WHERE CAN RISK BE REPAIRED? EVIDENCE ACQUISITION AND ESCALATION FOR SELECTIVE CLINICAL LLM DECISION-MAKING

Anonymous authors
Paper under double-blind review

ABSTRACT

In clinical decision making under incomplete evidence, current methods mainly answer whether the current decision might be wrong, but they cannot tell whether to repair an error by asking for additional evidence or deferring the case to a stronger model. We propose a repairability-aware decision framework that separates current decision risk from two distinct forms of repairability. We also construct a frozen, deployable DECIDE/ASK/DEFER router. On the MIMIC-IV held-out test cohort, Always DECIDE achieves 36.39% accuracy, the deployable ASK-only policy reaches 59.02%, and the joint router reaches 62.69%, improving over ASK-only by 3.67 percentage points. In a cross-family robustness evaluation using Qwen3-8B and Qwen3-32B, we again observe the non-redundant structure of ASK and DEFER repair sets. However, the joint router does not show a statistically significant advantage over ASK-only. This suggests that the repair channel structure can generalize across model families, while the incremental gain from joint routing is model-pair dependent. We also find that increasing the ASK budget does not monotonically improve accuracy. Learned stopping reduces the mean number of questions from 5.00 to 3.18 without evidence of a decrease in accuracy.

1 INTRODUCTION

Large language models are increasingly being used in scenarios such as clinical question answering, symptom explanation, and decision support (Singhal et al., 2023; Thirunavukarasu et al., 2023; Hager et al., 2024a). In these applications, models often need to make judgments when evidence is incomplete. Thus, the reliability of current predictions has become an important safety issue. Existing selective prediction, uncertainty estimation, and abstention methods primarily examine the conditions under which a model should not directly commit to its current prediction and identify states that may require intervention via risk or confidence signals (Geifman & El-Yaniv, 2017; 2019; Machcha et al., 2026). However, identifying a high-risk state does not determine what action the system should take next. For certain cases, the error may stem from the lack of critical patient evidence, in which case obtaining new vital signs, lab results, or imaging information might rectify the current judgment; for other cases, the existing evidence may be sufficient, but the base model lacks the ability to properly utilize this information, in which case escalation to a stronger model may be more useful. Therefore, in clinical decisions with incomplete evidence, the system not only needs to determine whether the current decision might be incorrect but also needs to answer a further question: *Where can the risk be repaired?*

Existing research on this issue primarily comes from three directions. Selective prediction and abstention methods focus on whether the current prediction is sufficiently reliable; the risk estimate itself does not indicate which subsequent intervention can truly correct the error. Active feature acquisition and information value methods investigate which additional observations can improve subsequent decisions (Saar-Tsechansky et al., 2009; Valancius et al., 2024; Guney et al., 2025), but these methods typically do not compare whether information acquisition or capability escalation affect different types of errors. Model routing and cascading methods allocate inputs to models of varying capabilities based on cost, confidence, or task difficulty (Ding et al., 2024; Gupta et al., 2024; Ong et al., 2025; Chen et al., 2024), but stronger models do not guarantee monotonic improvement in results, as escalation can both fix errors in the base model and potentially disrupt previously

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

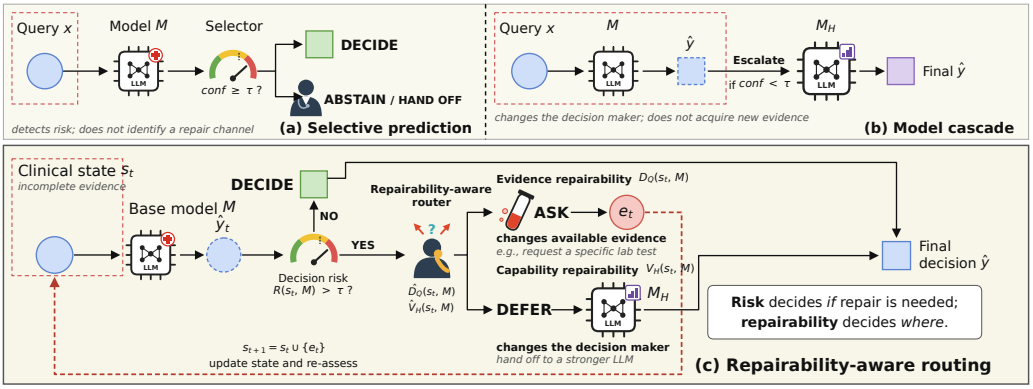


Figure 1: Overview of repairability-aware routing. Selective prediction identifies when a current decision may be unsafe, while model cascades change capability without acquiring new evidence. Our framework separates decision risk from evidence repairability and capability repairability. Risk determines whether repair is needed, while repairability determines whether to ASK for admissible evidence or DEFER to a stronger model.

correct judgments. Prior work has already distinguished DECIDE, ASK, and DEFER as three types of response actions in cases of incomplete evidence (dec, 2026); this distinction in the action space still does not answer which type of intervention can actually correct the current error. This gap means that a state can be correctly identified as risky while the system still chooses the wrong repair channel, spending additional evidence on cases that require capability escalation or escalating cases that could instead be repaired through admissible evidence acquisition.

To address this issue, we reformulate the decision-making under incomplete evidence as a repairability-aware decision problem. The system needs to determine not only whether the current state poses a high risk, but also whether additional patient evidence or stronger model capability can repair the error. This reformulation brings three main challenges. First, current risk and evidence repairability are not the same quantity. Even if two cases have similar error risks, their responses to new patient evidence may be completely different; risk cannot be directly used as a substitute signal for ASK. Second, ASK and DEFER may operate on different sets of errors, and capability escalation itself may also cause harm; therefore, the two interventions cannot be simply regarded as two interchangeable high-risk handling methods. Third, when ASK can be performed sequentially, more evidence does not guarantee a monotonic improvement in accuracy. Each time new information is added, it changes the current state, so the system also needs to determine when to continue ASK, when to stop, and how to choose between ASK and DEFER when they have different costs. As illustrated in Figure 1, the framework separates whether intervention is needed from where the current risk can be repaired, leading to distinct DECIDE, ASK, and DEFER actions.

To tackle these challenges, we propose a repairability-aware decision framework that separates the current decision risk from two distinct forms of repairability. We use $R(s, M)$ to represent the current decision risk, $D_Q(s, M)$ to represent the evidence repairability achieved by obtaining additional admissible patient evidence, and $V_H(s, M)$ to represent the harm-aware capability repairability through escalation to a stronger model. V_H simultaneously considers the repairs and new harms that may arise from capability escalation, thereby avoiding the perception of stronger model escalation as a monotonically safe operation. Based on these three quantities, we learn $\hat{R}$, $\hat{D}_Q$, and $\hat{V}_H$, and construct a frozen deployable DECIDE/ASK/DEFER router. All estimators, thresholds, and ASK priorities are determined solely from the development data and are frozen before final testing. For multi-step information acquisition, we further introduce sequential ASK and learned stopping, and use normalized intervention costs to analyze how ASK, DEFER, and DECIDE vary with relative intervention costs. As illustrated in Fig. 1, ASK changes the available evidence, whereas DEFER changes the decision maker.

We conduct the primary held-out evaluation on MIMIC-IV-Ext Clinical Decision Making using Llama-3.1-8B and Llama-3.1-70B, and perform an independent cross-family robustness evaluation

using Qwen3-8B and Qwen3-32B. The experiments first show that the correlation between current risk and realized evidence repairability is weak, and the repair sets of ASK and DEFER only partially overlap, with clear cases of ASK-only and DEFER-only. In the untouched MIMIC test cohort, the accuracy of Always DECIDE is 36.39%, the deployable ASK-only policy reaches 59.02%, and the joint router reaches 62.69%, an increase of 3.67 percentage points over ASK-only. In the Qwen3 evaluation, the non-redundant structure of ASK and DEFER repair sets is reproduced, but the additional gain of the joint router over ASK-only is smaller and does not reach statistical significance, which indicates that the repair-channel structure can be reproduced across model families, while the incremental gain from joint routing is model-pair dependent. Finally, in the DDXPlus sequential experiment, increasing the ASK budget does not monotonically improve accuracy. Learned stopping reduces the average number of questions from 5.00 to 3.18, with no observed decline in accuracy.

2 Related Work

Existing work mainly studies whether a model should trust its current prediction and output it. Uncertainty estimation uses signals such as confidence, predictive uncertainty, and calibration to assess the reliability of the current prediction. Selective prediction uses these reliability signals to balance coverage and error risk, and outputs predictions only for sufficiently reliable samples (Geifman & El-Yaniv, 2017; 2019). Abstention allows the model to withhold an answer when the current prediction is considered unreliable, avoiding direct commitment to a potentially incorrect prediction. Together, these approaches help determine whether the current prediction is reliable enough to commit to, which is particularly important in clinical decision support (Machcha et al., 2026). However, they do not directly answer which intervention can actually repair the underlying error.

A lot of work studies what additional information should be acquired when the current evidence is incomplete. Because acquiring information may incur costs, existing methods use criteria such as value of information, expected utility, or learned acquisition policies to select the most valuable observations (Saar-Tsechansky et al., 2009; Valancius et al., 2024; Guney et al., 2025). Prior work also treats ASK as an explicit response action under incomplete evidence (dec, 2026). This allows the model to acquire additional information rather than immediately making a final decision when the available evidence is insufficient, potentially improving subsequent prediction or decision quality. However, these studies mainly focus on what information should be acquired. They do not directly answer whether the current error should be repaired through additional evidence or through escalation to a stronger model.

Model routing and cascading methods study which model should handle a given input when models with different capabilities and computational costs are available (Ding et al., 2024; Gupta et al., 2024; Ong et al., 2025; Chen et al., 2024). These methods often use signals such as confidence, uncertainty, task difficulty, predicted quality, or cost to keep relatively easy cases in the base model and route more difficult or high-risk cases to a stronger model, thereby balancing model performance and computational cost. This line of work provides an important foundation for capability escalation. However, routing an input to a stronger model does not guarantee that the outcome for every individual case will improve. These methods only focus on when a model should be upgraded, rather than on whether the current error should be repaired with additional evidence or escalated to a stronger model.

3 Methodology

We formulate decision making under incomplete evidence as a repairability-aware routing problem over three response actions: DECIDE, ASK, and DEFER. The framework first separates current decision risk from two forms of repairability, corresponding to evidence acquisition and capability escalation. We then use these quantities to construct a deployable DECIDE/ASK/DEFER router and further extend the framework to sequential evidence acquisition and cost-aware routing.

3.1 Problem Setup and Current Decision Risk

Let s denote the current clinical state available to a base model M. The state contains the patient evidence available at the current decision point. Let H denote a stronger model that can be used for capability escalation, and let $\mathcal{Q}(s)$ denote the set of admissible evidence acquisition actions available from state s. An admissible action can only reveal evidence that is available at the current decision point and does not use information from the future.

At each state, the system can take one of three actions. DECIDE keeps the current evidence state unchanged and returns the prediction of the base model. ASK selects an admissible evidence item $q \in \mathcal{Q}(s)$, which produces an updated state s^q that is evaluated again by the base model. DEFER keeps the current evidence state unchanged but transfers the decision to the stronger model H. Thus, ASK changes the available evidence, whereas DEFER changes the decision maker.

Let

$$Y_M(s) \in \{0, 1\}$$

denote the error indicator of the base model at state s, where $Y_M(s) = 1$ indicates an incorrect prediction and $Y_M(s) = 0$ indicates a correct prediction. We define the current decision risk as

$$R(s, M) = P\big(Y_M(s) = 1 \mid s, M\big).$$

A larger value of $R(s, M)$ indicates that the current prediction is more likely to be incorrect and may therefore require intervention. However, current decision risk only indicates whether intervention may be needed. It does not determine whether the current error can be repaired through additional evidence or through escalation to a stronger model. We therefore introduce two intervention-specific notions of repairability.

3.2 Repairability

Evidence repairability. For an admissible evidence acquisition action $q \in \mathcal{Q}(s)$, acquiring q changes the current state from s to s^q. We define the realized change in prediction correctness as

$$\Delta_Q(q \mid s, M) = Y_M(s) - Y_M(s^q) \in \{-1, 0, +1\}.$$

A value of $+1$ indicates that the base model is incorrect at the original state but becomes correct after acquiring the additional evidence. In this case, ASK repairs the current error. A value of 0 indicates that the acquisition does not change prediction correctness. A value of -1 indicates that the model is originally correct but becomes incorrect after the additional evidence is introduced, meaning that the acquisition causes harm.

Because the realized outcome after acquisition is not available at decision time, we define the expected repair value of an admissible evidence action as

$$V_Q(q \mid s, M) = \mathbb{E}\left[\Delta_Q(q \mid s, M) \mid s, q, M\right].$$

The evidence repairability of the current state is then defined as

$$D_Q(s, M) = \max\left(0, \max_{q \in \mathcal{Q}(s)} V_Q(q \mid s, M)\right).$$

This quantity characterizes whether there exists an admissible evidence acquisition that is expected to improve the current decision. The zero option represents choosing no evidence acquisition when none of the admissible actions has positive expected repair value.

For hindsight analysis, we additionally define oracle evidence repairability as

$$D_Q^{\mathrm{HO}}(s, M) = \max\left(0, \max_{q \in \mathcal{Q}(s)} \Delta_Q(q \mid s, M)\right).$$

This oracle quantity uses the realized post-acquisition outcome and is therefore available only for supervision and analysis. It is never available to the deployed router.

Harm-aware capability repairability. Evidence acquisition is not the only way to repair an incorrect decision. A second option is to keep the available evidence unchanged and escalate the case from the base model M to a stronger model H.

Let $Y_H(s)$ denote the error indicator of the stronger model at the same state s. We define the realized effect of capability escalation as

$$A_H(s, M) = Y_M(s) - Y_H(s) \in \{-1, 0, +1\}.$$

A value of $+1$ indicates that the stronger model repairs an error made by the base model. A value of 0 indicates that escalation does not change prediction correctness. A value of -1 indicates that the base model is originally correct but the stronger model produces an incorrect prediction. Capability escalation can therefore repair some errors while introducing new harms in other cases.

We define harm-aware capability repairability as

$$V_H(s, M) = P(A_H = +1 \mid s, M) - P(A_H = -1 \mid s, M).$$

Unlike a measure that only considers whether the stronger model is correct, V_H explicitly accounts for both repairs and harms caused by escalation. A larger value indicates that the expected repair benefit of escalation outweighs the risk of changing an originally correct decision into an incorrect one.

Unlike D_Q, V_H is not truncated at zero. D_Q selects the best option among admissible evidence acquisitions and the no-acquisition option, whereas V_H represents the signed net value of escalating to the fixed stronger model H. A negative V_H therefore indicates that escalation is expected to introduce more harm than repair and remains informative for downstream routing. This signed value is also used by the cost-aware routing rule introduced in Section 3.4.

The framework therefore distinguishes three quantities:

$$R(s, M), \qquad D_Q(s, M), \qquad V_H(s, M).$$

Current decision risk characterizes whether the current prediction may require intervention. Evidence repairability characterizes whether changing the available evidence can repair the current decision, while capability repairability characterizes whether changing the decision maker can do so without ignoring the possibility of escalation-induced harm. These quantities separate the need for intervention from the location at which repair may be possible.

3.3 Deployable DECIDE/ASK/DEFER Routing

The repairability quantities defined above describe the potential value of different interventions, but their realized outcomes are not directly available at deployment time. We therefore learn three estimators,

$$\hat{R}(s, M), \qquad \hat{D}_Q(s, M), \qquad \hat{V}_H(s, M),$$

for current decision risk, evidence repairability, and harm-aware capability repairability, respectively.

All estimators are trained using development data only. The decision thresholds, admissible ASK selection rules, and action priorities are also selected using development data and are frozen before evaluation on the final test cohort. No estimator, threshold, or action priority is selected using the final test data.

At deployment time, the router first uses $\hat{R}$ to determine whether the current prediction should be considered for intervention. If the estimated risk is below the development-frozen risk threshold, the system selects DECIDE and returns the current prediction. For states that pass the risk gate, the router evaluates the estimated evidence and capability repairability. ASK can only be selected when an admissible evidence acquisition is available from the current state. The final action is determined by the development-frozen repairability thresholds and deterministic routing rule, yielding one of DECIDE, ASK, or DEFER.

This construction separates risk detection from repair selection. The risk estimator determines whether the current state should be considered for intervention, whereas the repairability estimators determine which available repair channel is appropriate for that state. The same routing rule is fixed before final testing and is applied without access to realized post-intervention outcomes.

3.4 Sequential and Cost-Aware Routing

Sequential evidence acquisition. A single ASK action changes the available evidence and therefore changes the decision state. Starting from an initial state s_0, repeated evidence acquisition generates a sequence

$$s_0 \to s_1 \to s_2 \to \cdots \to s_t.$$

Each acquisition can change the current prediction, the remaining admissible evidence, and the value of subsequent acquisitions. Additional evidence therefore does not imply monotonic improvement in prediction quality.

To support multi-step acquisition, we introduce a learned stopping mechanism that decides after each ASK whether another evidence acquisition should be performed. The stopping decision is conditioned on the current state after the newly acquired evidence has been incorporated. The goal is to stop acquisition when continuing ASK is unlikely to provide sufficient additional value, rather than using a fixed acquisition budget for every case.

This formulation allows the system to distinguish between cases that benefit from continued evidence acquisition and cases for which further ASK is unnecessary.

Cost-aware routing. The routing framework can also account for different intervention costs. Let c_Q denote the normalized cost of ASK and c_H denote the normalized cost of DEFER. These terms represent relative intervention costs rather than monetary costs.

For a state that passes the frozen risk gate, and subject to ASK legality, we define the net utilities of the three actions as

$$0, \qquad \hat{D}_Q(s, M) - c_Q, \qquad \hat{V}_H(s, M) - c_H,$$

corresponding to DECIDE, ASK, and DEFER, respectively. The cost-aware action is

$$a^*(s) = \arg\max \left\{ 0,\ \hat{D}_Q(s, M) - c_Q,\ \hat{V}_H(s, M) - c_H \right\}.$$

The zero utility associated with DECIDE provides a no-intervention option when neither repair channel has sufficient value after accounting for its cost. The formulation also makes the relative preference between ASK and DEFER depend explicitly on their intervention costs.

Together, the framework separates whether intervention may be needed from where the current risk can be repaired, while supporting both sequential evidence acquisition and cost-aware action selection.

4 Experiments and Results

We evaluate the proposed repairability-aware framework from three aspects. First, we test whether evidence acquisition and capability escalation correspond to distinct and non-redundant repair channels. Second, we evaluate the deployable DECIDE/ASK/DEFER router and examine whether the routing structure is robust across model families. Finally, we study sequential evidence acquisition and cost-aware routing under multi-step ASK and varying intervention costs.

4.1 Experimental Setup

Our primary evaluation uses the MIMIC-IV-Ext Clinical Decision Making dataset (Hager et al., 2024b), which is derived from MIMIC-IV (Johnson et al., 2023). We construct a cohort of 1,630 patient-level cases with patient-disjoint training, development, and test splits of 976, 327, and 327 cases. The task contains four diagnoses: appendicitis, cholecystitis, pancreatitis, and diverticulitis. The initial state contains chief complaint, age, and sex. ASK can acquire additional evidence through VITALS, LABS, and IMAGING. To prevent future information leakage, only evidence available before the earlier of admission time and emergency-department out time is admissible.

The primary model pair is Llama-3.1-8B-Instruct and Llama-3.1-70B-Instruct (Grattafiori et al., 2024). We also evaluate Qwen3-8B and Qwen3-32B (Yang et al., 2025) for cross-family robustness using the same MIMIC split, initial-state construction, ASK channels, temporal constraints,

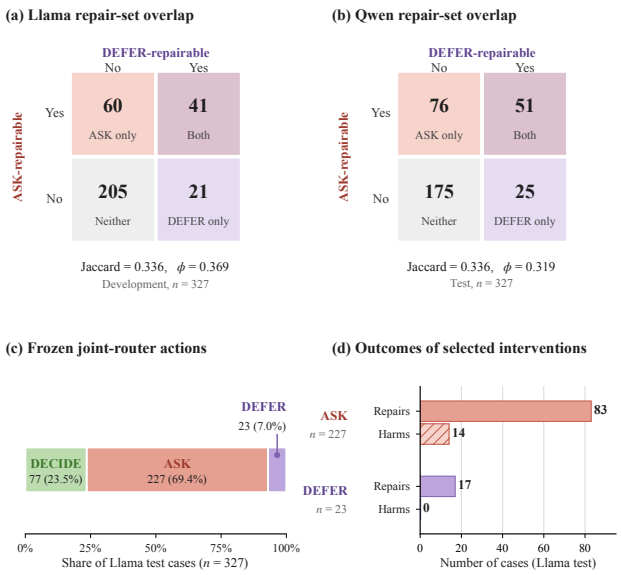


Figure 2: Repair-channel complementarity and routing behavior. (a,b) ASK and DEFER repair sets overlap but remain non-redundant across the Llama and Qwen evaluations. (c,d) The frozen Llama router primarily selects ASK while retaining a smaller DEFER branch, and both selected intervention channels produce repairs.

and output-parsing procedure. We select all estimators, thresholds, and ASK priorities using development data only and freeze them before final test evaluation. For Llama, the frozen thresholds for risk, evidence repairability, and capability repairability are 0.3, 0.2, and 0.4; for Qwen, they are 0.3, 0.2, and 0.5. We use paired bootstrap confidence intervals and exact McNemar tests for paired policy comparisons where applicable.

4.2 Repairability Structure

We first examine whether evidence acquisition and capability escalation repair the same cases. In the held-out Llama test cohort, the base model makes 208 errors, of which 104 (50%) can be repaired by at least one admissible ASK action. This shows that a substantial share of errors can be repaired by additional evidence without changing the decision maker. On the same cohort, escalation to Llama-3.1-70B repairs 67 base-model errors but also changes 18 originally correct predictions into incorrect ones. Thus, both evidence acquisition and capability escalation can repair errors, while escalation is not monotonically beneficial at the individual-case level. As shown in Fig. 2, ASK and DEFER repair sets overlap but remain substantially non-redundant across both model families.

To characterize the repair-channel mechanism without using the final Llama test cohort, we perform the repair-set analysis on the 327-case Llama development cohort. We observe 60 ASK-only cases, 21 DEFER-only cases, 41 cases repairable by both channels, and 205 repairable by neither. The Jaccard overlap is 0.3361, and the phi coefficient is 0.3690, indicating that the two repair channels are positively associated but clearly non-redundant.

As an independent cross-family robustness check, the same structure appears on the held-out Qwen test cohort. The Jaccard overlap is 0.3355, and the phi coefficient is 0.3191. Overall, these results show that evidence acquisition and capability escalation repair overlapping but distinct sets of cases across model families.

4.3 Deployable Repairability-Aware Routing

We next test whether learned repairability estimates can exploit this structure in a deployable routing policy. On the held-out Llama test cohort, Always DECIDE achieves 36.39% accuracy, the ASK-only policy reaches 59.02%, and the joint repairability-aware router reaches 62.69%. The joint

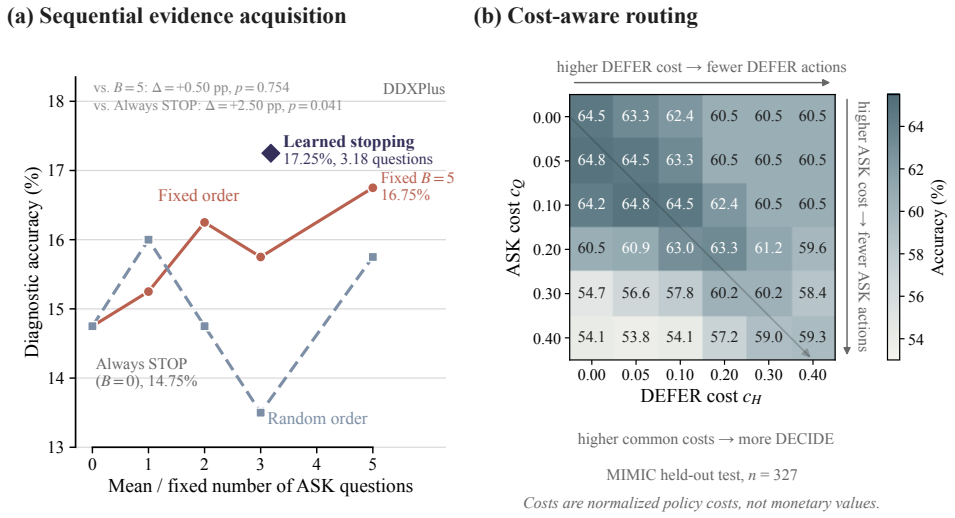


Figure 3: Sequential evidence acquisition and cost-aware routing. (a) Increasing the ASK budget does not monotonically improve accuracy, while learned stopping reduces the mean number of questions relative to fixed $B = 5$ without evidence of an accuracy decrease. (b) ASK and DEFER costs alter repair-channel allocation and the intervention pattern; the sweep characterizes policy behavior rather than selecting a held-out operating point.

router therefore improves over ASK-only by 3.67 percentage points, with a paired bootstrap 95% confidence interval of [1.53, 5.81] and exact McNemar $p = 0.001831$.

The joint router selects ASK for 227 cases, DEFER for 23 cases, and DECIDE for 77 cases. Among the 23 deferred cases, 17 base-model errors are repaired, and no originally correct prediction is changed into an incorrect one. Thus, the gain from joint routing comes from selective capability escalation rather than frequent use of the stronger model.

Oracle ASK-only, DEFER-only, and joint routing reach 68.20%, 56.88%, and 75.84%, respectively, indicating remaining headroom beyond the learned estimators.

For Qwen3-8B and Qwen3-32B, ASK-only reaches 62.08%, and joint routing reaches 63.00%. The difference is 0.92 percentage points, with a 95% confidence interval of [-0.61, 2.45] and $p = 0.453$. The improvement is therefore not statistically significant. Together with the repair-set analysis, this result suggests that the non-redundant repair-channel structure can be reproduced across model families, while the incremental gain from joint routing depends on the model pair.

4.4 Sequential and Cost-Aware Routing

We use DDXPlus (Tchango et al., 2022) to study sequential evidence acquisition. The Always DECIDE baseline achieves 14.75% accuracy. Under fixed-order ASK, accuracy is 15.25%, 16.25%, 15.75%, and 16.75% for budgets of 1, 2, 3, and 5 questions. Under random-order ASK, the corresponding accuracies are 16.00%, 14.75%, 13.50%, and 15.75%. Increasing the ASK budget therefore does not monotonically improve accuracy. Fig. 3(b) shows that intervention costs alter both the preferred repair channel and the overall intervention rate.

The learned stopping policy achieves 17.25% accuracy with 3.18 questions on average, compared with 16.75% using five fixed questions ($\Delta = +0.5$ pp, 95% CI $[-1, 2]$, $p = 0.754$). We therefore do not interpret learned stopping as improving accuracy. Instead, it reduces acquisition from 5.00 to 3.18 questions, a 36.4% reduction, without an observed accuracy decrease. This conclusion is unchanged under alternative treatments of invalid outputs.

Finally, we evaluate a 6×6 grid of normalized ASK and DEFER intervention costs. In the main text, we report two representative slices. When escalation cost is fixed at zero and ASK cost increases from 0 to 0.4, ASK decreases from 233 to 51 cases, while DEFER increases from 28 to 143 and

DECIDE from 66 to 133. When ASK cost is fixed at zero and escalation cost increases from 0 to 0.4, DEFER decreases from 28 to 4 cases, while ASK increases from 233 to 255.

These results show that the router responds systematically to relative intervention costs. Higher evidence-acquisition costs shift decisions away from ASK toward DEFER or DECIDE, while higher escalation costs shift decisions away from DEFER toward ASK. The appendix reports the full 6×6 grid.

5 Discussion and Limitations

Our results show that identifying whether a decision is high risk is a different problem from identifying how to repair that risk. Traditional selective prediction, uncertainty estimation, and abstention methods mainly determine whether the current prediction is reliable enough, but high risk alone does not indicate whether the next step should be to obtain more evidence or hand the case to a stronger model. Our experiments further show that a portion of errors can be repaired with additional evidence without changing the decision maker, while capability escalation may either repair errors or undermine an otherwise correct prediction. Meanwhile, the repair sets of ASK and DEFER overlap only partially. These results show that evidence acquisition and capability escalation should not be viewed as interchangeable responses to high risk. Separating current decision risk from evidence repairability and capability repairability provides the quantities needed to answer the second question after whether intervention is needed: where the current risk can be repaired.

The results across model families further illustrate the need to distinguish conclusions at two levels: repair-channel structure and joint routing gain. We observed ASK-only, DEFER-only, and jointly repairable cases in both model pairs. This indicates that the non-redundancy of evidence acquisition and capability escalation is not limited to a single model family. However, the joint router's additional gain over ASK-only does not show the same cross-model stability. This gain is statistically supported in the main Llama experiment, while the corresponding difference is not significant in the Qwen model pair. The capability channel also differs substantially across the two model pairs. On the held-out test cohort, escalation with Llama repairs 67 cases and harms 18, whereas escalation with Qwen repairs 76 cases but harms 65. The much smaller net capability gain for Qwen may contribute to the weaker incremental benefit of joint routing over ASK-only, although this observation does not establish a causal explanation. Therefore, we believe a more stable finding is that different repair channels correspond to different repairable cases, while the benefit that joint routing obtains from this complementarity depends on the specific model pair. The actual gain may depend on factors such as the quality of the repairability estimators, differences in behavior between the base model and the stronger model, and the proportion of cases in which DEFER can provide additional value.

The results of sequential ASK illustrate that more evidence does not necessarily lead to better final decisions. Each evidence acquisition changes the current state and may further change the model's response to subsequent information, so a fixed increase in ASK budget does not guarantee that accuracy increases monotonically. This result supports explicitly deciding when to continue asking and when to stop in multi-step evidence acquisition, rather than using the same fixed acquisition budget for all cases. Separately, the cost-aware experiment shows that repairability cannot be considered independently of intervention cost. When the relative costs of evidence acquisition and capability escalation change, even if the repairability estimates remain the same, the router also chooses different combinations of ASK, DEFER, and DECIDE. Therefore, repairability-aware routing is better viewed as a decision mechanism that depends on specific deployment conditions rather than as a fixed action ranking.

Several limitations remain. First, the primary experiments cover four MIMIC-IV-Ext diagnoses, with sequential acquisition evaluated separately on DDXPlus, so the results may not generalize to other clinical tasks or data sources. Second, ASK is restricted to temporally admissible VITALS, LABS, and IMAGING rather than unrestricted conversational questioning. Third, DEFER is implemented as stronger-model routing, although real systems may escalate to clinicians, specialists, multimodal models, or tool-augmented workflows. Finally, the oracle–deployable gap and model-pair dependence of joint-routing gains indicate that repairability estimation remains an important limitation.

REFERENCES

Decide, ask, or defer: Clinical llms under incomplete evidence, 2026. Manuscript under review.

Lingjiao Chen, Matei Zaharia, and James Zou. Frugalgpt: How to use large language models while reducing cost and improving performance. *Transactions on Machine Learning Research*, 2024.

Dujian Ding, Ankur Mallick, Chi Wang, Robert Sim, Subhabrata Mukherjee, Victor Rühle, Laks V. S. Lakshmanan, and Ahmed Hassan Awadallah. Hybrid llm: Cost-efficient and quality-aware query routing. In *International Conference on Learning Representations*, 2024.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. *arXiv preprint arXiv:1705.08500*, 2017.

Yonatan Geifman and Ran El-Yaniv. Selectivenet: A deep neural network with an integrated reject option. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2151–2159. PMLR, 2019.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Osman Berke Guney, Ketan Suhaas Saichandran, Karim Elzokm, Ziming Zhang, and Vijaya B. Kolachalama. Active feature acquisition via explainability-driven ranking. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 20748–20765. PMLR, 2025.

Neha Gupta, Harikrishna Narasimhan, Wittawat Jitkrittum, Ankit Singh Rawat, Aditya Krishna Menon, and Sanjiv Kumar. Language model cascades: Token-level uncertainty and beyond. In *International Conference on Learning Representations*, 2024.

Paul Hager, Friederike Jungmann, Robbie Holland, Kunal Bhagat, Inga Hubrecht, Manuel Knauer, Jakob Vielhauer, Marcus Makowski, Rickmer Braren, Georgios Kaissis, and Daniel Rueckert. Evaluation and mitigation of the limitations of large language models in clinical decision-making. *Nature Medicine*, 30:2613–2622, 2024a. doi: 10.1038/s41591-024-03097-1.

Paul Hager, Friederike Jungmann, and Daniel Rueckert. Mimic-iv-ext clinical decision making: A mimic-iv derived dataset for evaluation of large language models on the task of clinical decision making for abdominal pathologies. *PhysioNet*, 2024b. doi: 10.13026/ztyg-ah64. Version 1.1.

Alistair E. W. Johnson, Lucas Bulgarelli, Lu Shen, et al. Mimic-iv, a freely accessible electronic health record dataset. *Scientific Data*, 10:1, 2023. doi: 10.1038/s41597-022-01899-x.

Sravanthi Machcha, Sushrita Yerra, Sahil Gupta, Aishwarya Sahoo, Sharmin Sultana, Hong Yu, and Zonghai Yao. Knowing when to abstain: Medical llms under clinical uncertainty. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics*, 2026.

Isaac Ong, Amjad Almahairi, Vincent Wu, Wei-Lin Chiang, Tianhao Wu, Joseph E. Gonzalez, M. Waleed Kadous, and Ion Stoica. Routellm: Learning to route llms from preference data. In *International Conference on Learning Representations*, 2025.

Maytal Saar-Tsechansky, Prem Melville, and Foster Provost. Active feature-value acquisition. *Management Science*, 55(4):664–684, 2009. doi: 10.1287/mnsc.1080.0952.

Karan Singhal, Shekoofeh Azizi, Tao Tu, et al. Large language models encode clinical knowledge. *Nature*, 620:172–180, 2023. doi: 10.1038/s41586-023-06291-2.

Arsene Fansi Tchango, Rishab Goel, Zhi Wen, Julien Martel, and Joumana Ghosn. Ddxplus: A new dataset for automatic medical diagnosis. In *Advances in Neural Information Processing Systems*, volume 35, pp. 31306–31318, 2022.

Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. Large language models in medicine. *Nature Medicine*, 29:1930–1940, 2023. doi: 10.1038/s41591-023-02448-8.

Michael Valancius, Maxwell Lennon, and Junier Oliva. Acquisition conditioned oracle for non-greedy active feature acquisition. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 48957–48975. PMLR, 2024.

An Yang, Anfeng Li, Baosong Yang, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Table 1: Patient-level split used for the MIMIC-IV-Ext experiments. The validation split in the experimental artifacts is referred to as the development split in the paper.

Split	Total	Appendicitis	Cholecystitis	Pancreatitis	Diverticulitis
Train	976	437	272	178	89
Development	327	146	91	60	30
Test	327	146	91	60	30

APPENDIX

A DATASET CONSTRUCTION AND EXPERIMENTAL DETAILS

A.1 MIMIC-IV-EXT COHORT CONSTRUCTION

Our primary experiments use the MIMIC-IV-Ext Clinical Decision Making dataset (Hager et al., 2024b;a), which is derived from MIMIC-IV (Johnson et al., 2023). We construct a final cohort of 1,630 patient-level cases and divide it into patient-disjoint training, development, and test splits containing 976, 327, and 327 cases, respectively. The split labeled `validation` in the experimental artifacts is referred to as the development split throughout the paper.

The task contains four diagnostic categories: appendicitis, cholecystitis, pancreatitis, and diverticulitis. The training cohort contains 437 appendicitis cases, 272 cholecystitis cases, 178 pancreatitis cases, and 89 diverticulitis cases. The development cohort contains 146, 91, 60, and 30 cases from these four categories, respectively. The held-out test cohort has the same category counts. All model selection, estimator fitting, threshold selection, and ASK-priority selection are performed without using the held-out test cohort.

A.2 CLINICAL STATE CONSTRUCTION AND ASK ADMISSIBILITY

The initial clinical state contains only evidence available before active acquisition: chief complaint, age, and sex. Additional patient evidence can be acquired through three predefined ASK channels: VITALS, LABS, and IMAGING.

To prevent future information leakage, each candidate evidence item is checked against a temporal cutoff. We define the admissibility cutoff as

$$t_{\text{cutoff}} = \min\left(t_{\text{admission}}, t_{\text{ED out}}\right).$$

Only evidence available no later than this cutoff is eligible for acquisition. Evidence recorded after the cutoff is excluded from the admissible action set. ASK therefore operates over temporally available patient evidence rather than unrestricted question generation.

For a clinical state s, the admissible action set $\mathcal{Q}(s)$ contains only evidence acquisitions that satisfy these temporal constraints. When no admissible evidence item remains, ASK is unavailable to the router.

A.3 MODEL CONFIGURATION

The primary model pair consists of Llama-3.1-8B-Instruct as the base model and Llama-3.1-70B-Instruct as the stronger model (Grattafiori et al., 2024). We use Qwen3-8B and Qwen3-32B (Yang et al., 2025) as an independent cross-family robustness pair.

The Qwen evaluation uses the same MIMIC split, initial-state construction, ASK channels, temporal admissibility constraints, and output-parsing procedure as the primary Llama evaluation. Qwen models are evaluated using NF4 4-bit quantization with double quantization, float16 computation, and `enable_thinking=False`.

All deployable components are selected using development data only and frozen before final evaluation. For the Llama model pair, the frozen thresholds for current decision risk, evidence repairability,

Table 2: Frozen model and routing configuration used in the MIMIC-IV-Ext experiments. All thresholds and ASK priorities are selected using development data only.

Component	Llama	Qwen
Base model	Llama-3.1-8B-Instruct	Qwen3-8B
Stronger model	Llama-3.1-70B-Instruct	Qwen3-32B
τ_R	0.30	0.30
τ_{D_Q}	0.20	0.20
τ_{V_H}	0.40	0.50
ASK priority	IMAGING > VITALS > LABS	IMAGING > LABS > VITALS

and capability repairability are 0.3, 0.2, and 0.4, respectively. The ASK priority is IMAGING, followed by VITALS, then LABS. For the Qwen model pair, the corresponding thresholds are 0.3, 0.2, and 0.5, with ASK priority IMAGING, followed by LABS, then VITALS.

A.4 Evaluation Protocol

The held-out test cohort is not used for estimator training, model selection, threshold selection, or ASK-priority selection. All deployable choices are determined using the development split and then frozen before final evaluation.

The primary Llama evaluation uses the frozen development-selected configuration throughout the held-out test analysis. The Qwen evaluation follows the same split structure and admissibility protocol, with independently selected development thresholds and ASK priority. This separation between development mechanism analyses and held-out test results is maintained throughout the paper.

For paired comparisons between deployable policies, we report paired bootstrap confidence intervals and exact McNemar tests where applicable. Oracle quantities and oracle routing results use realized post-intervention outcomes and are included only for hindsight analysis and estimation of remaining headroom; they are not available to the deployed router.

B Repairability Estimation and Router Implementation

B.1 Estimation Targets

The deployable router uses three learned quantities corresponding to the repairability quantities defined in Section 3:

$$\hat{R}(s, M), \qquad \hat{D}_Q(s, M), \qquad \hat{V}_H(s, M).$$

The current-risk estimator $\hat{R}$ predicts whether the base model's current decision is incorrect. The evidence-repairability estimator $\hat{D}_Q$ estimates whether admissible evidence acquisition can repair the current decision. The capability-repairability estimator $\hat{V}_H$ estimates the net value of escalation to the stronger model while accounting for both repair and escalation-induced harm.

For evidence acquisition, supervision is derived from the realized change in correctness following an admissible ASK action. The hindsight evidence repairability quantity is

$$D_Q^{\mathrm{HO}}(s, M) = \max\left(0, \max_{q \in \mathcal{Q}(s)} \Delta_Q(q \mid s, M)\right).$$

This quantity is used for supervision and mechanism analysis only. It is not available to the deployed router.

For capability escalation, the realized transition is

$$A_H(s, M) = Y_M(s) - Y_H(s) \in \{-1, 0, +1\},$$

where $+1$ denotes repair, 0 denotes no change in correctness, and -1 denotes escalation-induced harm. At deployment time, the capability estimator produces probabilities for these three outcomes. The signed capability-repairability estimate is

$$\hat{V}_H(s, M) = \hat{P}(A_H = +1 \mid s, M) - \hat{P}(A_H = -1 \mid s, M).$$

Table 3: Five-fold out-of-fold estimator ablation on the Qwen development cohort. The capability-estimator AUROC and AUPRC evaluate prediction of the escalation-repair event. A dash indicates a metric not reported in the frozen ablation artifact.

Target	Estimator	AUROC	AUPRC	Brier	Log loss
R	Logistic	0.7591	0.8435	0.1935	0.5713
	MLP	0.5912	0.6830	0.2365	0.6659
D_Q	Logistic	0.7697	0.7247	–	–
	MLP	0.7086	0.6637	–	–
V_H repair event	Logistic	0.8068	0.6768	–	–
	MLP	0.6384	0.4461	–	–

The frozen estimator features are age, sex, chief complaint, and the base-model prediction. No feature is added or modified after the router configuration is frozen.

B.2 Estimator Training and Out-of-Fold Evaluation

Estimator development is performed without using the held-out test cohort. For the Qwen robustness evaluation, we use five-fold out-of-fold prediction on the development cohort to evaluate estimator quality without evaluating an estimator on the same observations used for fitting.

We compare logistic regression with an MLP estimator. Across all three prediction targets, the more expressive MLP does not improve performance over logistic regression. Table 3 reports the out-of-fold results.

For R, logistic regression achieves an AUROC of 0.7591 and an AUPRC of 0.8435, compared with 0.5912 and 0.6830 for the MLP. For D_Q, the corresponding AUROC/AUPRC values are 0.7697/0.7247 for logistic regression and 0.7086/0.6637 for the MLP. For the capability estimator, the reported AUROC and AUPRC evaluate prediction of the escalation-repair event. Logistic regression reaches 0.8068/0.6768, compared with 0.6384/0.4461 for the MLP.

These results show that the routing results are not driven by estimator complexity. We therefore retain the simpler development-selected estimator configuration for frozen evaluation.

B.3 Threshold Selection and Frozen ASK Priority

For the primary Llama router, all routing thresholds are selected on the 327-case development cohort before held-out test evaluation. The original threshold search uses

$$\tau_R \in \{0.3, 0.4, 0.5, 0.6, 0.7\},$$

$$\tau_{D_Q} \in \{0.2, 0.3, 0.4, 0.5, 0.6\},$$

and

$$\tau_{V_H} \in \{-0.1, 0, 0.1, 0.2, 0.3, 0.4, 0.5\}.$$

Candidate configurations are ranked lexicographically by the following development objectives:

1. maximize accuracy;
2. minimize intervention rate;
3. minimize DEFER rate;
4. maximize the risk threshold τ_R.

The resulting frozen primary thresholds are

$$\tau_R = 0.30, \qquad \tau_{D_Q} = 0.20, \qquad \tau_{V_H} = 0.40.$$

The Qwen robustness evaluation uses an independently frozen configuration with

$$\tau_R = 0.30, \qquad \tau_{D_Q} = 0.20, \qquad \tau_{V_H} = 0.50.$$

Table 4: Development-cohort ASK-channel statistics used to freeze the primary Llama ASK priority.

Channel	n	Mean Δ_Q	Repair rate	Harm rate
IMAGING	247	0.3522	0.3725	0.0202
VITALS	311	-0.0418	0.0740	0.1158
LABS	71	-0.1831	0.1268	0.3099

Table 5: Post hoc development-only sensitivity to the lower threshold-grid boundaries. The expanded-grid configuration is not used for held-out test evaluation.

Metric	Frozen grid	Expanded grid
τ_R	0.30	0.20
τ_{D_Q}	0.20	0.00
τ_{V_H}	0.40	0.40
Accuracy	68.50%	70.95%
Intervention rate	74.31%	85.63%
ASK count	232	269
DEFER count	11	11
DECIDE count	84	47

The complete frozen configurations for both model pairs are summarized in Appendix A.

ASK-channel priority is also frozen from development data. For each channel, we compute the mean realized correctness change Δ_Q. The development statistics for the primary Llama configuration are shown in Table 4.

Legal ASK channels are ranked by development-cohort mean realized Δ_Q, with exact ties broken lexicographically. This yields the frozen primary priority

$$\text{IMAGING} > \text{VITALS} > \text{LABS}.$$

At deployment time, the router selects the highest-priority channel that is legal for the current case.

The Qwen robustness evaluation independently freezes the priority

$$\text{IMAGING} > \text{LABS} > \text{VITALS}.$$

Post hoc boundary sensitivity. Because the frozen values $\tau_R = 0.30$ and $\tau_{D_Q} = 0.20$ lie at the lower boundaries of their original search grids, we perform a post hoc development-only sensitivity analysis with expanded lower bounds. We extend the risk grid to

$$\tau_R \in \{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7\}$$

and the evidence-repairability grid to

$$\tau_{D_Q} \in \{0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6\},$$

while leaving the τ_{V_H} grid unchanged.

Under the same lexicographic development objective, the expanded search selects

$$(\tau_R, \tau_{D_Q}, \tau_{V_H}) = (0.20, 0.00, 0.40).$$

Development accuracy increases from 68.50% under the original frozen configuration to 70.95%, while the intervention rate increases from 74.31% to 85.63%. The number of ASK decisions increases from 232 to 269. The number of DEFER decisions remains 11.

Table 5 summarizes this analysis.

This result indicates that the original operating point is partly constrained by the lower boundaries of the prespecified threshold grid. The expanded-grid solution also corresponds to a more intervention-heavy policy. We report this analysis only as a post hoc development sensitivity check. We do not replace the pre-frozen thresholds or reevaluate the held-out test cohort under the expanded-grid configuration.

B.4 Exact Frozen Deployable Routing Rule

The frozen router separates channel eligibility from final action selection. It first applies the current-risk gate. If

$$\hat{R}(s, M) < \tau_R,$$

the router selects DECIDE and no repair action is considered.

For states satisfying

$$\hat{R}(s, M) \geq \tau_R,$$

ASK and DEFER eligibility are evaluated independently. Define

$$E_Q(s) = \mathbb{I}\left[\mathrm{ASK}_{\mathrm{legal}}(s) \wedge \hat{D}_Q(s, M) \geq \tau_{D_Q}\right]$$

and

$$E_H(s) = \mathbb{I}\left[\hat{V}_H(s, M) \geq \tau_{V_H}\right].$$

Thus, ASK eligibility requires both an admissible evidence acquisition and sufficient estimated evidence repairability. DEFER eligibility depends only on the frozen capability-repairability threshold.

Eligibility does not by itself determine the final action when both repair channels qualify. If only ASK is eligible, the router selects ASK. If only DEFER is eligible, it selects DEFER. If both are eligible, the router compares their estimated repairability values and selects ASK when

$$\hat{D}_Q(s, M) \geq \hat{V}_H(s, M),$$

and DEFER otherwise. Exact equality is therefore resolved in favor of ASK.

The complete frozen policy is

$$\pi(s) = \begin{cases} \text{DECIDE}, & \hat{R} < \tau_R, \\ \text{ASK}, & \hat{R} \geq \tau_R,\ E_Q = 1,\ E_H = 0, \\ \text{DEFER}, & \hat{R} \geq \tau_R,\ E_Q = 0,\ E_H = 1, \\ \text{ASK}, & \hat{R} \geq \tau_R,\ E_Q = 1,\ E_H = 1,\ \hat{D}_Q \geq \hat{V}_H, \\ \text{DEFER}, & \hat{R} \geq \tau_R,\ E_Q = 1,\ E_H = 1,\ \hat{D}_Q < \hat{V}_H, \\ \text{DECIDE}, & \hat{R} \geq \tau_R,\ E_Q = 0,\ E_H = 0. \end{cases}$$

The threshold conditions therefore determine which repair channels enter the candidate set, while the comparison between $\hat{D}_Q$ and $\hat{V}_H$ resolves cases in which both channels are eligible.

Once ASK is selected, the router chooses the highest-ranked legal evidence channel under the frozen development-derived ASK priority described in Appendix B.3.

B.5 Held-Out Evaluation Safeguards

The held-out test implementation loads the frozen thresholds, estimator artifacts, feature specification, and ASK priority without modification. The following choices are prohibited after the router is frozen: threshold retuning, feature changes, ASK-priority changes, model-family changes based on test outcomes, and inspection of test labels before predictions are frozen.

The router does not have access to D_Q^{HO}, realized A_H, or any other hindsight intervention outcome at deployment time. These quantities are used only for supervision, evaluation, or oracle analysis.

The primary threshold-based router is distinct from the utility-based cost-aware router evaluated later in the paper. The cost-aware policy selects among

$$0, \qquad \hat{D}_Q - c_Q, \qquad \hat{V}_H - c_H$$

after the frozen risk gate and subject to ASK legality. Consequently, the zero-cost operating point of the cost-aware experiment is not the same policy as the primary threshold-based router.

Table 6: Repair-set decomposition for the primary Llama model pair and the Qwen cross-family robustness pair. The Llama decomposition uses the development cohort, whereas the Qwen decomposition uses the held-out test cohort.

Repair-set category	**Llama development**	**Qwen test**
ASK-only	60	76
DEFER-only	21	25
Both	41	51
Neither	205	175
ASK-repairable total	101	127
DEFER-repairable total	62	76
Union	122	152
Total	327	327

C Additional Repairability Analysis

C.1 Repair-Set Decomposition

We analyze whether ASK and DEFER repair the same cases or provide distinct repair opportunities. For each case, we define two binary indicators according to whether the realized intervention can repair an incorrect base-model decision:

$$Z_Q(s) = \mathbb{I}\left[D_Q^{\mathrm{HO}}(s, M) > 0\right],$$

and

$$Z_H(s) = \mathbb{I}\left[A_H(s, M) = +1\right].$$

These indicators partition cases into four mutually exclusive groups: ASK-only, DEFER-only, Both, and Neither.

By construction, $D_Q^{\mathrm{HO}}(s, M) > 0$ can occur only when the base-model decision is incorrect. Therefore, on a fixed cohort, the ASK-repairable total is exactly the number of baseline errors that can be repaired by at least one admissible ASK action. The two quantities differ in normalization only when reported as a fraction of all cases versus a fraction of baseline errors.

For the primary Llama analysis, we perform this decomposition on the 327-case development cohort rather than on the final held-out test cohort. This allows us to characterize the repair-channel mechanism without using the final Llama test set for mechanism discovery. As an independent cross-family robustness analysis, we perform the same decomposition on the held-out Qwen test cohort.

Table 6 reports the resulting counts.

For Llama, 101 of 327 development cases are ASK-repairable and 62 are DEFER-repairable. Their union contains 122 cases. Among these, 60 are repaired only by ASK, 21 only by DEFER, and 41 by both channels.

For Qwen, 127 of 327 held-out test cases are ASK-repairable and 76 are DEFER-repairable, with a union of 152 cases. Among these, 76 are ASK-only, 25 are DEFER-only, and 51 are repairable by both channels.

The overlap between the two repair sets is incomplete for both model pairs. The Jaccard similarity is

$$J_{\mathrm{Llama}} = \frac{41}{122} = 0.3361$$

and

$$J_{\mathrm{Qwen}} = \frac{51}{152} = 0.3355.$$

The corresponding phi coefficients are 0.3690 for Llama and 0.3191 for Qwen. ASK and DEFER repairability are therefore positively associated, but the two repair sets remain substantially non-redundant.

Table 7: Overlap statistics for ASK and DEFER repair sets.

Statistic	Llama development	Qwen test
ASK-repairable rate	30.89%	38.84%
DEFER-repairable rate	18.96%	23.24%
Union repairable rate	37.31%	46.48%
Jaccard similarity	0.3361	0.3355
Phi coefficient	0.3690	0.3191

Table 8: Hindsight evidence repairability among baseline errors on the held-out test cohorts.

Metric	Llama test	Qwen test
Test cases	327	327
Baseline errors	208	204
ASK-repairable errors	104	127
ASK-repairable error rate	50.00%	62.25%

Because the Llama and Qwen decompositions are reported on different split roles, these percentages are not intended as a direct comparison of model-family performance. Their purpose is to test whether the qualitative repair-set structure persists across independently evaluated model pairs.

C.2 Evidence Repairability Conditional on Baseline Error

The ASK-repairable counts in Table 6 can also be expressed conditional on the base model being incorrect. This changes the denominator, not the underlying definition of repairability.

On the held-out Llama test cohort, the base model makes 208 errors, of which 104 are repairable by at least one admissible ASK action:

$$\frac{104}{208} = 50.00\%.$$

This count differs from the 101 ASK-repairable cases reported for the Llama repair-set decomposition because the two quantities are computed on different cohorts. The repair-set decomposition uses the Llama development cohort, whereas the 104 repairable errors are measured on the held-out Llama test cohort.

For Qwen, both analyses use the same held-out test cohort. The 127 ASK-repairable cases in Table 6 are therefore the same 127 cases used in the conditional error analysis. Relative to all 327 test cases, they represent

$$\frac{127}{327} = 38.84\%.$$

Conditional on the 204 baseline errors, they correspond to

$$\frac{127}{204} = 62.25\%.$$

These two percentages use the same ASK-repairable numerator and differ only in denominator. They should therefore be interpreted as two views of the same repairable-case set rather than as independent measurements.

These quantities describe hindsight repairability rather than deployable router performance. They answer whether at least one admissible evidence acquisition exists that changes an incorrect decision into a correct one. The deployed router must instead estimate this opportunity from the current state using $\hat{D}_Q$.

The existence of a repairable ASK action also does not imply that every admissible evidence acquisition is beneficial. An ASK action may repair an error, leave correctness unchanged, or harm an originally correct decision. This distinction motivates estimating intervention-specific repairability rather than treating additional evidence as uniformly beneficial.

Table 9: Held-out capability-escalation transitions. Positive values denote repair, zero denotes no change in correctness, and negative values denote escalation-induced harm.

Escalation outcome	Llama test	Qwen test
$A_H = +1$ (repair)	67	76
$A_H = 0$ (unchanged)	242	186
$A_H = -1$ (harm)	18	65
Total	327	327

C.3 Capability-Repairability Transitions

DEFER is also not uniformly beneficial. We characterize escalation using

$$A_H = Y_M - Y_H \in \{-1, 0, +1\}.$$

Here,

$$A_H = +1$$

denotes a repair in which the base model is incorrect and the stronger model is correct,

$$A_H = 0$$

denotes no change in correctness, and

$$A_H = -1$$

denotes escalation-induced harm in which the base model is correct but the stronger model is incorrect.

The full held-out transition counts are reported in Table 9.

For Llama, escalation repairs 67 cases and harms 18 cases. The resulting net difference is 49 cases. For Qwen, escalation repairs 76 cases but harms 65, leaving a net difference of only 11 cases. Thus, although the Qwen stronger model repairs slightly more cases in absolute terms, those repairs are accompanied by substantially more escalation-induced harm.

This difference is also reflected in the aggregate Always-DEFER performance. For Llama, Always DEFER improves accuracy from 36.39% to 51.38%. For Qwen, the corresponding change is from 37.61% to 40.98%. These aggregate results do not determine the performance of selective DEFER routing, but they show that the net value of changing the decision maker differs substantially across the two model pairs.

This observation motivates the signed definition

$$V_H(s, M) = P(A_H = +1 \mid s, M) - P(A_H = -1 \mid s, M),$$

rather than defining capability repairability only as the probability of repair. A case can have a nonzero probability of escalation repair while still having limited or negative net escalation value when the probability of escalation-induced harm is also substantial.

C.4 Relationship Between Current Risk and Repairability

Current decision risk and repairability answer different questions. $R(s, M)$ describes whether the current decision is likely to be wrong. $D_Q(s, M)$ describes whether changing the available evidence can repair the decision. $V_H(s, M)$ describes whether changing the decision maker has positive expected value.

A high-risk state therefore need not be repairable by either channel. Similarly, knowing that a state is repairable does not determine which repair channel should be selected. The observed ASK-only, DEFER-only, Both, and Neither groups provide empirical evidence for this distinction.

The repair-set decomposition also shows why a single generic "intervene" score is insufficient for the routing problem considered here. ASK changes the available evidence, whereas DEFER changes the decision maker. Their realized repair sets overlap only partially in both model families examined here.

Table 10: Held-out accuracy of deployable routing policies for the primary Llama experiment and the Qwen3 robustness experiment. All thresholds and routing components are frozen using development data before held-out evaluation.

Policy	Llama	Qwen3
Always DECIDE	36.39%	37.61%
Always DEFER	51.38%	40.98%
R-only	50.46%	41.90%
$R + D_Q$ ASK-only	59.02%	62.08%
$R + V_H$ DEFER-only	46.79%	44.04%
$R + D_Q + V_H$ joint	**62.69%**	**63.00%**

Together, these results support the distinction

$$R(s, M) \neq D_Q(s, M) \neq V_H(s, M).$$

The three quantities describe current error risk, evidence repairability, and capability repairability, respectively, and each is required for a different part of the routing decision.

D Additional Routing Results

This appendix provides the complete deployable routing results for the primary Llama experiment and the Qwen3 cross-family robustness experiment. All reported deployable policies use estimators, thresholds, and ASK execution rules frozen on development data before evaluation on the held-out test cohort. Oracle results are reported separately and are used only to characterize remaining repair headroom.

D.1 Full Deployable Policy Comparison

Table 10 reports the complete deployable policy comparison for both model pairs. The same 327-case held-out test cohort is used within each model-pair experiment.

For the primary Llama experiment, direct prediction with the base model achieves 36.39% accuracy. Always escalating to the stronger model increases accuracy to 51.38%. A policy based only on current decision risk reaches 50.46%. Incorporating evidence repairability produces a stronger ASK-only policy at 59.02%. Using capability repairability alone yields 46.79%. The frozen joint router reaches 62.69%.

The Qwen3 robustness experiment shows a similar pattern for the evidence-acquisition policies but a smaller incremental benefit from joint routing. Qwen3-8B achieves 37.61% when always deciding, while Qwen3-32B achieves 40.98% under unconditional escalation. The risk-only policy reaches 41.90%, the ASK-only policy reaches 62.08%, the capability-repairability DEFER-only policy reaches 44.04%, and the joint router reaches 63.00%.

The comparison also illustrates why stronger capability alone is not equivalent to repairability-aware routing. For both model pairs, unconditional escalation remains substantially below the ASK-only and joint policies. The stronger model can repair some base-model errors while introducing new errors on other cases. Consequently, the availability of a stronger model does not imply that DEFER is beneficial for every state.

D.2 Paired Comparison with ASK-Only

The most relevant deployable comparison is between the joint router and the ASK-only policy because both policies already use current decision risk and evidence repairability. The joint router differs in its additional ability to select the capability-repair channel.

For the primary Llama model pair, the joint router improves accuracy from 59.02% to 62.69%, corresponding to an absolute improvement of 3.67 percentage points. The paired bootstrap 95% confidence interval for this difference is $[1.53, 5.81]$ percentage points. Across the 327 paired test

Table 11: Paired held-out comparison between the deployable ASK-only policy and the joint ASK/DEFER router.

Model pair	ASK-only	Joint	Δ accuracy	95% CI
Llama	59.02%	62.69%	+3.67 pp	$[1.53, 5.81]$
Qwen3	62.08%	63.00%	+0.92 pp	$[-0.61, 2.45]$

Table 12: Action distribution and realized outcomes of the primary Llama joint router on the 327-case held-out test cohort.

Action	Selected cases	Repairs	Harms
DECIDE	77	0	0
ASK	227	83	14
DEFER	23	17	0
Total intervention	250	100	14

cases, the joint policy obtains 13 gains relative to ASK-only and 1 loss. The exact McNemar test gives $p = 0.001831$.

For the Qwen3 model pair, ASK-only reaches 62.08% and the joint router reaches 63.00%, corresponding to an absolute difference of 0.92 percentage points. The paired bootstrap 95% confidence interval is $[-0.61, 2.45]$ percentage points, and the exact McNemar test gives $p = 0.453$. The incremental improvement is therefore not statistically significant in this second model pair.

These results separate two claims. First, evidence repairability and capability repairability identify non-redundant repair opportunities, as shown by the repair-set analyses in Appendix C. Second, the amount of additional held-out accuracy obtained by jointly routing between the two channels depends on the particular base and stronger model pair. The Llama experiment exhibits a statistically significant incremental benefit, whereas the Qwen3 experiment does not.

D.3 Primary Joint-Router Action Distribution

For the primary Llama experiment, the frozen joint router assigns 77 of 327 test cases to DECIDE, 227 to ASK, and 23 to DEFER. The resulting intervention rate is 76.45%.

Relative to the base-model decisions, the joint policy repairs 100 errors and introduces 14 new errors. The ASK branch accounts for 83 repairs and 14 harms, whereas the selected DEFER branch accounts for 17 repairs and no observed harms on the held-out cohort.

The action distribution is not intended to imply that ASK or DEFER should be used at a fixed population rate. These counts are induced by the frozen risk gate, repairability estimates, and routing thresholds for the evaluated model pair. Different model pairs can induce different intervention mixtures even when the same conceptual routing framework is used.

D.4 Selected DEFER Outcomes

We further examine the realized outcomes of cases assigned to the DEFER branch by the frozen joint router. This analysis differs from unconditional escalation, which applies the stronger model to every case regardless of the estimated capability-repairability value.

For the primary Llama model pair, the joint router selects DEFER for 23 of 327 held-out test cases. Among these selected cases, 17 repair an incorrect base-model decision, while none changes an originally correct base-model decision into an incorrect one. By comparison, unconditional escalation over the full test cohort produces 67 repairs and 18 harms.

The same analysis yields a different operating regime for the Qwen3 robustness experiment. The frozen joint router selects DEFER for 13 of 327 test cases. Of these cases, 10 repair an incorrect Qwen3-8B decision and 1 changes an originally correct base-model decision into an incorrect one.

Table 13: Realized outcomes of capability escalation on the held-out test cohorts. Selected DEFER evaluates only cases assigned to DEFER by the frozen joint router. Unconditional escalation applies the stronger model to all 327 cases.

Model pair	Escalation setting	Cases	Repairs	Harms
Llama	Selected DEFER	23	17	0
Llama	Unconditional escalation	327	67	18
Qwen3	Selected DEFER	13	10	1
Qwen3	Unconditional escalation	327	76	65

Table 14: Deployable and hindsight-oracle routing performance for the primary Llama experiment. Oracle policies are used only to characterize repair headroom and are not deployable.

Policy	Accuracy
Deployable ASK-only	59.02%
Deployable DEFER-only	46.79%
Deployable joint	62.69%
Oracle ASK-only	68.20%
Oracle DEFER-only	56.88%
Oracle joint	75.84%

Unconditional escalation from Qwen3-8B to Qwen3-32B produces 76 repairs and 65 harms over the full test cohort.

These results illustrate why capability escalation should not be treated as uniformly beneficial. For both model pairs, the joint router applies DEFER to only a subset of the cases for which stronger capability is available. The selected subsets have more favorable realized escalation outcomes than unconditional escalation over the complete cohort.

The two model pairs also differ substantially in the overall value of the capability channel. Unconditional escalation with Llama produces 67 repairs and 18 harms, corresponding to a net difference of 49 favorable transitions. The corresponding Qwen3 transition counts are 76 repairs and 65 harms, giving a net difference of 11. This difference is consistent with the smaller incremental benefit of joint routing over ASK-only in the Qwen3 robustness experiment, although it does not by itself establish a causal explanation for that difference.

The purpose of $\hat{V}_H$ is therefore not to identify whether the stronger model is globally more accurate. Instead, it estimates whether changing the decision maker is expected to provide positive value for the current state. This distinction is important because a stronger model can repair some base-model errors while simultaneously introducing new errors on other cases.

D.5 Oracle Headroom

We additionally evaluate hindsight oracle policies to characterize the amount of repair headroom available under the current intervention space. These results are not deployable because they use realized post-intervention outcomes to select actions.

On the primary Llama held-out test cohort, oracle ASK-only reaches 68.20% accuracy, oracle DEFER-only reaches 56.88%, and oracle joint routing reaches 75.84%. The gap between the deployable joint router and the oracle joint policy therefore indicates that substantial headroom remains in estimating where each repair channel will be beneficial.

The oracle comparison should not be interpreted as an estimate of future deployable performance. It instead isolates the value of improved repairability estimation under the currently defined ASK and DEFER action spaces. In particular, the oracle joint policy shows that the two repair channels provide additional complementary repair opportunities beyond either channel alone.

Table 15: Diagnostic accuracy under fixed-order and random-order sequential ASK budgets on DDXPlus. Increasing the ASK budget does not monotonically improve accuracy.

ASK budget	Fixed order	Random order
$B = 0$	14.75%	14.75%
$B = 1$	15.25%	16.00%
$B = 2$	16.25%	14.75%
$B = 3$	15.75%	13.50%
$B = 5$	16.75%	15.75%

D.6 Summary of Routing Results

The additional routing results support three observations.

First, evidence acquisition provides the dominant deployable improvement over direct prediction in both evaluated model pairs. ASK-only increases accuracy from 36.39% to 59.02% for Llama and from 37.61% to 62.08% for Qwen3.

Second, capability escalation is not uniformly beneficial. Applying the stronger model to all cases produces both repairs and harms. The realized balance differs substantially across model pairs: unconditional Llama escalation produces 67 repairs and 18 harms, whereas unconditional Qwen3 escalation produces 76 repairs and 65 harms.

Third, joint routing can exploit repair opportunities that are not captured by ASK alone, but the size of this incremental benefit is model-pair dependent. The primary Llama joint router improves over ASK-only by 3.67 percentage points with a statistically significant paired difference. The corresponding Qwen3 improvement is 0.92 percentage points and is not statistically significant.

Together with the repair-set analysis, these results motivate the separation of current decision risk, evidence repairability, and capability repairability. Risk determines whether intervention may be needed, while the repairability signals distinguish where that intervention is expected to be useful.

E Sequential Evidence Acquisition Details

The primary MIMIC experiments evaluate a single-step intervention: the system either decides from the current state, acquires additional evidence, or defers to a stronger model. We additionally study a sequential setting on DDXPlus in which multiple ASK actions may be available before a final decision is made.

These experiments address two questions. First, does increasing the number of available ASK steps necessarily improve diagnostic accuracy? Second, can a learned stopping policy reduce unnecessary evidence acquisition while preserving diagnostic performance?

E.1 Fixed and Random ASK Budgets

We evaluate fixed ASK budgets $B \in \{0, 1, 2, 3, 5\}$ under two evidence-ordering conditions. In the fixed-order condition, evidence follows the predefined trajectory order. In the random-order condition, the acquisition order is randomized.

Table 15 reports the resulting diagnostic accuracies.

The relationship between budget and accuracy is non-monotonic in both conditions. Under fixed ordering, accuracy rises from 15.25% at $B = 1$ to 16.25% at $B = 2$, decreases to 15.75% at $B = 3$, and then increases to 16.75% at $B = 5$. Under random ordering, the highest accuracy among the evaluated nonzero budgets occurs at $B = 1$.

These results do not imply that additional evidence is generally harmful. They show instead that increasing the ASK budget alone does not guarantee improved diagnostic performance.

Table 16: Comparison of stopping strategies on the fixed-order DDXPlus trajectories. Always STOP terminates at the initial state.

Policy	Accuracy	Mean questions
Always STOP ($B = 0$)	14.75%	0.00
Learned stopping	17.25%	3.1825
Fixed $B = 5$	16.75%	5.00

E.2 Learned Stopping

A fixed evidence budget cannot adapt to the amount of information required by an individual trajectory. We therefore evaluate a learned stopping policy that predicts whether evidence acquisition should continue from the current observable state.

The stopping target is defined on a realized five-question behavior trajectory. At state s_t,

$$\text{CONTINUE}(s_t) = 1$$

if and only if some later prefix on the same trajectory attains lower realized diagnostic error than stopping at the current state. Future outcomes and error labels are used only to define the supervision target and are not used as inference-time features.

On the fixed-order trajectories, learned stopping achieves 17.25% diagnostic accuracy while using an average of 3.1825 questions. The fixed $B = 5$ policy uses exactly 5.00 questions and achieves 16.75% accuracy.

The observed accuracy difference is +0.50 percentage points. The paired bootstrap 95% confidence interval is $[-1.00, 2.00]$ percentage points. Across the 400 paired trajectories, learned stopping produces 6 gains and 4 losses relative to fixed $B = 5$, yielding an exact McNemar test of

$$p = 0.754.$$

We therefore find no evidence of an accuracy decrease relative to the five-question policy.

To determine whether the reduction in acquisition can be explained by the strong class imbalance alone, we additionally compare against a trivial Always STOP policy. Because stopping is permitted at the initial state, Always STOP is equivalent to the same-path $B = 0$ reference. It uses no additional questions and achieves 14.75% accuracy.

Learned stopping improves over Always STOP by 2.50 percentage points. The paired bootstrap 95% confidence interval is $[0.50, 4.75]$ percentage points. Learned stopping produces 15 paired gains and 5 paired losses over the 400 trajectories, corresponding to an exact McNemar test of

$$p = 0.0414.$$

Relative to fixed $B = 5$, learned stopping reduces the mean number of questions by approximately 36.4%.

The continuation target is highly imbalanced. Only 3.48% of the evaluated decision states are labeled as requiring another acquisition. The learned continuation score achieves an AUROC of 0.592 and an AUPRC of 0.043.

These estimator-level metrics place an important limit on the interpretation of the stopping experiment. The stopping estimator has only limited discrimination of rare continuation states, and part of its behavior is likely influenced by the strong prior toward stopping. However, the resulting policy is not equivalent to a trivial majority-class strategy. Always STOP reaches 14.75% accuracy, whereas learned stopping reaches 17.25% and shows a statistically significant paired advantage over that baseline.

Our supported conclusion is therefore narrower than claiming strong state-level stopping prediction. Under the evaluated protocol, learned stopping reduces acquisition relative to fixed $B = 5$ without evidence of lower accuracy, while also outperforming a trivial Always STOP policy.

E.3 Interpretation of Sequential Acquisition

Sequential ASK changes the observable state after every acquisition. If evidence e_t is acquired at step t, then

$$s_{t+1} = s_t \cup \{e_t\}.$$

The value of continuing acquisition should therefore be reconsidered after each state update. Evidence that is useful at one stage does not imply that all subsequently available evidence remains useful.

The fixed-budget results show that simply increasing the number of questions does not monotonically improve diagnostic accuracy. The learned-stopping results further show that a state-dependent stopping policy can reduce acquisition relative to the maximum evaluated budget, although accurate identification of rare continuation states remains difficult.

F Full Cost-Aware Routing Results

We examine how explicit intervention costs change the allocation between DECIDE, ASK, and DEFER in the primary MIMIC setting.

The cost-aware rule is

$$a^*(s) = \arg\max \left\{ 0,\ \hat{D}_Q(s, M) - c_Q,\ \hat{V}_H(s, M) - c_H \right\},$$

subject to the frozen risk gate and ASK-legality constraint. Here, c_Q denotes the normalized cost of evidence acquisition and c_H denotes the normalized cost of capability escalation.

We evaluate

$$c_Q, c_H \in \{0,\ 0.05,\ 0.10,\ 0.20,\ 0.30,\ 0.40\},$$

yielding 36 cost configurations. These values are normalized policy costs and should not be interpreted as monetary quantities.

F.1 Full Cost Grid

Table 17 reports held-out accuracy for all 36 cost configurations. Table 18 reports the corresponding DECIDE, ASK, and DEFER counts.

The full grid shows that costs change both whether the router intervenes and which repair channel it selects. Increasing the relative cost of ASK shifts actions away from evidence acquisition. Increasing the relative cost of DEFER shifts actions away from capability escalation. When neither intervention retains sufficient estimated net value, DECIDE becomes more frequent.

Accuracy itself is not monotonic in either cost. We therefore do not interpret the grid as a held-out cost-selection procedure. Its purpose is to characterize how the frozen repairability estimates translate into different action mixtures under different relative intervention costs.

F.2 Effect of ASK Cost

To isolate the effect of evidence-acquisition cost, we fix $c_H = 0$ and vary c_Q.

As c_Q increases from 0 to 0.40 while c_H remains fixed at zero, ASK decreases from 233 to 51 cases. DEFER increases from 28 to 143, while DECIDE increases from 66 to 133.

Table 17: Held-out accuracy across the complete 6×6 intervention-cost grid. Rows vary the ASK cost c_Q and columns vary the DEFER cost c_H.

c_Q \ c_H	0.00	0.05	0.10	0.20	0.30	0.40
0.00	64.53%	63.30%	62.39%	60.55%	60.55%	60.55%
0.05	64.83%	64.53%	63.30%	60.55%	60.55%	60.55%
0.10	64.22%	64.83%	64.53%	62.39%	60.55%	60.55%
0.20	60.55%	60.86%	63.00%	63.30%	61.16%	59.63%
0.30	54.74%	56.57%	57.80%	60.24%	60.24%	58.41%
0.40	54.13%	53.82%	54.13%	57.19%	59.02%	59.33%

Table 18: Action composition across the complete cost grid. Each entry reports DECIDE / ASK / DEFER counts over the 327-case held-out cohort.

c_Q \ c_H	0.00	0.05	0.10	0.20	0.30	0.40
0.00	66/233/28	66/240/21	66/247/14	67/253/7	67/255/5	68/255/4
0.05	66/226/35	66/233/28	66/240/21	67/253/7	67/254/6	68/255/4
0.10	66/219/42	66/226/35	66/233/28	67/247/13	67/253/7	68/255/4
0.20	73/180/74	73/201/53	74/212/41	75/225/27	76/239/12	77/244/6
0.30	112/88/127	113/118/96	115/141/71	116/171/40	118/184/25	120/197/10
0.40	133/51/143	135/55/137	138/66/123	139/118/70	142/147/38	146/158/23

This pattern illustrates substitution between repair channels. When evidence acquisition becomes relatively more costly, some states that would otherwise be assigned to ASK move to DEFER when capability repair remains valuable. Other states return to DECIDE when neither intervention provides sufficient estimated net value.

F.3 Effect of DEFER Cost

We next fix $c_Q = 0$ and vary c_H.

Increasing c_H reduces DEFER from 28 cases at zero cost to 4 cases at $c_H = 0.40$. ASK absorbs most of these interventions, increasing from 233 to 255 cases, while DECIDE changes only slightly.

The two slices therefore exhibit complementary behavior. Increasing c_Q suppresses ASK and makes DEFER relatively more competitive, whereas increasing c_H suppresses DEFER and makes ASK relatively more competitive.

F.4 Repairs and Harms Across the Cost Grid

The action changes induced by the cost grid also alter the realized balance between repairs and harms.

At $(c_Q, c_H) = (0, 0)$, the cost-aware policy produces 105 repairs and 13 harms relative to the base model, corresponding to 92 net repairs. At $(0.05, 0)$, the policy produces 106 repairs and 13 harms, for 93 net repairs.

At $(0.20, 0.20)$, the policy produces 101 repairs and 13 harms, corresponding to 88 net repairs. At $(0.40, 0.40)$, it produces 84 repairs and 9 harms, corresponding to 75 net repairs.

These values are reported descriptively. We do not use the held-out cost sweep to select a new operating point.

Table 19: Cost-aware routing with $c_H = 0$ while varying the ASK cost c_Q.

c_Q	DECIDE	ASK	DEFER	Accuracy
0.00	66	233	28	64.53%
0.05	66	226	35	64.83%
0.10	66	219	42	64.22%
0.20	73	180	74	60.55%
0.30	112	88	127	54.74%
0.40	133	51	143	54.13%

Table 20: Cost-aware routing with $c_Q = 0$ while varying the DEFER cost c_H.

c_H	DECIDE	ASK	DEFER	Accuracy
0.00	66	233	28	64.53%
0.05	66	240	21	63.30%
0.10	66	247	14	62.39%
0.20	67	253	7	60.55%
0.30	67	255	5	60.55%
0.40	68	255	4	60.55%

F.5 Relation to the Primary Frozen Router

The zero-cost cost-aware policy reaches 64.53% accuracy. This differs from the 62.69% accuracy of the primary frozen joint router because the two experiments use different decision rules.

The primary router first applies the frozen risk threshold and then uses separately frozen repairability thresholds

$$\tau_R, \qquad \tau_{D_Q}, \qquad \tau_{V_H}$$

to determine intervention eligibility.

The E6 cost-aware analysis instead applies

$$\arg\max \left\{0, \hat{D}_Q - c_Q, \hat{V}_H - c_H\right\}$$

after the frozen risk gate and ASK-legality constraint.

Consequently,

$$c_Q = c_H = 0$$

does not recover the primary E4 threshold router. The cost sweep is therefore interpreted as a cost-sensitive policy-behavior analysis, not as a replacement for or post hoc optimization of the primary held-out router.

G Invalid-Output Robustness

Cross-family robustness, repair-set overlap, estimator ablations, and threshold sensitivity are reported in earlier appendices. We therefore restrict this section to the additional robustness issue specific to the sequential DDXPlus experiment: invalid model outputs.

G.1 Invalid Output Frequency

Across the sequential experiments, we observe 34 invalid-output occurrences. These affect 26 trajectories, corresponding to 12 patients and 17 unique trajectory states.

Table 21: Invalid outputs observed in the sequential DDXPlus experiments.

Statistic	Count
Invalid-output occurrences	34
Affected trajectories	26
Affected patients	12
Unique affected states	17

All observed invalid outputs occur in the random-order condition. No invalid outputs are observed in the fixed-order trajectories under the evaluated protocol.

We retain these failures in the robustness analysis rather than silently dropping affected states or trajectories.

G.2 Sensitivity to Invalid-Output Handling

We evaluate alternative treatments of invalid model outputs to test whether the sequential conclusions depend on a single fallback or parsing convention.

Across the evaluated alternatives, the principal qualitative findings are unchanged. Increasing the ASK budget still does not monotonically improve diagnostic accuracy. The reduction in acquisition produced by learned stopping relative to the fixed five-question policy is also preserved.

The sequential conclusions are therefore not an artifact of one specific invalid-output treatment. This does not imply that invalid outputs are unimportant for deployment. It indicates only that the reported comparison among the evaluated acquisition strategies is robust to the tested handling alternatives.

G.3 Scope

The robustness analyses in the paper address distinct sources of variation. The Qwen3 experiment tests whether the non-redundant structure of ASK and DEFER appears under a second model family. Estimator ablations test whether development-set repairability prediction depends on the logistic-regression estimator. Threshold-boundary analysis tests sensitivity of the development operating point to the original search grid. The analysis above tests whether sequential conclusions depend on one specific treatment of invalid model outputs.

These experiments support the reported mechanism-level findings under the evaluated alternatives. They do not establish invariance across arbitrary model families, datasets, clinical tasks, or prompting protocols.

H Prompt Templates and Output Parsing

This appendix documents the model-visible information used in the MIMIC experiments and the provenance available for the sequential DDXPlus experiments.

H.1 MIMIC Initial-State Diagnosis Prompt

All MIMIC diagnostic calls use a single `user` message. No additional system message is supplied.

The frozen initial-state prompt has the following form:

```
You are given an emergency-department presentation
from a patient who was ultimately diagnosed with one
of four conditions.
Choose the single most likely diagnosis using only the
information provided.
Allowed diagnoses:
```

```
1. appendicitis
2. cholecystitis
3. pancreatitis
4. diverticulitis
Patient information:
Age: {age}
Sex: {sex}
Chief complaint: {chief complaint}
Return exactly one diagnosis from the allowed list and
nothing else.
```

The values in braces denote patient-specific fields inserted into the frozen template.

For Llama inference, this user message is passed through the model-specific tokenizer chat template with `add_generation_prompt=True`. Generation is deterministic with sampling disabled in the primary direct-diagnosis inference.

H.2 MIMIC ASK-UPDATED DIAGNOSIS PROMPT

ASK preserves the base model and diagnostic instruction while adding the selected admissible evidence to the observable state.

The frozen prompt has the form:

```
You are given an emergency-department presentation
from a patient who was ultimately diagnosed with one
of four conditions.
Choose the single most likely diagnosis using only the
information provided.
Allowed diagnoses:
1. appendicitis
2. cholecystitis
3. pancreatitis
4. diverticulitis
Patient information:
Age: {age}
Sex: {sex}
Chief complaint: {chief complaint}
Additional evidence:
{selected evidence}
Return exactly one diagnosis from the allowed list and
nothing else.
```

The VITALS, LABS, and IMAGING channels use the same prompt structure. They differ only in the contents inserted into the `Additional evidence` block.

For example, the VITALS block contains the admissible vital-sign observations. The LABS block contains the admissible laboratory measurements together with units and abnormality indicators when present. The IMAGING block contains the admissible imaging report text.

Only evidence satisfying the frozen temporal-admissibility rule is eligible for insertion.

H.3 PROMPT MATCHING UNDER DEFER

DEFER does not modify the observable patient state or the diagnostic prompt. The stronger model receives the same frozen prompt used to evaluate the base model at that state.

For the primary experiment, Llama-3.1-8B-Instruct and Llama-3.1-70B-Instruct therefore receive the same model-visible clinical information. Only the decision maker changes.

This implements the intended distinction

ASK: change available evidence,

whereas

DEFER: change the decision maker.

The Qwen3 robustness experiment follows the same direct-diagnosis protocol. Qwen3 inference uses a single user message and explicitly disables reasoning mode with

`enable_thinking=False`,

so that the evaluation remains aligned with the original direct-diagnosis setup.

H.4 Sequential DDXPlus Provenance

The sequential DDXPlus experiment uses protocol

`STEP17E-DDXPLUS-SEQUENTIAL-ASK-V1`

with Qwen2.5-14B-Instruct.

Each sequential trajectory contains the initial model prediction and the predictions after each acquired question. The trajectory artifact records, for every inference state,

- the model identifier,
- the frozen model revision,
- the observable-state hash,
- the prompt hash,
- the normalized diagnosis,
- the raw model response,
- parser status and parser method,
- input and output token counts, and
- the selected question and updated state hash.

The executed literal prompt string is not stored verbatim in the trajectory artifact. Instead, the corresponding prompt hash is stored for each inference state. We therefore do not reconstruct a literal sequential prompt after the fact.

This distinction is important for reproducibility: the manuscript reports the frozen observable-state protocol and the recorded prompt provenance rather than presenting a manually reconstructed string as if it were the executed prompt.

H.5 Sequential Output Format

The Step17E inference records use structured diagnostic responses. Successful examples take the form

```
{"diagnosis": "Bronchitis"}
```

with the diagnosis value varying by case.

The recorded parser method for successful Step17E trajectory predictions is `json_exact`. Invalid outputs are retained and handled according to the robustness analysis in Appendix G rather than silently removed.

H.6 State Updates and Stopping Features

After a sequential ASK action acquires evidence e_t, the observable state changes according to

$$s_{t+1} = s_t \cup \{e_t\}.$$

The model is then evaluated at the updated state.

The learned stopping model in Appendix E.2 does not receive gold labels, realized diagnostic errors, future outcomes, or oracle repairability information as features. Its inference-time features are restricted to the current model prediction, parse status, acquisition policy, step index, number of remaining legal questions, and identifiers of previously asked questions.

The continuation target is defined using future prefixes only for supervision:

$$\text{CONTINUE} = 1$$

if a later prefix on the same realized five-question trajectory attains lower diagnostic error than stopping at the current state.

Thus, the sequential stopping analysis is behavior-conditioned on the observed acquisition trajectory. It is not an oracle search over all possible future ASK sequences.