGemma-4B family (§4). The MedGemma-4B family also shows a larger relative lift on the calibration-style external benchmarks than on *CLR-POMDP* (§4.1). Characterising how this base-family-conditional behaviour varies with model scale and base-family pre-training distribution is a direction for future work.

E Verbatim clinician qualitative feedback and recurring failure modes

This appendix walks the full R2–R3 clinician-annotation export of §5: 230 rater-annotated cases, 2,498 rater–criterion pairs, and 387 free-text rationales across the four raters split into the spine and obesity cohorts. Raters are referred to throughout as **Spine A**, **Spine B**, **Obesity A**, and **Obesity B** to match the convention already used in Appendix D.7. Subject identifiers below are de-identified MIMIC-IV admission keys and no protected health information appears in this appendix. The annotation interface exposes criterion ordering, criterion text, criterion addition, criterion removal (`not_relevant`), and per-criterion verdicts (`verified`) to the clinician, but does not expose criterion point weights, so the analysis below treats the clinician-implied importance signal as the reorder distribution rather than as a numerical re-weight.

Summary. The 387 free-text rationales explain the spine–obesity gap of Table 4. The spine cohort values exhaustive enumeration (*"missing something is medical negligence"*), while the obesity cohort wants answer-first, resource-aware recommendations, which argues for persona-aware training and clinician-controllable response behaviour. Raters also flagged recurring clinical-judgement gaps, namely under-weighting of advanced age, omitting previously prescribed medications when proposing additions (a drug–drug interaction hazard), and over-inference of acute cardiac events. These motivate drug-interaction and age-aware rubric criteria and a faithfulness reward anchored to the past timeline (Appendices E and E.6).

E.1 Four rater profiles

The four raters span a recognisable clinician-annotation spectrum. **Spine A** is a heavy editor: across 52 cases they modified 33 criteria and added 30 new ones, with at least one rubric-side edit on 44 of 52 cases and at least one new criterion on 17 of 52. Their rationales are short and surgical, typically a single clause that anchors the verdict to a specific past-timeline event. **Spine B** is a light editor: 5 modifications, no additions, 51 criteria marked not relevant, and 69 rationales that read as concise evidence pointers ("targets also provided", "BMP not mentioned", "fall risk identified"). **Obesity A** is the verbose commentator: zero modifications and zero additions, but 275 of the 387 free-text rationales (71%) and every cohort-level CommunicationQuality observation in the export. Their rationales are long, pattern-tagged, and reproducible across cases. **Obesity B** is a minimal commentator: zero modifications, zero additions, one rationale across 676 criteria, and reliance on the binary `verified` field to deliver judgement. The R2–R3 free-text rationale corpus is therefore

effectively an Obesity-A-plus-Spine-B-plus-Spine-A pooled signal, with Obesity B contributing the binary-verdict baseline.

E.2 Cohort-philosophy split, evidenced four ways

The single most consistent finding across the four raters is a philosophy split on what a good model response ought to look like. The split is evidenced four independent ways. The **edit volume** differs: spine raters edit aggressively (85% of Spine-A cases include an edit; 43% of Spine-B cases include a not-relevant marking), while obesity raters never modify or add criteria. The **reorder priority** differs: spine raters move newly-added criteria to the leading positions on 11.7% of cases, while 96.1% of obesity cases simply lead with Accuracy. The **per-axis verified-rate** differs: spine raters score the model 20.2 pp lower on Accuracy and 13.1 pp lower on ContextAwareness than obesity raters do; the obesity cohort scores the model 24.3 pp lower on CommunicationQuality. The **rationale themes** differ: every "verbose monologue", "fabricated content", "Chinese intrusion", "tangent beyond scope", and "antibiotic over-use" rationale comes from the obesity cohort, and spine raters never tag these as failure modes.

The qualitative narrative behind these splits is captured in the spine-side and obesity-side raters' own words on the same model behaviour. Spine-A frames exhaustive enumeration as a desired CDSS property:

> *"The lengthy ones covered most differential diagnosis as well, in addition to the definite diagnosis. It listed treatment for all differential diagnosis, rather than just one, which is good for a clinical decision support tool — then the clinician can choose what to consider. . . . Did you find the models giving overly aggressive recommendations at any point? It tries to be comprehensive. Because missing something is medical negligence, and the doctor can face charges. So good for the model to be comprehensive. Most of the CDSS in use are like this, as the doctor starts typing, possibilities, diagnosis and treatment options pop up. But nothing is benchmark yet. So the field is evolving — now with LLMs this could get refined."*

Obesity-A reads the same property as overhead:

> *"Most responses answered the question, although main point was often after multiple paragraphs of explanation. I would prefer the answer to be given directly at the start. Most models also give detailed responses (e.g., a detailed management plan when prompted). Responses for management / treatment plans tend to be all-rounded (i.e., consider all possibilities). In real clinical setting, I believe it's not the wisest to send the patient for as many tests as possible due to resource constraint. Instead, the most crucial diagnostics tests can be highlighted with the other low-yield tests be suggested as optional."*

Read together, neither cohort is right and the other wrong. Both cohorts give a coherent and internally-consistent quality signal on the same model behaviour. The per-case rubric machinery captures both signals separately rather than collapsing them, which is the entire point of grading the model against per-case rubrics rather than a static checklist. The dichotomy is the qualitative form of the cohort-level R1 split (Table 4): *CLR-voyance*-8B wins decisively in the spine cohort that values comprehensiveness, and is at parity with GPT-5 in the obesity cohort that prefers an answer-first style.

E.3 Recurring per-case failure modes

Five recurring failure modes account for the bulk of the obesity-cohort negative verdicts and for the model-side gaps the spine cohort captured by adding new criteria. All are documented with verbatim rationales below.

1. Under-weighting of advanced patient age. This is the only mixed-result theme that appears in both cohorts. Sixty-six of the 387 rationales reference age (eleven from spine, fifty-five from obesity), and only three percent of the criteria carrying these rationales are marked MET. Verbatim from Spine A on subject `12605254`: *"Patient is 91 yrs old, and this should have been explicitly considered by the model. However, i don't see it."* On subject `12152963` the same rater notes: *"Patients age is 91,*

but the model identifies it as 45", where the age miss co-occurs with the contamination-frame failure documented in §E.2. Spine B records *"advanced age not referenced"* as a typical short-form pointer.

2. Failure to surface previously-prescribed medications when proposing additions (DDI hazard). Twenty-seven rationales reference previously-prescribed medications or DDI risk; eighty-five percent come from Obesity A. Verbatim: *"model does not mention Aspirin EC or Clopidogrel as antiplatelet agents from the prescriptions; lists different medications instead"*, *"model recommends switching to Ticagrelor as primary addition rather than recognizing Metoprolol as the new beta-blocker addition"*, *"model mentions constipation risk from opioids but does not explicitly recommend management with stool softeners or laxatives"*. The candidate satisfies the underlying criterion in 14.8% of these cases.

3. Vagueness on antibiotic class or drug name. Spine B records both directions: *"ceftriaxone suggested"* when the model is specific, but on cases where the model defaults to a class label the same rater is terse. Obesity A supplies the longer-form complaint: *"Does not mention the specific 14-day antibiotic course duration."* Antibiotic over-use itself is a separate obesity-cohort theme: ten rationales flag broad-spectrum or sepsis-coverage recommendations on patients without a documented infection source.

4. Fabricated content on the obesity cohort. Thirty-seven rationales document content the model emits without past-timeline support. The single most reproducible signature is *"45-year-old male inferior MI ST elevation"*, which appears in fourteen Obesity-A cases (versus 74- or 70-year-old female patients in the actual past). Other documented fabrications include *"Indomethacin as cause of GI bleeding when not in past prescriptions"*, *"hypertension and obesity history not in past records"*, fabricated arterial-blood-gas values that flip the diagnosis from respiratory acidosis to alkalosis, and a fabricated *Ceftriaxone-Simvastatin* interaction that is not clinically established. None of these rationales accompany a verified-true criterion, identifying these as pure model-side failure modes rather than as legitimate clinical-judgement disagreements.

5. Stylistic and language drift on the obesity cohort. Three sub-patterns recur. (i) *Verbose internal-monologue leakage*: the `<think>` block content with conversational filler ("Okay lets tackle", "Wait", "Hmm") leaks into the user-facing answer on 34.8% of Obesity-A cases. (ii) *Tangents beyond reference scope*: the response covers the reference's recommendations and adds extensive workup tangents (*"adds tangential discussion of CTPA, V/Q scan, and echocardiogram beyond the question scope"*; *"tangents into HLA-B27, RF, CCP, uric acid, blood cultures, DEXA, peripheral neuropathy workup"*). (iii) *Chinese-character intrusion*: ten Obesity-A cases contain Chinese ideograms mid-text (*"Chinese characters for splenomegaly"*, *"Chinese character for prolong"*), an artefact of the Qwen3-8B base's pre-training distribution.

Positive cases. The same raters also recorded explicitly positive verdicts. Spine B, on subject `10661376`: *"Follow-up timeline provided was optimal and useful for clinicians"* on a question about post-operative monitoring; the same response received a content-level critique that the model's choice to lead with DVT / PE prophylaxis was not the most clinically salient first step in a young, low-risk post-operative patient. Spine A, on a case with the candidate's structured response: *"While the direct answer would be point 3, the rest of the response provides a more comprehensive narrative"* (InstructionFollowing, MET). These cases are a useful reminder that within-cohort within-axis preferences differ even when the rubric-level verdict is unambiguous.

E.4 Rubric-author failure modes flagged in R2–R3

R2–R3 also surfaced two rubric-author failure modes that we report transparently. *Future-timeline leakage*: a small number of oracle-authored rubric criteria are written in a way that quotes a future event verbatim, which can penalise an entirely reasonable past-only answer. The clinician edits from R2 are inserted as ICL exemplars into the rubric-author prompt (Appendix D.5, Use 1), tightening this constraint on subsequent rubric-author calls. *Duplicate criteria*: the only two duplicate-criterion rationales in the entire export are from Spine B (subject `11384537`, criterion noted as a repeated question), and both are marked MET on the candidate, so the duplication produces a denominator-inflation artefact rather than a content failure. The rubric-author post-processing now includes a de-duplication pass that mechanically catches this case.

E.5 Quantitative cross-references

The qualitative themes above admit clean cross-references to the quantitative tables of §4 and §5. The **Completeness** $+21.4$ **pp** lift over GPT-5 (Figure 2) is the quantitative form of the spine-cohort "CDSS comprehensiveness" observation. The -3.4 **pp CommunicationQuality and** -1.0 **pp InstructionFollowing** deltas vs. GPT-5 are consistent with the obesity-cohort flags on answer-shape, vague antibiotic class names, and verbose-monologue patterns. The **cohort-level R1 split** (spine 85.6% vs. obesity 48.0% against GPT-5; Table 4) is the quantitative form of the comprehensiveness-vs-overhead dichotomy. The **length-aligned win rate** (88.9% spine, 64.3% obesity; Appendix D.7) is the quantitative form of the same length-as-content confound that the obesity cohort treats as a failure mode (over-aggressive recommendations) and the spine cohort treats as a feature (defensive coverage). The **negative-points trigger rate** of 1.30% vs. 1.79% for GPT-5 is the quantitative envelope on the safety-side concerns: the rubric's negative-points criteria explicitly cover spurious diagnostic confidence and unsupported aggressive treatment, and the merged model trips them less often, even though specific cases of over-inference still occur.

E.6 Comprehensive analysis of the raw R2–R3 rubric-edit data

The R2–R3 "rubrics_final" export is the canonical clinician-curation artefact for this paper. It contains 230 rater-annotated cases (101 unique patients), $2{,}498$ rater–criterion pairs, and 387 free-text rationales across the four raters. The remainder of this appendix unpacks four analyses on the raw export that the headline tables in §5 and Appendix D.6 collapse: what clinicians did to the AI-generated rubric, where the negative-points criteria fired, which recurring failure modes the rationales surface against the candidate model, and how those qualitative findings line up with the cohort-stratified verification rates.

Edits made to the AI-generated rubric

Three patterns dominate the edit log. **Reordering is essentially universal.** The oracle-authored rubric uses a deliberate five-axis cycle (Accuracy, Completeness, ContextAwareness, CommunicationQuality, InstructionFollowing and back), intended to keep every axis covered before depth-first descent. Clinicians override this layout almost universally: 63.1% of spine cases lead with Accuracy after Step 1, 14.6% with Completeness, 11.7% with a newly-added criterion, and the remaining 10.6% with one of the other three axes; 96.1% of obesity cases lead with Accuracy. Both cohorts move the two clinical-content axes to the top of the ranked rubric, and the spine cohort additionally promotes new content-axis criteria into the leading positions on a non-trivial fraction of cases.

Modifications and additions are exclusively a spine-cohort phenomenon, and are dominated by Spine A. All 38 modifications and all 30 additions come from the spine cohort, with 33 of the 38 modifications (87%) and 30 of the 30 additions (100%) attributable to Spine A. The five Spine-B modifications cluster around a single pharmacology theme (tricyclic-antidepressant management of musculoskeletal pain after an initial NSAID trial, plus one chest-X-ray dating modification), and are too narrow a sample to read as a generalisable rater preference. The thirty Spine-A additions span three reproducible thematic groups: examination-first additions (eight criteria of the form "calls for clinical examination as the first step" / "suggest examination of the spine and neurological evaluation"), post-surgical / post-operative awareness additions (seven criteria such as "Recognises the postoperative status of the lumbar spine", "Mentions prophylaxis for DVT", and "Mentions the need for the comprehensive post-operative pain management"), and structural / response-format additions (three criteria of the form "Provides structured response with points, including rationale for each management strategy"). Twenty-five of the thirty additions (83%) were verified-true on the candidate response that prompted the addition, suggesting these are oracle-rubric coverage gaps for spine-cohort cases rather than candidate-model failures.

Edits concentrate on the actionable axes. Every modification sits on Accuracy, Completeness, or ContextAwareness; CommunicationQuality and InstructionFollowing receive zero modifications and zero additions across both cohorts. Spine raters mark Accuracy criteria as not relevant on 27.4% of cases versus 8.1% in the obesity cohort, with similar three-to-four-fold ratios on Completeness and ContextAwareness; the conversational-axis not-relevant rates are similar across cohorts.

Numerical reweighting of points is absent by construction. Across all $2{,}498$ rater-criterion pairs there is no recorded change in criterion point value. The annotation UI does not expose point editing

to clinicians, so this absence is a measurement-platform property rather than a finding about clinician indifference. Inferring rater-implied reweighting from the reorder signal alone is a possible post-hoc analysis, but the current pipeline treats the reorder distribution as the importance signal.

Cohort split on edit volume, restated. The two cohorts edit very differently. The spine cohort modifies 3.4% of AI-generated criteria, adds 2.7%, and marks 16.9% as not relevant; the obesity cohort modifies 0.0%, adds 0.0%, and marks only 6.0% as not relevant. We read this as specialty conservatism on the obesity side (clinicians treat the AI rubric as a screening list and use the binary verified field to deliver judgement) and willingness to rewrite on the spine side (clinicians actively rewrite a knowledge-domain criterion they consider clinically incorrect for this patient).

Negative-points (safety) criteria

The rubric template includes negative-point criteria designed to penalise unsafe behaviour (spurious diagnostic confidence, recommendations contraindicated by the patient context, etc.). Seventy-one such criteria appear across the 230 annotations. Clinicians keep 27 (38%) as relevant after Step 1, and on the candidate response the negative criteria trigger (clinician marked verified-true on a negative criterion, indicating the bad behaviour was detected) 14 times. The axis breakdown is Completeness 4, CommunicationQuality 4, ContextAwareness 5, Accuracy 1, InstructionFollowing 0. The zero InstructionFollowing triggers are consistent with the model reliably emitting the required `<think>`-then-answer envelope. The ContextAwareness and Completeness triggers map cleanly onto the omission and missed-workup buckets discussed in §E.2.

Recurring failure modes documented in the rationales

Of the 387 rationales, 343 (89%) come from the obesity cohort and specifically from Obesity A; the remaining 44 are predominantly short Spine B evidence pointers and a handful of Spine A contextual notes. Obesity B contributed exactly one rationale across 676 criteria. We tag each rationale with regular-expression patterns and report the per-pattern frequency together with the per-cohort split and the verified-rate of the criteria the rationales accompany. Patterns at 0% verified-rate identify pure model-side failure modes (the clinician marked the criterion as not-MET on every case where the pattern fired); intermediate verified-rates identify issues the model gets right some of the time.

Table 22: Failure-mode patterns surfaced by the 387 R2–R3 free-text rationales. *n* is the number of rationales matching the regex pattern (one rationale can match multiple patterns); *verified* is the clinician's MET-rate on those rationales' criteria. Patterns with 0% verified-rate are uniformly model-side failures. Per-cohort split: *S* = spine, *O* = obesity.

Pattern	n	S	O	verified-rate
Pure model-side failures (verified=0%)				
verbose internal-monologue leakage ("Okay lets tackle", "Wait", "Hmm")	54	0	54	0%
fabricated content (45-year-old male, drug, lab)	37	0	37	0%
cardiac over-inference (STEMI / PCI / thrombolysis)	27	0	27	0%
tangents beyond reference scope	12	0	12	0%
paediatric / pregnancy mis-framing	11	0	11	0%
antibiotic over-use	10	0	10	0%
Chinese-character intrusion	10	0	10	0%
lab-value misuse / ignored	3	0	3	0%
vague / generic phrasing	3	0	3	0%
Mixed-result patterns				
imaging recommendations (relevance, modality, timing)	149	41	108	18.8%
under-weighting of advanced patient age	66	11	55	3.0%
positive acknowledgement / "covers all"	31	6	25	12.9%
previously-prescribed meds / DDI hazards	27	3	24	14.8%
risk consideration	23	5	18	13.0%
post-op / surgical context	18	8	10	38.9%
incomplete workup	15	3	12	13.3%
differential-diagnosis enumeration	8	3	5	25.0%
dose / titration specifics	7	1	6	14.3%
duplicate / repeated criterion	2	2	0	100%

The pure-model-side-failure rows are the strongest qualitative finding of the R2–R3 study. They identify specific behaviours of the *CLR-voyance*-8B candidate response, not generic clinical-judgement disagreements, and every one of them is an obesity-cohort observation. Verbose internal-monologue leakage is the largest at 54 rationales (twenty-two of them carrying the exact phrasing *"verbose internal monologue with Okay lets tackle; not concise"*), corresponding to 34.8% of all Obesity-A cases. The 37 fabricated-content rationales are dominated by the *"45-year-old male inferior MI ST elevation"* pattern that the model emits even when the patient is a 70- or 74-year-old female, and which co-occurs with the 27 cardiac-over-inference rationales (STEMI / PCI / thrombolysis framing on patients without supporting past-timeline data). Both signatures point at a single underlying memorised case in the Qwen3-8B base. Chinese-character intrusion is benign for content but reproducible at 15.2% of Obesity-A cases. Tangents beyond reference scope (CTPA, V/Q, HLA-B27, DEXA) are the qualitative form of the comprehensiveness-vs-overhead dichotomy: the spine cohort never tags this behaviour as a failure mode, treating the same prose as defensive coverage.

The mixed-result rows surface clinical-judgement issues the model gets right some of the time. The strongest signal is on previously-prescribed-meds / DDI hazards (n=27, 14.8% MET): the model proposes new medications without surfacing the patient's existing prescription list, which the clinician explicitly flags as a DDI hazard particularly for opioids. Under-weighting of advanced patient age (n=66, 3.0% MET) is the only mixed-result theme that appears in both cohorts, with verbatim Spine A evidence (*"Patient is 91 yrs old, and this should have been explicitly considered by the model"*) joining the larger Obesity-A corpus.

Cross-reference back to the verified-rate matrix

Per Table 12, the per-axis verified-rates split as follows. Spine cohort: CommunicationQuality 91.3%, InstructionFollowing 84.3%, Completeness 67.1%, ContextAwareness 58.9%, Accuracy 55.6%. Obesity cohort: Accuracy 75.8%, InstructionFollowing 76.5%, Completeness 74.9%, ContextAwareness 72.0%, CommunicationQuality 67.0%. The CommunicationQuality cell flips sharply between cohorts (a 24.3 pp swing), and the rationale-pattern data above is the qualitative form of that swing: the obesity cohort penalises the model on CommunicationQuality for the verbose-monologue, tangent, and fabrication patterns, while the spine cohort accepts the same prose style as professionally-styled exposition. Conversely, the spine cohort penalises the model harder on Accuracy (-20.2 pp vs. obesity), which matches their willingness to rewrite an Accuracy criterion when they consider the clinical detail wrong. The two cohorts therefore deliver different quality signals on the same model behaviour, and the per-case rubric machinery preserves both signals separately rather than collapsing them.

F Hospital deployment: extended three-stage rollout

Stage 1: Pilot trial (5 months, 3 rounds). A structured note-drafting trial built on a summarisation pipeline [Nagar et al., 2025] ran across three rounds with a pilot cohort of physicians at the partner network, who defined their own note and summary-template elements, drafted candidate notes, and imported approved notes into their hospital information system (HIS). The goal was to characterise day-to-day drafting requirements and isolate failure modes by source (input data context, prompts, or model capability), initially using a model fine-tuned on the partner network's internal note corpus.

Stage 2: The reasoning-element gap. Once input-context and clinician-customisable prompt issues were resolved, an SFT + DPO model trained on the internal note corpus captured drafting style and formatting reliably, but consistently failed on what clinicians designated as *reasoning-heavy* elements even after substantial prompt engineering. Two failure modes recurred: (i) *complex instruction-following*, e.g., *Chief Complaint* requires traversing a clinician-defined data-source priority list (consult the latest ED note; if absent, the OPD note within the last X weeks; etc.); and (ii) *clinical-reasoning-heavy sections* (*Impressions*, *Assessment*, *Review of Systems*, *Plan*) that integrate partial-observable signals with comorbidities and intended next actions, the regime the upstream pipeline of this paper (§3) is designed to supervise.

Stage 3: *CLR-voyance* in production. The per-case adaptive rubric (§3.2) was designed precisely to capture quantitative physician feedback on these reasoning-heavy elements; *CLR-voyance*-8B was trained against that signal, and after positive trial feedback was deployed network-wide. In

production, clinicians can edit and maintain their own note templates, optionally assign individual elements of the template to the reasoning model, and accept or edit the result before saving into the HIS. On-premise GPU compute together with concurrency, latency, and reliability SLAs make *CLR-voyance*-8B the largest model the system can support without violating those constraints, a hard infrastructure ceiling that motivates the 8B parameter scale of the headline model.

G Qualitative Case Studies from *CLR-POMDP*

This appendix shows four representative items from the *CLR-POMDP*test split, graded by the Grounded Judge against the rubric generated for that item. We select cases from the four canonical action spaces, each scored under a rubric of ten criteria spanning all five axes. The excerpts are drawn from our *CLR-voyance*-8B (Qwen3-8B + GRPO + DELLA-Linear merge) and are intended to convey two properties of the benchmark that raw numbers do not: (i) the *scale and rawness of the EHR context* the policy must reason over at each step, and (ii) the *granularity* of a patient-specific rubric over a single, long-form clinical response. All prose is verbatim from the evaluation records and truncated only for space.

A.1 Diagnosis Assistance (Condition Identification)

Patient context (paraphrased for space). ED-Observation admission at `2121-02-21 23:39:00`. Total ED stay $\approx$ 38 h; the observation segment is only $\approx$ 1.7 h. During the stay the medication-administration records show administrations of *DiphenhydrAMINE*, *Haloperidol*, *Lorazepam*, *FoLIC Acid*, *Multivitamins*, and *Thiamine*. No ED-level ICD codes are exposed to the model (stripped to avoid label leakage).

Query. *"What is the most appropriate action to determine the correct diagnosis for a patient admitted under the Emergency Department Observation unit for approximately 1.7 hours?"*

Reference answer (summary). Confirm diagnosis via focused ED assessment; order targeted labs/imaging; monitor for deterioration; document rationale for not pursuing inpatient admission. Evidenced by the `Transfers` discharge event at `2121-02-23T07:19:59Z`.

Candidate answer (excerpt, Qwen3-8B GRPO+DELLA+AIM). The candidate integrates the medication pattern into the differential: Haloperidol + Lorazepam + Diphenhydramine $\Rightarrow$ acute agitation, psychosis, or seizure management; Thiamine + FoLIC Acid + Multivitamins $\Rightarrow$ nutritional / metabolic encephalopathy, including explicit consideration of *Wernicke's encephalopathy* in the setting of possible alcohol use. It then proposes short-interval reassessment of vitals and mental status, a CIWA-based withdrawal pathway, toxicology screen, metabolic panel, and explicit criteria for escalation vs. discharge.

Rubric (10 criteria, theme *Context Seeking*).

Axis	Pts	Criterion	Met
ContextAwareness	+7	Mentions the patient's admission via ED Observation unit based on timeline data.	✓
CommunicationQuality	+8	Provides a clear and concise explanation of the appropriate actions.	✓
Completeness	+6	Includes multiple steps for diagnostic evaluation and documentation.	✓
Accuracy	+7	Correctly identifies the need for diagnostic tests based on initial symptoms and vitals.	✓
InstructionFollowing	+5	Bases reasoning solely on the past timeline data.	✓
ContextAwareness	+6	Notes the unusually long ED stay and its diagnostic implications.	×
CommunicationQuality	+7	Uses appropriate clinical terminology for healthcare professionals.	✓
Completeness	+5	Addresses both diagnostic evaluation and documentation requirements.	✓
ContextAwareness	+6	References medication administration patterns as part of the diagnostic approach.	✓
CommunicationQuality	+5	Organizes recommendations logically with numbered points.	✓

Score. $56/62 = \mathbf{0.903}$. The single missed criterion is precisely the one that would have required the model to draw attention to the 38-hour ED dwell time as a diagnostic red flag in its own right, a miss the rubric was *designed* to catch.

A.2 Treatment Recommendations (Medication Suggestions)

Patient context. Urgent Labor & Delivery admission; lab events in the past show Hct $= 34.0\%$ (ref. 36–48), Hb $= 11.7$ g/dL (ref. 12–16), RBC $= 3.85 \times 10^6/\mu$L, WBC $= 12.6 \times 10^3/\mu$L, MCH $= 30.3$, MCV $= 88$, RDW $= 14.2$. Transfer to Obstetrics (postpartum & antepartum).

Query. *"Given the patient's admission to the labor and delivery unit with abnormal hemoglobin and hematocrit levels, what medications would you recommend initiating to manage potential complications and support postpartum recovery?"*

Reference answer (summary). Oxytocin IV; oral ferrous sulfate; Lactated Ringer's; lactation-safe analgesia (acetaminophen). Evidenced by future prescriptions recorded the following morning.

Candidate answer (excerpt). Parses the hematology panel (low Hb/Hct, leukocytosis) into a normocytic-normochromic anemia with concurrent inflammation, enumerates PPH and postpartum endometritis as the leading differentials, and recommends (i) Oxytocin IV infusion, (ii) oral ferrous sulfate, (iii) crystalloid resuscitation with attention to transfusion threshold, (iv) lactation-safe acetaminophen, and (v) empiric broad-spectrum coverage if sepsis is confirmed, with dosing and contraindication notes inline.

Rubric (10 criteria, theme *Health Data Tasks*). Axes covered: Accuracy ($\times 2$, +8/+10), Completeness ($\times 2$, +7/-5), CommunicationQuality ($\times 2$, +6/+7), ContextAwareness ($\times 2$, +9/+8), InstructionFollowing ($\times 2$, +5/+10). The two negatively-signed items were *"mentions appropriate dosing or administration routes for each medication"* (actually positively phrased, judged `true`) and an axis-check on structure. All 10 of 10 criteria met.

Score. $\mathbf{0.929}$. The model's hedging around the NSAID contraindications and the explicit lactation-safety reasoning are exactly what the rubric axis for ContextAwareness was graded on; these are behaviours that rule-based clinical rewards would not register, yet they drive physician preference.

A.3 Responding under Uncertainty (Risk–Benefit Analysis)

Patient context. Post-surgical admission; past-timeline prescriptions include Ketorolac 15–30 mg IV q4h PRN, HYDROmorphone 0.5–1 mg IV q4h PRN with oral 2–4 mg PO/NG q4h PRN, plus acetaminophen 1000 mg PO/NG q4h PRN and ibuprofen 600 mg PO q4h PRN.

Query. *"Given the patient's administration of multiple analgesics including Ketorolac and HYDROmorphone, what clinical strategy should be employed to manage potential adverse effects while ensuring adequate pain control?"*

Reference answer (summary). Monitor renal function; assess opioid use and respiratory-depression risk; adopt a multimodal approach; review Ketorolac duration ($\leq$5 d); plan safe IV-to-oral transition.

Candidate answer (excerpt). The candidate produces a structured plan: (i) explicit risk assessment for COX-mediated renal impairment and GI bleeding, opioid-induced respiratory depression, and NSAID–anticoagulant interactions; (ii) titration protocol for hydromorphone with defined rescue thresholds; (iii) hydration strategy to mitigate nephrotoxicity; (iv) ibuprofen/acetaminophen opioid-sparing ladder; and (v) explicit acknowledgement of irreducible uncertainty about the patient's baseline renal function given limited past data, with a recommendation to obtain a creatinine trend before continued dosing.

Rubric (10 criteria, theme *Health Data Tasks*). Nine of ten criteria met, including Accuracy (+10) for correctly identifying Ketorolac/HYDROmorphone risks, ContextAwareness (+9) for addressing

concurrent-drug interactions, and Completeness (+7) for multimodal strategy. The missed criterion (InstructionFollowing, +5: *"references the patient's prescriptions from the past timeline"*) caught a real and subtle failure: the candidate describes the pharmacology well but under-cites the patient-specific prescription rows.

Score. $\mathbf{0.903}$. This case demonstrates the rubric's ability to distinguish *domain-level* from *patient-grounded* reasoning on an otherwise near-ideal answer, a distinction that rule-based or unanchored-judge rewards routinely collapse.

A.4 Procedural Decision Making (Diagnostic Procedures)

Patient context. Medicine/Cardiology unit under an Ambulatory Observation admission; past-timeline prescriptions include Nitroglycerin SL, Aspirin EC, Clopidogrel, Potassium Chloride, Sodium Chloride 0.9% flush, Simethicone, Acetaminophen, and an influenza vaccine; lab events show a mildly elevated Chloride and Creatinine with Potassium 4.9. HCPCS coding flags cardiovascular services.

Query. *"Based on the patient's admission to the Medicine/Cardiology unit and the prescribed medications, which procedural or diagnostic intervention is warranted to monitor potential cardiac complications?"*

Reference answer (summary). Cardiac enzyme panel with Troponin T; electrocardiogram. Evidenced by a future Troponin T lab event the following morning.

Candidate answer (excerpt). Continuous telemetry with ST-segment and arrhythmia surveillance; serial troponin (including an explicit reference to troponin T as the ACS workhorse assay); electrolyte and renal panel to monitor for Ketorolac/Nitroglycerin-adjacent risk; and bedside ECG with a documented trigger list for escalation. The reasoning explicitly links the Nitroglycerin + Aspirin EC + Clopidogrel trio to the ACS/unstable-angina working diagnosis.

Rubric (10 criteria, theme *Health Data Tasks*). Nine of ten criteria met; the single miss (CommunicationQuality, +7) was a stylistic one: the judge did not credit "clearly differentiates between procedural and diagnostic interventions", which the candidate's structure blurs.

Score. $\mathbf{0.899}$.

A.5 What these cases jointly demonstrate

Three observations cut across the four cases above and motivate the design choices in §3.1–§3.3.

Raw-EHR reasoning, at scale. Each input exposes the model to a full, event-level patient slice (vitals, labs with units and reference ranges, prescriptions with dose/route, transfers, discharge notes, and HCPCS-coded events), rather than a hand-distilled vignette. Candidate responses are themselves long, grounded, Markdown-formatted clinical documents of roughly 1,500–2,000 tokens each, routinely citing specific lab values, drug interactions, and care-unit transitions. This is the shape of reasoning our deployment demands.

Rubric granularity matches clinical judgement. Across the four cases, the items the judge marked `false` were never cosmetic: they were a missed causal link (length of ED stay as diagnostic signal, A.1), a failure to cite patient-specific prescription rows (A.3), and a structural conflation of diagnostic and procedural intervention (A.4). These are exactly the error modes that rule-based metrics (ROUGE, clinical NER overlap) cannot express and that unanchored judge rewards cannot consistently locate.

Outcome-anchored rubrics stay past-verifiable. Although each rubric was authored with access to the future timeline (so that the criteria are *about the right things for this patient*), every criterion is still a pass/fail statement that the judge can evaluate using only the past context plus the candidate response. This is the empirical form of the grounded-judge guarantee laid out in §3.2.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction enumerate the five contributions (CLR-POMDP formulation, outcome-grounded per-case rubrics with the Grounded Judge, clinician validation across two cohorts with reusable judge / preference-model benchmarks, GRPO + weight-space-merge post-training, and production hospital deployment). Each is concretely mapped to a paper section and is supported by quantitative results: §4 for the headline 84.95% CLR-POMDP score and the GPT-5 / MedGemma-27B / HuatuoGPT-o1 comparisons, §4.1 for the orthogonal-benchmark gains, §5 for the clinician validation, and §6 for the deployment evidence.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Limitations are discussed in Appendix D.9, covering that observed care is not guaranteed optimal, single-turn optimisation, teacher sensitivity, the two-cohort scope of the clinician study, the length-vs-identity confound in the blinded preference, the on-premise SLA-driven 8B parameter ceiling, and the base-family-conditional behaviour of the merge step on the 4B family. The qualitative analysis (§5) and Appendix E additionally include specific model failure modes (verbose-monologue leakage, fabricated-content signatures, cardiac over-inference, DDI hazards) flagged by the rater pool.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.

- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

   Justification: The paper does not include theoretical results that require proof. The POMDP framing in §3.1 is a problem-setting definition rather than a theorem, and the rubric reward r_{rub} in §3.2 is a closed-form scoring rule whose definition is reproduced verbatim with all assumptions stated inline.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Appendix B is an implementation-neutral specification of the entire pipeline. It expands the data extraction and timeline-splitting procedure (§B.1), the verbatim query-and-rubric generation prompts and JSON schemas (§B.2, §B.5), the grader prompt (§B.6), every reward function used during GRPO (§B.7), the per-base hyper-parameter table (§B.8), the merge configurations (§B.9), and the evaluation protocol (§B.10). To stress-test reproducibility, an independent researcher not involved in the study used commodity coding agents to replicate the pipeline end-to-end in under a day from the moment MIMIC-IV access was granted, working only from the paper and this appendix (Appendix B "Artefact release statement").

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often

one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.

- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: We are unable to publicly release *CLR-voyance*-specific artefacts (model weights, generated rubrics, candidate responses, curated patient timelines, or clinician annotations) due to non-disclosure agreements and patient-privacy contracts signed with the partner public health network under which the deployment runs (Appendix B "Artefact release statement"). To support review-time verification while honouring those constraints, our experiment code, the curated CLR-POMDP train/test artefact, and the full clinician-annotation export are submitted as anonymised supplementary material visible to reviewers, area chairs, and ethics reviewers but not redistributed publicly. The MIMIC-IV source data is available under PhysioNet credentialing, and Appendix B provides every prompt, hyper-parameter, and configuration needed to reproduce every CLR-voyance artefact from scratch given that access; an independent third-party reproduction confirmed this end-to-end.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The cohort-selection procedure, timeline-splitting rule, and admission-wise train/test split are specified in §3.1 and Appendix B.1. The full per-base hyper-parameter table (learning rate, batch size, group size G, gradient accumulation, max prompt and completion lengths, schedule, warm-up, precision, attention kernel, and distributed strategy) is in Appendix B.8. Reward functions are specified in Appendix B.7 and merge configurations in Appendix B.9. The evaluation protocol with sampling parameters, grading rubric, and aggregation rule is in Appendix B.10.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: The aggregate gap to GPT-5 in §4 is supported by a paired Wilcoxon signed-rank test on matched per-admission scores ($p < 10^{-300}$, n=7,927). All R1 win rates in Table 4 carry Wilson 95% CIs and per-pair p-values are reported in §5 and Appendix D.7 together with Bradley–Terry rank intervals. Inter-rater and judge–clinician agreement are reported using Cohen's κ, Fleiss' κ, and Krippendorff's α as specified in Appendix D.4, and bootstrap 95% CIs (1,000 resamples of cases with replacement) are used for pooled rates. The ablations in §4.2 and Appendix C report paired-bootstrap 95% confidence intervals.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: Per-base GRPO and SFT hyper-parameters, including precision (bf16), attention kernel (FlashAttention-2), and distributed strategy (DeepSpeed ZeRO-3 with vLLM rollouts), are reported in Appendix B.8. The rollout-serving stack (vLLM) is named in §3.3. The deployed hardware envelope (on-premise GPU concurrency, latency, and reliability SLAs) is described in §6 as the constraint that motivates the 8B headline scale; specific cluster identifiers and per-run wall-clock are anonymised at submission time per double-blind review and will be added to the camera-ready version on acceptance.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

Answer: [Yes]

Justification: The research conforms with the NeurIPS Code of Ethics. Patient data is sourced from MIMIC-IV under PhysioNet credentialed access, with subject identifiers de-identified upstream. The clinician validation is conducted with four board-certified physicians under an on-site research collaboration with the partner network. Patient-trajectory artefacts and clinician annotations are not publicly redistributed because of the non-disclosure and privacy contracts with the partner network (Appendix B).

Guidelines:

- The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: The positive impact is the deployable inpatient-reasoning surface evidenced by the production rollout in §6 and the trillion-token-scale clinician usage signal to generate thousands of patients notes daily. The qualitative analysis in §5 and Appendix E explicitly documents specific negative-impact failure modes the model still exhibits (fabricated patient frames, cardiac over-inference, antibiotic over-use, DDI hazard from missing prior-medication awareness, age-under-weighting), and §5 flags two concrete next-step rubric criteria (DDI-check and age-weighted management) and a patient-centric faithfulness reward as mitigations. The deployment is gated behind a clinician edit-and-accept loop into the hospital information system rather than autonomous action, which is the operational mitigation against incorrect-output harm.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [Yes]

    Justification: *CLR-voyance* model weights and patient-trajectory artefacts are not released publicly; access is restricted to the partner-network deployment under a non-disclosure agreement and to anonymised supplementary materials visible only to reviewers (Appendix B). In production, every reasoning-element draft passes through a clinician edit-and-accept gate before being saved to the hospital information system, so the model never autonomously commits to the patient record (§6, Appendix F). The negative-points criteria of the per-case rubric (Appendix B.5) explicitly penalise spurious diagnostic confidence and contraindicated recommendations during training, and the qualitative analysis tracks the model's negative-flag trigger rate as a safety-side regression diagnostic.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All existing assets are cited at first use. Patient data is sourced from MIMIC-IV [Johnson et al., 2023c, 2024, 2023a,b, Goldberger et al., 2000] under PhysioNet credentialed access. Base models include Qwen3-8B [Yang et al., 2025] and MedGemma-4B [Sellergren et al., 2025]; comparator and judge models include GPT-5, GPT-4.1, GPT-4.1-mini, Qwen3-32B, MedGemma-27B [Sellergren et al., 2025], HuatuoGPT-o1-7B [Chen et al., 2024], DeepSeek-R1 variants [Shao et al., 2024], and Gemma-3-4B. Library citations cover PyTorch

[Paszke et al., 2019], DeepSpeed [Rasley et al., 2020], vLLM [Kwon et al., 2023], and HuggingFace TRL [von Werra et al., 2020]. Each asset is used in compliance with its respective license; redistribution of derivative artefacts is governed by the upstream licences plus the partner-network NDA where applicable.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [No]

Justification: New assets (the *CLR-voyance*-8B and *CLR-voyance*-4B model checkpoints, the curated CLR-POMDP train/test artefact, the per-case rubrics, and the clinician-annotation export) are not publicly released because of the non-disclosure agreements and patient-privacy contracts signed with the partner public health network (Appendix B "Artefact release statement"). The artefacts are submitted as anonymised supplementary material for reviewer access, and Appendix B fully documents the construction of every artefact (data pipeline, prompts, hyper-parameters, scoring rules, merge configurations) so that an independent team can recreate them from MIMIC-IV; an independent third-party reproduction confirmed this end-to-end.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [Yes]

Justification: The clinician validation involves four board-certified physicians as study participants. Their verbatim instructions for study framing, task setting, R2 (rubric editing), R3 (rubric grading), and R1 (A/B preference) are reproduced in Appendix D.3. Annotation-interface screenshots for both R2–R3 (rubric editing and grading) and R1 (A/B preference) are provided in Figures 3–5 of Appendix D.2. Raters participated as collaborators on the on-site research programme rather than as crowdsourced workers; no fee-based compensation was provided.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [Yes]

    Justification: The clinician validation and the production hospital deployment were conducted under the partner public health network's institutional approvals and the corresponding research-collaboration agreement (Appendix D.1). Participants are clinically active board-certified physicians annotating de-identified MIMIC-IV admissions; the patient data is already de-identified upstream by PhysioNet, so no additional patient-side risk is incurred during annotation. Specific institutional names, approval reference numbers, and ethics-board correspondence are anonymised at submission per double-blind requirements and will be added to the camera-ready version on acceptance.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: LLMs are core to the methodology and are described in detail. The Grounded Judge is a Qwen3-32B oracle that authors per-case rubrics and grades policy responses (§3.2, Appendix B.5 for the rubric-author prompt and Appendix B.6 for the grader prompt). The post-trained policies are Qwen3-8B and MedGemma-4B (§3.3), and the comparator panel includes GPT-5, GPT-4.1, GPT-4.1-mini, MedGemma-27B, HuatuoGPT-o1-7B, DeepSeek-R1 variants, Gemma-3-4B, and the Qwen3-8B base. LLMs were used only as study artefacts and as comparators. They were not used for the manuscript text in any way that would affect originality.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.