CARE: A CONSULTING, ACQUIRING, RESOURCE-AWARE, EXPLAINING AGENT FOR CLINICAL DECISION SUPPORT

Anonymous authors
Paper under double-blind review

ABSTRACT

Clinical decision support is based on a record in progress: tests are ordered as the diagnosis evolves, each with medical resource consumption, and the diagnosis process is personal per patient condition. We present CARE, a generic framework for this setting. A reinforcement learning agent for predicting medical outcome, trained on the medical center's own patients, works in a human-like process, deciding per patient which test to order next, when to consult a medical language model, and when to stop and predict, under test-specific costs and turnaround times and within resource budget limits. The language model supplies medical knowledge and a learned fusion model weighs its assessment by the evidence behind it and fuses it into the locally trained prediction. Every episode is rendered into a clinical account of what was acquired and what it showed, so a clinician can see how the recommendation was reached. On MIMIC-IV, CARE reaches higher accuracy than dynamic acquisition agents and static test rankings at an earlier diagnosis stage, while keeping its workups individual to the patient. The framework is generic and can be applied to any medical data and any medical outcome, a holistic approach that combines local learning, medical knowledge, resource awareness, and self-explanation in one system.

1 INTRODUCTION

A clinician's decision-making process uses the existing data to decide what further data is needed for the decision, relies on vast medical knowledge, considers the medical center's resources, and is always explainable by the clinician. Decision support for this setting must have the same four properties: informed by medical knowledge, self-explanatory, personalized, and resource-aware.

Decision support systems today meet some of these properties, but none all four. Risk models over the electronic health record presuppose a complete feature vector, yet at admission most of the record does not exist (Escobar et al., 2020; Mahmoudi et al., 2020). Learned acquisition methods choose tests per patient, and recent clinical systems add resource awareness, as priced tests (Yu et al., 2023) or as turnaround time with a language model in the loop (Sun et al., 2024), but they carry no verified explanation of the workup and, where a language model is used, rely on it directly as predictor or policy. Language models carry broad medical knowledge but are not trained on the deploying institution's population and cannot be trusted directly in either role (Singhal et al., 2023; Brown et al., 2025; Bani-Harouni et al., 2026). Explanation methods describe a prediction from a fixed input rather than a sequence of acquisitions (Tonekaboni et al., 2019).

We propose CARE, an approach that gives a holistic solution meeting all four. An acquisition agent trained on the deploying institution's own patients decides per patient which feature to acquire next and when to stop, conditioned on what earlier results showed. Consulting a medical language model is one of the actions it may take: the agent learns when the model's assessment is worth obtaining, and a learned fusion model weighs that assessment by the evidence behind it and fuses it into the local prediction, so medical knowledge enters the decision. Each feature carries its own cost and turnaround time, and the agent is trained under a budget for each with both shaping its reward, and hence is resource-aware. Finally, each episode is rendered into a clinical-language account and verified against the record, so a clinician can see what the system acquired and why it matters.

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

Our contributions are:

- We present a single framework for clinical decision support that is personalized, resource-aware, informed by medical knowledge, and self-explanatory.
- We propose a generic approach to represent features as text, structured or free-text alike, embedded by one clinical encoder, so that various test formats can enter a single acquisition process.
- We model test turnaround time as a per-action duration through a semi-Markov formulation, separate from monetary cost and under its own budget.

On MIMIC-IV, CARE's personalized selections outperform static baselines, and consulting the language model raises accuracy while reducing the number of acquired tests. Cost- and duration-aware training reduce spend and time at negligible loss in accuracy, and the agent's acquisitions come with explanations that pass clinical verification and a faithfulness check.

2 Related Work

Two lines of work bear on this problem: how to decide which evidence to acquire, and how to predict from it. Acquisition methods split into static feature selection, which fixes one subset for the whole population, and dynamic feature acquisition, which chooses per patient based on what is already known. Prediction is done either by a model trained on the institution's own records or by a language model trained on medical text. The following paragraphs review each and our position relative to it.

Dynamic feature acquisition. Dynamic methods select sequentially per instance. Greedy methods pick the next feature by expected information gain (Ma et al., 2019; Covert et al., 2023; Gadgil et al., 2024), while reinforcement-learning methods plan beyond the next step, including when to stop (Shim et al., 2018; Janisch et al., 2019; Li & Oliva, 2021; Valancius et al., 2023). Clinical work follows this line and adds a resource: priced tests (Yu et al., 2023; Kashani et al., 2025) or turnaround time (Sun et al., 2024). A recent direction makes the language model itself the acquisition policy (Sun et al., 2024; Bani-Harouni et al., 2026), which inherits the model's limits as a predictor. This is the domain this paper works in: our agent is a reinforcement-learning policy that consults the language model as one of its actions and learns when its assessment is worth obtaining.

Static feature selection. Static methods rank features on importance or information and keep the top ones for every patient (Guyon & Elisseeff, 2003; Wang et al., 2025), as in the SHAP ranking we use as a baseline.

Prediction models. Locally trained neural networks are the basis of EHR risk models (Rajkomar et al., 2018; Jiang et al., 2023). Language models carry broad clinical knowledge and can be prompted with a patient record to estimate risk (Singhal et al., 2023; Yang et al., 2022) or to write clinical text (Thirunavukarasu et al., 2023; Van Veen et al., 2024), but they trail locally trained models in discrimination and calibration, and their validation on EHR data remains thin (Brown et al., 2025; Lehman et al., 2023; Wornow et al., 2023). We use both: the prediction comes from a locally trained model, and a learned fusion model fuses the language model's assessment into it.

Cost and time in acquisition. Acquisition methods account for test cost either in the predictor, as in cost-sensitive learning, which yields one operating point per trained model (Xu et al., 2013; Nan & Saligrama, 2017), or in the sequential decision, as a step penalty (Janisch et al., 2019) or an episode budget (Kachuee et al., 2019). We place cost in the decision, scaling the step reward by cost under an episode budget, with stopping learned from the discount on the terminal reward; we price tests from published fee schedules and treat turnaround time as a second resource modeled as action duration.

Clinical explanations. Explanations for clinical models come as attribution, which assigns importance to the inputs of a single prediction (Lundberg & Lee, 2017; Ribeiro et al., 2016) and has been used within acquisition to rank the next feature (Guney et al., 2025), or as a narrative that follows the workup, which clinicians prefer (Tonekaboni et al., 2019; Ghassemi et al., 2021; Rudin, 2019) and which language models can now write well (Thirunavukarasu et al., 2023; Van Veen et al.,

2024). Generated text may assert facts that never occurred, so faithfulness must be checked, and LLM judges do this at scale (Ji et al., 2023; Zheng et al., 2023; Croxford et al., 2025). We take the narrative path and add the check: each account describes the acquisition trajectory itself, is verified against the episode record, scored by an independent judge, and tested against mismatched accounts.

Taken together, our work sits at the junction of these lines: a reinforcement-learning acquisition agent that consults a language model and fuses its knowledge into a locally trained prediction, under separate cost and time resources, with verified narrative explanations of the resulting trajectories.

3 Proposed Approach

3.1 Rationale

CARE aims to provide a complete solution for clinical decision support. We do this by improving the medical decision prediction using feature acquisition, and doing this optimally under cost and duration considerations. We model the acquisition process as an episode in which an agent chooses the next test at each step, rewarded by how much each acquisition moves the prediction. The prediction of the medical outcome is done by the reward model, trained on the medical center's data. A medical language model is used to integrate medical knowledge into the decision on top of the patient data; the agent learns when its assessment is worth obtaining, and a fusion model combines that assessment with the local prediction. In order to allow the clinician to interpret system actions, each episode is rendered into a clinical account and verified against what actually happened.

3.2 Data Representation and Feature Acquisition

We represent the patient state in each admission as the group of features revealed, where each feature represents a test or test group. Features are rendered as text and then embedded by Clinical ModernBERT (Lee et al., 2025); feature construction, embedding, and reduction are detailed in Appendix B. The probabilities requested are the prediction targets, in-hospital mortality and 30-day readmission. Ordering a test and receiving its result is represented as feature acquisition, and the agent is trained to acquire features optimally in order to improve prediction accuracy. The agent's reward is calculated from the prediction probabilities. Figure 1 shows one acquisition episode.

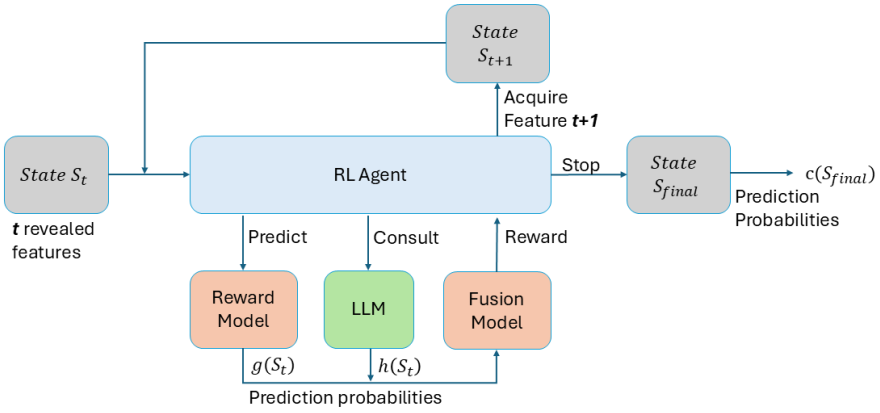


Figure 1: CARE acquisition process. For each state S_t the RL agent predicts the probabilities of the target outcomes and either acquires a feature or consults the LLM, yielding S_{t+1} as the next state. When the agent decides to stop, the final state is reached and the prediction for that state is reported.

3.3 Problem Formulation

Let $F = \{f_1, \ldots, f_d\}$ be the full set of features describing a patient, partitioned into a free subset $S_0 \subset F$ observed at admission and an acquirable subset $Q = F \setminus S_0$. Each acquirable feature $f_j \in Q$ is associated with a pair $(c_j, \tau_j) \in \mathbb{R}^+ \times \mathbb{R}^+$, its monetary cost and its turnaround time.

In addition to the acquirable features, a consultation action l queries a language model g over the currently revealed subset $S_t \subseteq F$, returning its outcome probabilities $g(S_t)$. Like a feature it carries a cost and a turnaround time (c_l, τ_l). An acquisition trajectory is a sequence $S_0 \subset S_1 \subset \cdots \subset S_T$, growing by at most one feature per step, that terminates by choice or when the accumulated resources $\sum_{f_j \in S_T \setminus S_0} c_j$ and $\sum_{f_j \in S_T \setminus S_0} \tau_j$ reach the budgets B_c and B_τ. The outcomes are $y = (y_1, y_2) \in \{0,1\}^2$, in-hospital mortality and 30-day readmission, with y_2 undefined when $y_1 = 1$. For a binary label y and a predicted probability q, we write $\Pr(q = y)$ for the probability assigned to the observed label, q if $y = 1$ and $1 - q$ if $y = 0$. A reward model h maps any revealed subset S_t to a probability pair $h(S_t)$, and a fusion model $c\colon (h(S_t), g(S_t)) \mapsto [0,1]^2$ fuses the reward model's probabilities with the consultation's results into the reported prediction $c(S_t)$. We seek an acquisition policy π that, for each patient, produces a trajectory maximizing $\Pr\big(c_i(S_T) = y_i\big)$ on both outcomes while minimizing the cost and time consumed, with every step conditioned on the values in S_t.

3.4 MDP Formulation

We cast acquisition as an episodic Markov decision process, one episode per patient, with an optional semi-Markov extension by feature durations.

State. The state s_t is a fixed-length vector representing the revealed subset S_t, in which free features are populated at start and each acquirable feature occupies a slot that is zero until revealed; a binary reveal mask m_t maps the revealed features; and, once the LLM has been consulted, its two probabilities and a consult-time map recording which features were revealed at the time of consultation.

Action. $\mathcal{A} = Q \cup \{l, \text{stop}\}$: one reveal action per acquirable feature, the consultation, and a terminal action. Actions for revealed features are invalid; l is valid once per episode; actions exceeding the remaining cost or time budget are invalid as well. Budgets are enforced by masking; the resources spent are a function of the reveal mask m_t, so the remaining budget is implicit in the state.

Transition. Acquiring f_j sets $S_{t+1} = S_t \cup \{f_j\}$, fills its slot, and deducts c_j and τ_j; an absent feature yields a revealed zero. Costs are tiered from publicly published price ranges and normalized so that the cheapest reveal costs 1, with the consultation priced at the minimum; the tiers are listed in Appendix K. Stop terminates the episode, as does the maximum length or an exhausted budget. The episode is not a clock-time simulation: features are admission-level panels, and time enters only through each acquisition's turnaround duration. The consultation earns no step reward; its value reaches the agent only through the terminal reward.

Reward. Rewards use the reported prediction $c(S_t)$, the output of the fusion model over the reward model h and, when consulted, the language model g. A reveal earns the change in probability achieved by the acquired feature, scaled by its cost, and stopping is learned from the discount in the Bellman equation: each further acquisition delays the terminal reward and discounts it by γ.

The step reward r_t and the label-aware terminal reward R_T are

$$r_t = \frac{1}{W c_j} \sum_i w_i \, \delta_i \, \big| c_i(S_{t+1}) - c_i(S_t) \big|, \qquad R_T = \frac{1}{W} \sum_i w_i \, \delta_i \, \Pr\big(c_i(S_T) = y_i\big), \tag{1}$$

where f_j is the feature acquired at step t and c_j its cost, set to 1 when cost does not shape the reward, $\delta_i = 1$ if y_i is defined for the patient, 0 otherwise, w_i is the target weight, and $W = \sum_i w_i \delta_i$.

3.5 Reward Model and Fusion Model

Reward model. The reward model h is a feed-forward network over the feature vector with one output head per target, trained with a class-asymmetric binary cross-entropy summed over the defined targets,

$$\mathcal{L} = \sum_i \delta_i \frac{1}{N_i} \sum_{n=1}^{N_i} \Big[-\alpha_i \, y_i^{(n)} \log h_i^{(n)} - (1 - \alpha_i)\big(1 - y_i^{(n)}\big) \log\big(1 - h_i^{(n)}\big) \Big], \tag{2}$$

where $\alpha_i \in (0, 1)$ counters class imbalance, N_i is the number of training samples with y_i defined, and $h_i^{(n)}$ is the predicted probability for target i on sample n; the readmission head is trained on survivors only. The reward model is pretrained on masked inputs; training details are in Appendix C.

Fusion model. The fusion model c is a network over the reward model's probabilities, the language model's probabilities, and the consult-time map, trained with the same loss; training details are in Appendix C.

3.6 RL AGENT

The agent selects actions with a dueling double DQN (van Hasselt et al., 2016; Wang et al., 2016). Its Q-network takes as input the state vector s_t, the binary reveal mask m_t, which disambiguates a revealed zero from an unrevealed slot, and the current probabilities: the reward model's, the fused prediction's, and, if a consultation has occurred, the language model's together with the consult-time map, so the agent can learn when it is best to consult. The network has one Q-value per action, $|Q| + 2$ in total, and invalid actions are masked after the forward pass so the valid set varies per step without changing the architecture. Training uses the double DQN target, with the online network selecting the successor action and the target network evaluating it, generalized to the semi-Markov form when duration discounting is enabled. Both h and c are frozen during agent training, keeping the reward stationary. Transitions are stored in a prioritized replay buffer (Schaul et al., 2016) shared by asynchronous actors and learner (Horgan et al., 2018); training details are in Appendix H.

3.7 LLM CONSULTATION

Consultation as an acquirable action. The consultation l is an action alongside the features in Q: acquiring it queries MedGemma 27B (Sellergren et al., 2025), a locally hosted open medical model run with deterministic decoding, over the evidence revealed so far, S_t. The prompt presents the revealed features - structured values rendered as text and clinical notes passed verbatim - and asks for the model's assessment of each outcome's probability given the available information. Consultation is taken at most once per episode, so the agent decides whether to consult and at which point of the episode. It is priced at the minimum cost and duration.

Combining the two predictions. The model's answer is parsed into the two outcome probabilities $g(S_t) = (g_1, g_2)$; an unreadable answer fails the parse, leaving the consultation uncharged. The reward model h never sees the consultation: it scores the clinical features alone. The fusion model c fuses the two assessments, taking as input the reward model's probabilities $h(S_t)$, the language model's probabilities $g(S_t)$, and a consult-time map recording which features were revealed when the model was asked, so that it can weigh the model's opinion by how much evidence it was given. The output of c is the reported prediction $c(S_t)$, which enters the reward equations and serves as the final prediction.

3.8 DURATION AS SMDP

Each acquisition occupies its turnaround time τ_j rather than a unit step, and we model it as the action's duration. The process then becomes a semi-Markov decision process (Sutton et al., 1999) in which an action of duration τ discounts its successor value by γ^τ and its step reward materializes only when the result returns, so that between two similarly informative tests the faster one retains more discounted future value. The step reward is credited when the result returns, after τ_t time units, and the successor value is discounted by the full duration:

$$y_t = \gamma^{\tau_t - 1} r_t + \gamma^{\tau_t} \max_a Q(s_{t+1}, a).$$

Durations are measured from MIMIC-IV turnaround times and bucketed into integer units, scaled and normalized accordingly; the procedure is given in Appendix J. As an alternative to discounting, durations can replace the cost vector so that the step reward is divided by the feature's normalized duration. A time budget B_τ bounds the cumulative turnaround of an episode.

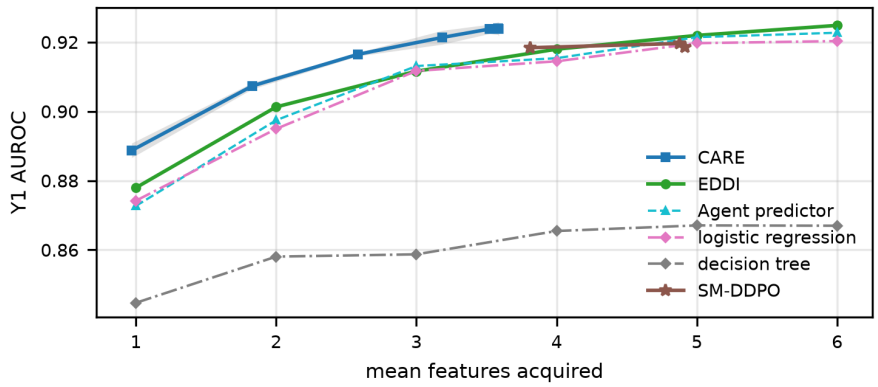


Figure 2: Mortality AUROC against mean panels acquired on the test split. EDDI and the static rankings use exactly k panels for every patient; CARE is a mean over 3 seeds, band ± 1 standard deviation.

3.9 Explainability

Motivation. A clinician must be able to see what the system did and what each test contributed (Thirunavukarasu et al., 2023). We generate a trajectory-grounded clinical account for each episode and verify it against what actually happened.

Account generation and verification. Each episode is serialized into a text trace of the acquired tests, their results, the consultation, and the prediction at each step (Appendix E). A structured prompt asks the model to write an account of the session, citing only material in the trace: what each acquired test assesses, what it returned, and what the session established. Generation is deterministic, using MedGemma 27B (Sellergren et al., 2025), and since generated text can assert facts that never occurred (Ji et al., 2023), each account is checked against the episode record. Passing accounts are scored by a judge model from a different family (Zheng et al., 2023) on faithfulness to the trace and clinical correctness.

4 Experiments

We evaluate on MIMIC-IV (Johnson et al., 2023) with its emergency department and notes modules, 471,384 admissions from 207,305 patients, each admission one episode with 5 free features and 50 acquirable panels and in-hospital mortality as the target; cohort construction, the feature layout, splits, and statistics are in Appendix B. We compare CARE against two dynamic acquisition agents that choose per patient, EDDI (Ma et al., 2019), greedy by expected information gain, and SM-DDPO (Yu et al., 2023), a reinforcement-learning policy with a cost penalty, both adapted to the MIMIC-IV panels; the comparison shows that the agent learns to acquire without the language model and that the consultation carries it beyond them. Two static rankings reveal the same top-k SHAP-ranked panels for every patient, predicted by a logistic regression or by CARE's own predictor h; the gap to CARE shows that a per-patient acquisition policy is meaningful, and with h as the predictor that the gain is in the selection and not the predictor. A decision tree of depth k is a personalization baseline. Implementation details are in Appendix D.

4.1 Impact of Medical Knowledge

Every method is placed at its mean number of acquired panels on the test split (Gadgil et al., 2024; Schütz et al., 2026): EDDI and the static rankings use exactly k panels for every patient; CARE is a single trained policy evaluated under caps $k = 1$ to 6 and stops well before the cap, and can be trained to a longer episode length to acquire more panels; each SM-DDPO point is a separate run with its own cost penalty. Figure 2 shows mortality AUROC against panels acquired.

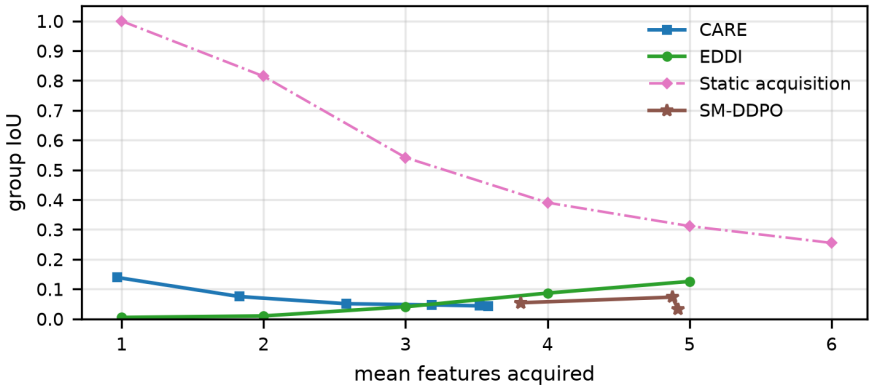


Figure 3: Intersection-over-union of acquired panel sets within random groups of ten test patients, against mean panels acquired.

Table 1: Mortality AUROC for the language model, the reward model, and the fused prediction on the full record.

Predictor	Evidence	AUROC
Language model	full record	0.9045
Reward model	full record	0.9264
Fused	full record	**0.9369**

We show that the impact of the language model is in reaching a decision with less information in hand. At every depth CARE is above every other method at the same number of panels, and its accuracy rises faster per panel: it is above the best other method, EDDI, by 0.011 with one panel, 0.006 with two, and 0.005 with three, and it reaches 0.9240 ± 0.0018 at 3.58 ± 0.66 panels, a level EDDI reaches at six and the static rankings at seven. SM-DDPO peaks at 0.9196 at 4.9 panels, below CARE's accuracy at 3.2. The consultation adds information that further panels supply only slowly, so the agent stops earlier at a higher accuracy. The workups stay individual to the patient: Figure 3 reports the mean intersection-over-union of acquired panel sets within random groups of ten test patients.

The experiment shows that medical knowledge combined with an RL dynamic acquisition agent reaches a decision faster, with higher accuracy at fewer panels, while keeping acquisition individual to the patient.

4.2 LLM Ablation

We isolate the consultation by training CARE without the consultation action, and by comparing the language model and the reward model as standalone predictors on the full record.

Agent without consultation. Figure 4 adds CARE without the consultation action to Figure 2. Without consultation the agent is in line with the other dynamic acquisition agents, within 0.003 of EDDI and 0.002 above SM-DDPO at matched panels, reaching 0.9213 ± 0.0005 at 4.80 ± 0.13 panels. With consultation CARE is above it at every cap, by 0.008 on average over $k = 1$ to 3 and by 0.005 over $k = 4$ to 6, and reaches at 3.2 panels the accuracy the agent without consultation reaches at 4.8.

Prediction on the full record. Table 1 compares the language model prompted with the entire record, the reward model at full reveal, and the fused prediction. The fused prediction exceeds both.

The consultation is the source of the gain: without it the agent is one dynamic acquisition agent among others, and with it the agent reaches a higher accuracy with fewer panels.

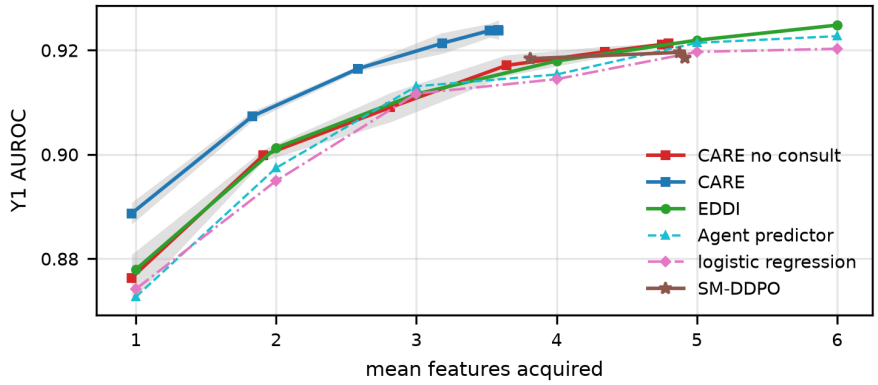


Figure 4: Mortality AUROC against mean panels acquired for CARE with and without consultation, alongside the baselines of Figure 2; bands are ± 1 standard deviation over seeds, 3 with consultation and 2 without.

4.3 Self-Explanation

We evaluate the explanations on 1,000 test-split episodes sampled per seed: for each sampled episode the explainer writes its clinical account, and every metric is computed over those accounts; Table 2 reports the results. Verification against the episode record is the primary check: 98.9% of accounts pass, and the remainder are withheld for citing a value that is not in the trace. Passing accounts are scored by gpt-oss-120b, a judge from a different model family than the writer (Croxford et al., 2025), on faithfulness to the trace and clinical correctness of the stated interpretations; 88% score at least 8 of 10 on both.

As a control, each trace is also paired with the account written for a different episode of similar length, the mismatched rows of Table 2; the judge then scores near zero on both, so the high scores reflect the specific trajectory and not generic clinical prose. Appendix E shows the generated account for one admission.

Table 2: Explanation evaluation on 1,000 sampled test episodes per seed, agent without consultation, mean $\pm$ standard deviation over 3 seeds.

Metric	Agent without consultation
Verification pass rate	$98.9 \pm 0.2\%$
Judge faithfulness	9.51 ± 0.05
Judge clinical correctness	9.12 ± 0.06
Both categories ≥ 8	$87.9 \pm 1.4\%$
Faithfulness, mismatched	0.10 ± 0.02
Clinical correctness, mismatched	0.64 ± 0.04

The accounts are verified, faithful, and specific to their trajectories, so a clinician can read what the system acquired and what it showed.

4.4 Resource Awareness

We evaluate duration and cost awareness on CARE without consultation against a base agent that tracks duration and cost without reward shaping. Figure 5 shows mortality AUROC against mean episode duration and against mean cost; each configuration, including budget-aware, is a separate trained model evaluated under caps $k = 1$ to 6 and placed at its mean spend.

Duration. Semi-Markov discounting shifts the curve left, 23% less time for 0.002 ± 0.001 AUROC at the natural stop, mainly by choosing faster tests, as time per panel falls from 2.4 to 2.0 units. Substituting duration for cost stays on the base agent's curve, 8% less time for 0.004 ± 0.001

AUROC. A time budget of 7 follows the discounted agent's curve and halves the base agent's time for 0.008 ± 0.001 AUROC, binding in 42% of episodes.

Cost. Dividing the step reward by cost shifts the curve left, 11% less spend for 0.001 ± 0.001 AUROC at the natural stop, by fewer panels and slightly cheaper ones, as cost per panel falls from 1.6 to 1.5 units. A cost budget of 5 cuts the spend by 46% for 0.011 ± 0.001 AUROC, binding in 54% of episodes; trained under the cap, it is above both unbudgeted agents at every spend below it.

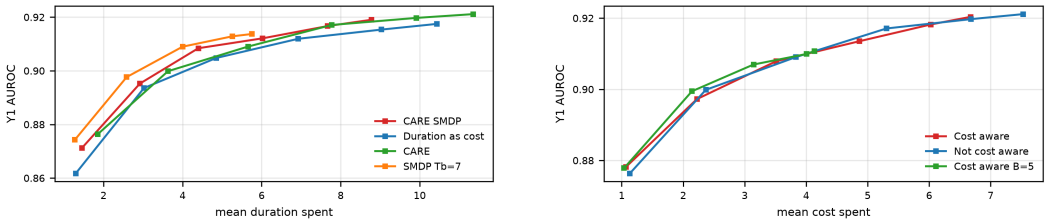


Figure 5: Mortality AUROC against mean episode duration (left) and mean cost (right) for CARE without consultation; each configuration is evaluated under caps $k = 1$ to 6 and placed at its mean spend. All configurations are means over 2 seeds.

Shaping the reward by duration or by cost shifts the accuracy curve toward less of that resource, and a budget bounds it at a controlled loss.

5 Conclusion

We presented CARE, an acquisition agent for clinical decision support that learns from the medical center's own patients and consults a medical language model as one of its actions. The agent learns when a consultation is worth asking for, and a learned fusion model, given the map of features the language model saw, learns how much to rely on its assessment when fusing it into the local prediction. With this medical knowledge the agent is able to reach higher accuracy with fewer tests than dynamic acquisition agents and static test rankings, while keeping its workups individual to the patient. Every episode yields a trajectory-grounded clinical account, verified against the record and judged faithful and clinically sound, that shows clinicians what the system acquired and what it showed. Representing tests as text and embedding them into the agent's state lets unstructured and complex EHR data enter the acquisition process, and modeling turnaround time as an action duration in a semi-Markov formulation, separately from monetary cost, shortens the agent's episodes and lets the medical center's resources and patient discomfort enter the choice of tests. Overall, this is a step toward a clinical decision-support system that combines medical knowledge with learning from local data.

Reproducibility Statement

MIMIC-IV is available to credentialed users through PhysioNet, and the cohort is reconstructed from it by the preprocessing pipeline described in Appendix B, which fixes the patient-level split, the panel grouping (Appendix I), the cost tiers (Appendix K), and the measured durations (Appendix J). The language model, MedGemma 27B, and the judge, gpt-oss-120b, are open weights and were run locally with deterministic decoding; prompts and rubrics are given in the appendix. All reported numbers are means over training seeds, with the seed count stated per experiment. The pipeline, training code, panel grouping, cost tiers, and measured durations are provided as supplementary material.

References

David Bani-Harouni, Chantal Pellegrini, Ege Özsoy, Nassir Navab, and Matthias Keicher. Language agents for hypothesis-driven clinical decision making with reinforcement learning. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.),

International Conference on Learning Representations, volume 2026, pp. 156240–156258, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/fcedb4cd8eea1ac0f139228376e4348d-Paper-Conference.pdf.

Katherine E. Brown, Chao Yan, Zhuohang Li, Xinmeng Zhang, Benjamin X. Collins, You Chen, Ellen Wright Clayton, Murat Kantarcioglu, Yevgeniy Vorobeychik, and Bradley A. Malin. Large language models are less effective at clinical prediction tasks than locally trained machine learning models. *Journal of the American Medical Informatics Association*, 32(5):811–822, 2025. doi: 10.1093/jamia/ocaf038.

Centers for Medicare & Medicaid Services. Clinical laboratory fee schedule. https://www.cms.gov/medicare/payment/fee-schedules/clinical-laboratory-fee-schedule-clfs, 2025a. Accessed September 2026.

Centers for Medicare & Medicaid Services. Physician fee schedule. https://www.cms.gov/medicare/payment/fee-schedules/physician, 2025b. Accessed September 2026.

Ian Connick Covert, Wei Qiu, Mingyu Lu, Na Yoon Kim, Nathan J White, and Su-In Lee. Learning to maximize mutual information for dynamic feature selection. In *International Conference on Machine Learning*, pp. 6424–6447. PMLR, 2023.

Emma Croxford, Yanjun Gao, Elliot First, Nicholas Pellegrino, Miranda Schnier, John Caskey, Madeline Oguss, Graham Wills, Guanhua Chen, Dmitriy Dligach, et al. Evaluating clinical AI summaries with large language models as judges. *npj Digital Medicine*, 8(1):640, 2025.

Gabriel J. Escobar, Vincent X. Liu, Alejandro Schuler, Brian Lawson, John D. Greene, and Patricia Kipnis. Automated identification of adults at risk for in-hospital clinical deterioration. *New England Journal of Medicine*, 383(20):1951–1960, 2020. doi: 10.1056/NEJMsa2001090. URL https://www.nejm.org/doi/full/10.1056/NEJMsa2001090.

Soham Gadgil, Ian Covert, and Su-In Lee. Estimating conditional mutual information for dynamic feature selection. In *International Conference on Learning Representations (ICLR)*, 2024.

Marzyeh Ghassemi, Luke Oakden-Rayner, and Andrew L Beam. The false hope of current approaches to explainable artificial intelligence in health care. *The Lancet Digital Health*, 3(11): e745–e750, 2021.

Osman Berke Guney, Ketan Suhaas Saichandran, Karim Elzokm, Ziming Zhang, and Vijaya B. Kolachalama. Active feature acquisition via explainability-driven ranking. In *International Conference on Machine Learning (ICML)*, 2025.

Isabelle Guyon and André Elisseeff. An introduction to variable and feature selection. *Journal of machine learning research*, 3(Mar):1157–1182, 2003.

Dan Horgan, John Quan, David Budden, Gabriel Barth-Maron, Matteo Hessel, Hado van Hasselt, and David Silver. Distributed prioritized experience replay, 2018. URL https://arxiv.org/abs/1803.00933.

Jaromír Janisch, Tomáš Pevný, and Viliam Lisý. Classification with costly features using deep reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38, 2023.

Lavender Yao Jiang, Xujin Chris Liu, Nima Pour Nejatian, Mustafa Nasir-Moin, Duo Wang, Anas Abidin, Kevin Eaton, Howard Antony Riina, Ilya Laufer, Paawan Punjabi, et al. Health system-scale language models are all-purpose prediction engines. *Nature*, 619(7969):357–362, 2023.

Alistair EW Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, et al. MIMIC-IV, a freely accessible electronic health record dataset. *Scientific data*, 10(1):1, 2023.

Mohammad Kachuee, Orpaz Goldstein, Kimmo Kärkkäinen, Sajad Darabi, and Majid Sarrafzadeh. Opportunistic learning: Budgeted cost-sensitive learning from data streams. In *International Conference on Learning Representations (ICLR)*, 2019.

Naama Kashani, Mira Cohen, and Uri Shaham. P-CAFE: Personalized cost-aware incremental feature selection for electronic health records. *arXiv preprint arXiv:2508.08646*, 2025.

Simon A. Lee, Anthony Wu, and Jeffrey N. Chiang. Clinical ModernBERT: An efficient and long context encoder for biomedical text, 2025. URL https://arxiv.org/abs/2504.03964.

Eric Lehman, Evan Hernandez, Diwakar Mahajan, Jonas Wulff, Micah J Smith, Zachary Ziegler, Daniel Nadler, Peter Szolovits, Alistair Johnson, and Emily Alsentzer. Do we still need clinical language models? In Bobak J. Mortazavi, Tasmie Sarker, Andrew Beam, and Joyce C. Ho (eds.), *Proceedings of the Conference on Health, Inference, and Learning*, volume 209 of *Proceedings of Machine Learning Research*, pp. 578–597. PMLR, 22 Jun–24 Jun 2023. URL https://proceedings.mlr.press/v209/eric23a.html.

Yang Li and Junier Oliva. Active feature acquisition with generative surrogate models. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 6450–6459. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/li21p.html.

Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Chao Ma, Sebastian Tschiatschek, Konstantina Palla, José Miguel Hernández-Lobato, Sebastian Nowozin, and Cheng Zhang. EDDI: Efficient dynamic discovery of high-value information with partial VAE, 2019. URL https://arxiv.org/abs/1809.11142.

Elham Mahmoudi, Neil Kamdar, Noa Kim, Gabriella Gonzales, Karandeep Singh, and Akbar K Waljee. Use of electronic medical records in development and validation of risk prediction models of hospital readmission: systematic review. *BMJ*, 369, 2020. doi: 10.1136/bmj.m958. URL https://www.bmj.com/content/369/bmj.m958.

Feng Nan and Venkatesh Saligrama. Adaptive classification for prediction under a budget. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/d9ff90f4000eacd3a6c9cb27f78994cf-Paper.pdf.

Alvin Rajkomar, Eyal Oren, Kai Chen, Andrew M Dai, Nissan Hajaj, Michaela Hardt, Peter J Liu, Xiaobing Liu, Jake Marcus, Mimi Sun, et al. Scalable and accurate deep learning with electronic health records. *NPJ digital medicine*, 1(1):18, 2018.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016.

Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, 2019.

Tom Schaul, John Quan, Ioannis Antonoglou, and David Silver. Prioritized experience replay. In *International Conference on Learning Representations (ICLR)*, 2016.

Valter Schütz, Han Wu, Reza Rezvan, Linus Aronsson, and Morteza Haghir Chehreghani. Afabench: A generic framework for benchmarking active feature acquisition. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, KDD '26, pp. 9765–9776, New York, NY, USA, 2026. Association for Computing Machinery. ISBN 9798400722592. doi: 10.1145/3770855.3817493. URL https://doi.org/10.1145/3770855.3817493.

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.

Uri Shaham, Yutaro Yamada, and Sahand Negahban. Understanding adversarial training: Increasing local stability of supervised models through robust optimization. *Neurocomputing*, 307:195–204, 2018.

Hajin Shim, Sung Ju Hwang, and Eunho Yang. Joint active feature acquisition and classification with variable-size set encoding. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/e5841df2166dd424a57127423d276bbe-Paper.pdf.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180, 2023.

Liwen Sun, Abhineet Agarwal, Aaron Kornblith, Bin Yu, and Chenyan Xiong. ED-Copilot: Reduce emergency department wait time with language model diagnostic assistance, 2024. URL https://arxiv.org/abs/2402.13448.

Richard S Sutton, Doina Precup, and Satinder Singh. Between MDPs and semi-MDPs: A framework for temporal abstraction in reinforcement learning. *Artificial intelligence*, 112(1-2):181–211, 1999.

Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. Large language models in medicine. *Nature medicine*, 29(8):1930–1940, 2023.

Sana Tonekaboni, Shalmali Joshi, Melissa D. McCradden, and Anna Goldenberg. What clinicians want: Contextualizing explainable machine learning for clinical end use. In Finale Doshi-Velez, Jim Fackler, Ken Jung, David Kale, Rajesh Ranganath, Byron Wallace, and Jenna Wiens (eds.), *Proceedings of the 4th Machine Learning for Healthcare Conference*, volume 106 of *Proceedings of Machine Learning Research*, pp. 359–380. PMLR, 09–10 Aug 2019. URL https://proceedings.mlr.press/v106/tonekaboni19a.html.

Michael Valancius, Max Lennon, and Junier Oliva. Acquisition conditioned oracle for nongreedy active feature acquisition, 2023. URL https://arxiv.org/abs/2302.13960.

Hado van Hasselt, Arthur Guez, and David Silver. Deep reinforcement learning with double Q-learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2016.

Dave Van Veen, Cara Van Uden, Louis Blankemeier, Jean-Benoit Delbrouck, Asad Aali, Christian Bluethgen, Anuj Pareek, Malgorzata Polacin, Eduardo Pontes Reis, Anna Seehofnerová, et al. Adapted large language models can outperform medical experts in clinical text summarization. *Nature medicine*, 30(4):1134–1142, 2024.

Hongnian Wang, Mingyang Zhang, Liyi Mai, Xin Li, Abdelouahab Bellou, and Lijuan Wu. An effective multi-step feature selection framework for clinical outcome prediction using electronic medical records. *BMC medical informatics and decision making*, 25(1):84, 2025.

Ziyu Wang, Tom Schaul, Matteo Hessel, Hado van Hasselt, Marc Lanctot, and Nando de Freitas. Dueling network architectures for deep reinforcement learning. In *International Conference on Machine Learning (ICML)*, 2016.

Michael Wornow, Yizhe Xu, Rahul Thapa, Birju Patel, Ethan Steinberg, Scott Fleming, Michael A Pfeffer, Jason Fries, and Nigam H Shah. The shaky foundations of large language models and foundation models for electronic health records. *npj digital medicine*, 6(1):135, 2023.

Zhixiang Xu, Matt Kusner, Kilian Weinberger, and Minmin Chen. Cost-sensitive tree of classifiers. In Sanjoy Dasgupta and David McAllester (eds.), *Proceedings of the 30th International Conference on Machine Learning*, volume 28, No. 1 of *Proceedings of Machine Learning Research*, pp. 133–141, Atlanta, Georgia, USA, 17–19 Jun 2013. PMLR. URL https://proceedings.mlr.press/v28/xu13.html.

Xi Yang, Aokun Chen, Nima PourNejatian, Hoo Chang Shin, Kaleb E Smith, Christopher Parisien, Colin Compas, Cheryl Martin, Anthony B Costa, Mona G Flores, et al. A large language model for electronic health records. *NPJ digital medicine*, 5(1):194, 2022.

Zheng Yu, Yikuan Li, Joseph Kim, Kaixuan Huang, Yuan Luo, and Mengdi Wang. Deep reinforcement learning for cost-effective medical diagnosis, 2023. URL https://arxiv.org/abs/2302.10261.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

A LLM USAGE

Large language models were used in three ways. In the method, MedGemma 27B serves as the consultation action and as the explanation writer, and gpt-oss-120b scores the explanations; these uses are described in Sections 3.7 and 3.9 and Appendix E. In preprocessing, an LLM summarizes prior discharge summaries into a fixed template before embedding (Appendix B). In preparing the paper, LLM assistants were used for code development, for literature search, and for drafting and editing text; all design decisions, experiments, results, and claims are the authors' own, and every reference was verified against its published source.

B PREPROCESSING PIPELINE

Overview. Each admission is converted into a fixed-length feature vector by a nine-phase Snakemake pipeline over MIMIC-IV v3.1 and its emergency department and notes modules. The layout is F1 demographics, F2 diagnosis history, F3 discharge summary history, F4 triage, F5 chief complaint, F6–F26 laboratory panels, F27–F33 radiology categories, F34–F55 microbiology panels, and F56 the consultation; F1–F5 are always visible, F6–F56 acquirable. History features use only admissions strictly earlier than the current one.

Identifier linkage. A large fraction of clinical rows carry a null admission identifier (10–20% of lab events, 56% of microbiology events, the latter mostly outpatient). Each such row is assigned to an admission by a two-stage cascade: a source-specific window around admission time (-24h/$+24$h for labs and radiology, -24h/$+96$h for microbiology), with an overlapping ED stay disambiguating consecutive admissions and the earliest admission as fallback; otherwise the row is linked through an ED stay, and failing that dropped and counted.

Demographics (F1). Age at admission, gender, height, weight, and BMI . Age is recomputed per admission, since MIMIC-IV shifts dates while preserving intervals. About 25% of the body measurements are missing, informatively so; missingness flags are therefore excluded, as they would correlate with both targets. Missing values are derived from whatever is present before any sampling - BMI is never sampled, weight only when BMI is absent too - with sampling drawn from age-and-gender strata and all values clipped to physiological bounds.

Triage, chief complaint, and history (F2–F5). Triage fields are rendered as a sentence so the same encoder applies; the chief complaint is raw text. Diagnosis history concatenates prior admissions' codes chronologically without dates, which are meaningless once shifted. Discharge history is summarized by an LLM into a fixed template of complaint, diagnosis, treatment, and status, capped at 1024 tokens; the summary is what is embedded.

Laboratory panels (F6–F26). Of 976 lab items with events, 908 are routed into 21 panels and 68 excluded, the latter about 6.6% of 158M events - instrument flags, specimen-quality codes, PCR controls, and ventilator settings miscategorised as blood gas. Panels are keyed first by body fluid and then by clinical ordering pattern, which in several cases crosses the source dictionary's own categories. Events within 24 hours either side of admission become one text block per panel, one

line per measurement, timestamped as elapsed time and flagged abnormal from the source flag or reference range, with repeats preserved so trajectories remain visible.

Radiology categories (F27–F33). Reports are assigned by exam name to chest X-ray, head CT, other CT, MRI, ultrasound, plain film, or other, so imaging can be acquired selectively at the cost of the individual study. Administrative text above the examination line is removed and the clinical content preserved.

Microbiology panels (F34–F55). Events are grouped into 22 panels keyed by test and specimen type together, since the same test on a different specimen is a distinct workup; post-mortem specimens are excluded to prevent target leakage. The window extends 96 hours past admission because these are investigations rather than monitoring, and 28.3% of first microbiology events arrive after day one; the order time determines window membership, so a panel is attributed to the admission in which it was ordered, while result availability is modeled separately by the panel's turnaround duration (Appendix J), which the agent pays before the result is credited. Each event renders test, specimen, organism, an aggregated susceptibility string, and a cleaned comment, since comments are the primary result field for serology, antigen, molecular, gram stain, and parasitology panels.

Embedding. Text features are embedded with Clinical ModernBERT, whose 8,192-token window accommodates multi-visit histories without truncation in most cases. Non-padding hidden states are mean-pooled into one 768-dimensional vector per feature; mean pooling is preferred to the classification token because the encoder is frozen rather than fine-tuned. Absent features produce zero vectors.

Dimensionality reduction. Unreduced vectors are 41,477 floats and the training split occupies roughly 60 GB, exceeding the compute nodes. Reduction is per column, since a global reduction would mix distinct clinical domains. Mean-centred PCA is unusable because zero is the sentinel for absent and masked features and centring would displace it, so the pipeline fits a PCA restricted to non-zero rows, written back into a zero matrix. Transforms are fitted on training rows only; the default target is 128 components, with twelve low-variance columns raised to reach 95% explained variance, giving $D = 7{,}378$.

Targets. Mortality is taken from the admission record. Readmission is 1 when the patient is readmitted between 24 hours and 30 days after discharge; the lower bound excludes administrative back-to-back admissions, about 11% of otherwise-positive labels. Readmission is NaN for deceased patients so the masked loss trains that head on survivors only, and survivors with unusable discharge data are excluded rather than masked, since imputing an outcome would fabricate ground truth. Both invariants are asserted before the dataset is written.

Splits. Admissions are split at the patient level, 80/10/10, stratified on mortality alone, since readmission is undefined for the deceased subset. The split is computed on the admission registry before embedding, so one assignment governs every downstream artefact. Imputation statistics use all admissions, being population parameters the model never sees; train-only fitting applies to the reduction, where components could otherwise encode evaluation variance.

Source-text store. The text behind each embedded feature is kept in a lookup store keyed by admission and feature, so the consultation and explanation modules can recover the evidence behind any state.

C Reward and Fusion Model Training

Reward model. The reward model is pretrained on masked inputs: random masking conceals a random subset of acquirable features per sample, and adversarial masking conceals the features with the largest loss gradients (Shaham et al., 2018).

Fusion model. Consultations are generated online during fusion-model training. Each epoch draws one random reveal mask per training admission. The language model is queried once per

distinct revealed set and its answer cached with the consult-time map. The fusion model is trained on the resulting tuples of $h(S), g(S)$; during agent training and evaluation, a revealed set without a cached answer triggers one query. Training shares the inference load across multiple processes, each running its own language model instance on a GPU with vLLM.

D BASELINES

SHAP ranking. One XGBoost classifier per target is fitted on the training split over all embedding columns. TreeSHAP attributions are computed on the training rows; each panel is scored by the sum of the mean absolute SHAP values of its columns, and panels are ranked by that score. The ranking is fixed across patients: the top-k panels are revealed for every admission, and a ranked panel absent from an admission's record is revealed as zero, as for the agent.

Logistic regression. For each k, a logistic regression is fitted on the free features and the top-k panels' columns, standardized on the training split, one model per target.

Agent's predictor. For each k, the top-k panels are revealed and the agent's reward model h, unchanged, predicts from them. This baseline shares the predictor with the agent and differs only in which panels are revealed, so the gap between them is the contribution of the agent's selection.

Decision tree. A depth-k tree is fitted on all embedding columns, one model per target. Each patient's acquired panels are the distinct panels among the columns tested on that patient's path, excluding free-feature columns; the leaf probability is the prediction.

EDDI. EDDI selects the next panel greedily by expected information gain, estimated with a partial VAE (Ma et al., 2019), while the prediction comes from the agent's own reward model and fusion model, so the baseline differs from the agent only in the acquisition rule. The partial VAE is a set encoder over the revealed slots with a 16-dimensional latent, fitted on the training split and evaluated as one model at exactly k panels.

SM-DDPO. SM-DDPO (Yu et al., 2023) trains a cost-penalized acquisition policy with its own classifier, with discount 1 and the misclassification penalty ratio set to 19 to match the 5% mortality prevalence. Its depth is set by the cost weight at training time, so each point in Figure 2 is a separate run, at cost weights 0, 0.05, and 1, placed at its mean panels acquired on the test split.

E EXPLAINABILITY DETAILS

The trace for one test-split admission as given to the explainer, followed by the generated account (judge faithfulness 10, clinical correctness 9; the account calls a bicarbonate of 27 mEq/L, range 22–32, "at the upper limit of normal," the one mild overstatement). Clock stamps on laboratory lines are omitted.

```
Admission record:  - Demographics:  age 72, sex male, height 193 cm,
weight 95 kg, BMI 25.7 Baseline predicted probabilities:  In-hospital
mortality 0.213, 30-day readmission 0.515
Step 1:  Acquired Lab blood chemistry routine (cost 1.0, duration
1.0) Value:  Bicarbonate:  27 mEq/L (ref:  22-32) Chloride:  103
mEq/L (ref:  96-108) Creatinine:  1.1 mg/dL (ref:  0.5-1.2) Anion
Gap:  15 mEq/L (ref:  8-20) Glucose:  175 mg/dL (ref:  70-100)
[HIGH] Sodium:  141 mEq/L (ref:  133-145) Potassium:  3.5 mEq/L
(ref:  3.3-5.1) Urea Nitrogen:  19 mg/dL (ref:  6-20) Probabilities:
In-hospital mortality 0.213 -> 0.228, 30-day readmission 0.515 ->
0.418
Step 2:  Acquired Lab blood complete blood count (cost 1.0, duration
1.0) Value:  MCHC: 35.4 % (ref:  31-35) [HIGH] MCV: 89 fL (ref:
82-98) Basophils:  0.4 % (ref:  0-2) Hemoglobin:  11.9 g/dL (ref:
14-18) [LOW] Lymphocytes:  29.2 % (ref:  18-42) MCH: 31.3 pg (ref:
27-32) Eosinophils:  5.8 % (ref:  0-4) [HIGH] Hematocrit:  33.6 %
(ref:  40-52) [LOW] Platelet Count:  189 K/uL (ref:  150-440) RDW:
14 % (ref:  10.5-15.5) Monocytes:  6.1 % (ref:  2-11) Neutrophils:
58.6 % (ref:  50-70) Red Blood Cells:  3.8 m/uL (ref:  4.6-6.2) [LOW]
White Blood Cells:  6.2 K/uL (ref:  4-11) Probabilities:  In-hospital
mortality 0.228 -> 0.238, 30-day readmission 0.418 -> 0.387
Step 3:  Acquired Chest X-ray (cost 2.0, duration 1.0) Value:
INDICATION: Syncope.  COMPARISON: None.  PA AND LATERAL VIEWS OF THE
CHEST: The patient is status post median sternotomy, CABG, and aortic
valve replacement.  The cardiac silhouette is normal in size.  The
mediastinal and hilar contours are unremarkable.  The lungs are clear
without focal consolidation.  No pleural effusion or pneumothorax
is seen.  The pulmonary vascularity is normal.  There are no acute
osseous abnormalities.  Multilevel degenerative changes of the
thoracic spine are present.  IMPRESSION: No acute cardiopulmonary
process.  Probabilities:  In-hospital mortality 0.238 -> 0.139,
30-day readmission 0.387 -> 0.303
Step 4:  Consulted the language model (cost 1.0, duration 1.0)
LLM answer:  Differential:  Anemia, Hyperglycemia, Eosinophilia
In-hospital Mortality:  0.0150 30-day Readmission:  0.3500
Probabilities:  In-hospital mortality 0.139 -> 0.015, 30-day
readmission 0.303 -> 0.278
Step 5:  Stopped and predicted (cost 0.0, duration 0.0)
Probabilities:  In-hospital mortality 0.015 -> 0.015, 30-day
readmission 0.278 -> 0.278
Session:  cost spent 5.0, duration spent 4.0.  Ended:  terminal_action
```

The patient is a 72-year-old male with a history of cardiac surgery (median sternotomy, CABG, aortic valve replacement) presenting with syncope. His initial vital signs and demographics show a BMI of 25.7.
The first acquisition was a Lab blood chemistry routine. This test assesses metabolic function and electrolyte balance, which is relevant in evaluating syncope, particularly to check for dehydration, kidney function, or metabolic disturbances. The results showed elevated glucose at 175 mg/dL (ref: 70-100) and a bicarbonate level of 27 mEq/L (ref: 22-32), which is at the upper limit of normal. Other values, including creatinine, electrolytes, anion gap, and urea nitrogen, were within their respective reference ranges. The elevated glucose may indicate hyperglycemia, potentially related to diabetes or stress response.
Next, a Lab blood complete blood count was acquired. This test evaluates red blood cells, white blood cells, and platelets, providing information about oxygen-carrying capacity, infection, and bleeding risk, all relevant in the context of syncope. The results showed low hemoglobin (11.9 g/dL, ref: 14-18), low hematocrit (33.6%, ref: 40-52), and low red blood cells (3.8 m/uL, ref: 4.6-6.2), indicating anemia. There was also an elevated MCHC (35.4%, ref: 31-35) and elevated eosinophils (5.8%, ref: 0-4). The anemia could contribute to symptoms like fatigue or potentially syncope, while the eosinophilia might suggest an allergic reaction or parasitic infection, though its significance here is unclear.
A Chest X-ray was then acquired. This imaging test assesses the lungs, heart, and major blood vessels, looking for acute cardiopulmonary processes that could cause syncope, such as pneumonia, pneumothorax, or heart failure. The report showed no acute cardiopulmonary process, with clear lungs and a normal cardiac silhouette, although it noted the patient's prior cardiac surgery and degenerative spinal changes.
Finally, the agent consulted a language model. This step synthesizes the available data to generate potential diagnoses and updated risk estimates. The model suggested differential diagnoses including anemia, hyperglycemia, and eosinophilia, and provided updated risk estimates.
The session ended after the language model consultation. The agent has identified potential contributing factors to the patient's syncope, including anemia and hyperglycemia, and noted eosinophilia. However, the underlying cause of the syncope remains uncertain, and further investigation may be needed to clarify the significance of the findings and determine the appropriate management.

Judge configuration. Accounts are scored by gpt-oss-120b served with vLLM at temperature 0, with a 16,384-token context and a 2,048-token reply budget including reasoning; faithfulness and clinical correctness are scored in separate passes, each a single prompt containing the rubric, the trace, and the account, ending in a required `Score:` line. Replies without that line, most plausibly where reasoning exhausted the reply budget, are excluded from the means (4 of 1,958 for faithfulness and 6 of 1,958 for clinical correctness for the consult-enabled agent, leaving 975 and 973 episodes with both arms scored).

F Step Reward Ablation

We compare the label-blind step reward r_t of Equation 1 against three alternatives, with the terminal reward R_T unchanged in all four: a signed reward, positive when an acquisition moves the prediction toward the true label and negative when it moves away; a positive-gain reward, positive toward the label and zero away; and no step reward, the terminal-only form. All four agents are trained without the consultation and averaged over 2 seeds; Figure 6 shows mortality AUROC against mean panels acquired under caps $k = 1$ to 6.

The four forms are within seed variation of each other at every matched depth; the difference is how far they acquire. On average, signed and terminal-only reach lower AUROC (0.9175 ± 0.0013 and 0.9180 ± 0.0049) and stop earlier. Label-blind and positive-gain have equal mean AUROC within noise (0.9213 ± 0.0005 and 0.9215 ± 0.0006), and label-blind reaches it with fewer panels (4.80 ± 0.13 against 4.96 ± 0.11), so label-blind was selected over positive-gain.

G Replay Buffer Ablation

We retrain CARE without consultation with uniform sampling in place of prioritized replay, holding all other settings fixed, over 2 seeds. Uniform replay reaches an AUROC of 0.9186 ± 0.0003 at 4.70 ± 0.02 panels against 0.9213 ± 0.0005 at 4.80 ± 0.13 for prioritized replay, 0.003 lower at a similar depth. Prioritization is retained for accuracy and for throughput: the actors generate far more transitions than the learner consumes, and prioritized sampling selects which of them the updates are spent on.

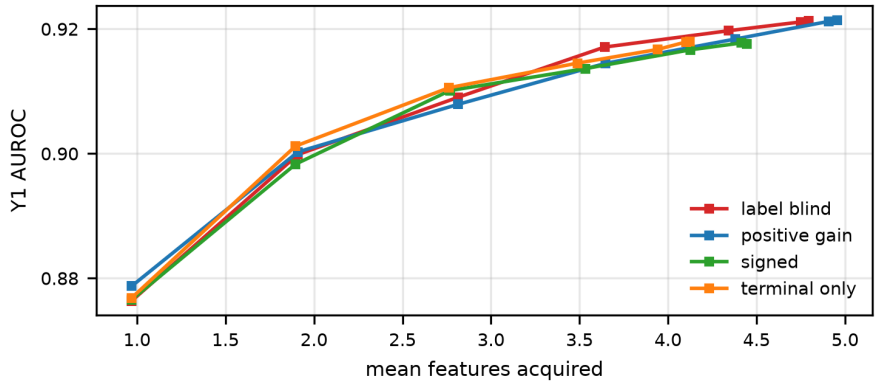


Figure 6: Mortality AUROC against mean panels acquired for four step-reward forms, each evaluated under caps $k = 1$ to 6, mean over 2 seeds; terminal reward unchanged.

H RL Agent Training

Acting and learning run as separate asynchronous processes sharing the replay buffer, as in Ape-X (Horgan et al., 2018), with two actors under a common ϵ-greedy exploration schedule and priorities assigned by the learner. Transitions are sampled in proportion to their temporal-difference error, with importance-sampling weights correcting the induced bias (Schaul et al., 2016); with the actors generating far more transitions than the learner consumes, prioritization selects which transitions the updates are spent on. Each transition's reward is computed from $c(S_t)$ by Equations 1 and stored with it; h and c are frozen, so the reward is stationary. Exploration is warm-started from the static SHAP ranking and annealed to uniform over the first third of training (Appendix D).

I Feature Groups

Each acquirable feature corresponds to a test a clinician would order as a unit. Laboratory items are grouped first by body fluid, so that the same analyte measured in blood, urine, CSF, stool, or bone marrow never shares a panel, and then by clinical ordering pattern; microbiology events are grouped by test and specimen type together, since the same test on a different specimen is a distinct workup; radiology reports are grouped by exam modality. The grouping is derived from ordering practice in the data; where ED-Copilot's clinician-defined laboratory groups (Sun et al., 2024) and ours overlap, the panels coincide or nest (their chemistry, electrolyte, lipase, liver-function, and inflammation groups fall within our routine chemistry panel), and clinician validation of the full grouping remains future work. Table 3 lists the resulting 5 free features and 50 acquirable panels. Laboratory items that are instrument flag codes, tube-hold and specimen-problem codes, PCR controls, and ventilator settings recorded on the blood-gas analyzer are excluded, as are post-mortem microbiology cultures, cytogenetic analyses, and tissue-culture requests; item-level mappings are provided with the code.

J Duration Estimation

For each panel, the turnaround is the median interval between specimen chart time and result store time in MIMIC-IV, normalized to 12-hour units with the fastest laboratory tier defining one unit, rounded, and capped at six units (72 hours), since beyond roughly three days store time reflects protocol-driven negative-culture sign-off rather than decision-relevant turnaround. Imaging durations are ordered by modality (radiography, ultrasound, CT, MRI) and the consultation is assigned the minimum, reflecting inference latency of seconds. Our medians agree in order and magnitude with the mean completion times ED-Copilot reports for its twelve laboratory groups (Sun et al., 2024), all within a factor of two, with their means above our medians on the slow groups as expected for skewed turnaround distributions. The 12-hour unit is chosen for an inpatient horizon in

Table 3: Feature layout by modality and body fluid. Free features are observed at admission; all others are acquirable.

Group	Features
Free (5)	Demographics; diagnosis history; discharge-summary history; triage; chief complaint
Laboratory, blood (13)	Routine chemistry (electrolytes, renal, hepatic, lipids, HbA1c, iron, vitamins, inflammation); complete blood count with differential; coagulation; blood gas; smear morphology; hemolysis and hemoglobinopathy; cardiac markers; endocrine; toxicology and drug monitoring; autoimmune serology and immunoglobulins; infectious serology; tumor markers; immunophenotyping
Laboratory, urine (4)	Urinalysis; urine chemistry; urine pregnancy; urine drug screen
Laboratory, other fluids (4)	Cerebrospinal fluid; other body fluids (ascitic, pleural, synovial, respiratory PCR); stool; bone marrow
Radiology (7)	Chest X-ray; plain film; head CT; other CT; MRI; ultrasound; other
Microbiology, cultures (9)	Routine blood culture; urine culture; wound and tissue culture; non-invasive respiratory; invasive respiratory (bronchoscopic, AFB/TB, Pneumocystis, Legionella, fungal); peritoneal fluid; CSF culture and cryptococcal antigen; stool bacterial; vaginal and genital flora
Microbiology, stains and screens (6)	Blood-bottle Gram stain; respiratory Gram stain; wound and tissue Gram stain; MRSA/S. aureus screen; Legionella urinary antigen; C. difficile
Microbiology, serology and molecular (6)	Hepatitis and HIV viral loads; syphilis serology; herpesvirus serology (CMV, EBV, VZV); respiratory viral panel; stool parasitology; miscellaneous serology (Lyme, ASO, toxoplasma, rubella, rubeola, mumps, malaria)
Microbiology, other (1)	Rare cultures and screens (hardware, joint fluid, bone marrow, skin and nail specimens, resistance screens, throat)
Consultation (1)	LLM consultation (acquirable action)

which cultures dominate; all laboratory panels fall within one unit, so within-day differences among them do not enter the discount. The resulting per-feature durations are listed in Table 4.

Table 4: Feature durations in 12-hour units, derived from MIMIC-IV median result turnaround (store time minus chart time), capped at six units. Imaging is ordered by modality; the consultation is assigned the minimum.

Units	Median turnaround	Features
1	< 15 h	Blood gas, urine pregnancy, complete blood count, urinalysis, urine toxicology, coagulation, routine chemistry, blood toxicology, cardiac markers, CSF analysis, blood smear morphology, hemolysis panel, urine chemistry, other body fluids, gram stain (wound/tissue, respiratory), tumor markers, endocrine, stool analysis, autoimmune serology, C. difficile, urinary antigens; chest X-ray, plain film; LLM consultation
2	18–27 h	Syphilis serology, respiratory viral panel, infectious serology, bone marrow, immunophenotyping, blood-bottle gram stain; ultrasound, other radiology
3	–	Head CT, other CT
4	43–48 h	Vaginal/genital flora, hepatitis/HIV, urine culture, stool bacterial culture, stool parasitology, miscellaneous serology, herpesvirus serology; MRI
5	57 h	MRSA/staph screen
6	≥ 68 h (cap)	Non-invasive respiratory culture, other respiratory culture, other CSF, peritoneal culture, wound culture, routine blood culture, other microbiology

K Cost Estimation

Each acquirable feature is assigned a cost tier derived from publicly published United States price ranges for its test category (Centers for Medicare & Medicaid Services, 2025a;b), expressed as a

multiple of the cheapest acquisition. Routine blood and urine panels and gram stains form the base tier; cultures, cardiac and endocrine markers, and plain radiography the second; send-out serology, molecular assays, CT, and fluid analyses the third; flow cytometry and MRI the fourth; and a full bone-marrow workup the highest. The consultation is priced at the base tier, reflecting compute-only expense. Exact prices vary by institution; the tier structure rather than the dollar values enters the model. The resulting per-feature costs are listed in Table 5.

Table 5: Feature cost tiers as multiples of the cheapest acquisition, with the approximate US price range each tier reflects.

Cost	Price range	Features
1	\$10–40	Routine chemistry, complete blood count, coagulation, blood gas, blood smear morphology, hemolysis panel, urinalysis, urine chemistry, urine pregnancy, blood-bottle gram stain, gram stain (respiratory, wound/tissue); LLM consultation
2	\$40–150	Chest X-ray, plain film, other radiology, cardiac markers, endocrine, routine blood culture, urine culture, wound culture, non-invasive respiratory culture, urinary antigens, MRSA/staph screen, stool bacterial culture, stool analysis, C. difficile, syphilis serology, vaginal/genital flora
4	\$150–400	Head CT, other CT, ultrasound, blood toxicology, urine toxicology, autoimmune serology, infectious serology, tumor markers, hepatitis/HIV, herpesvirus serology, miscellaneous serology, respiratory viral panel, other respiratory culture, stool parasitology, CSF analysis, other body fluids, other CSF, peritoneal culture, other microbiology
6	\$400–1,000	Immunophenotyping (flow cytometry), MRI
10	\$1,500+	Bone marrow workup