EVIDIAG: POSTERIOR-GUIDED EVIDENCE SEEKING FOR AUDITABLE TIME-SERIES DIAGNOSIS

Anonymous authors
Paper under double-blind review

ABSTRACT

Time-series diagnosis commonly maps observed signals directly to a fault or root-cause label, leaving evidence acquisition and decision making implicit. Although large language models (LLMs) have enabled readable diagnostic reasoning, existing approaches typically lack explicit mechanisms for sequentially planning evidence acquisition and quantitatively revising fault beliefs. This challenge is compounded in multivariate time series, where fault evidence is often buried in context-dependent temporal dynamics rather than directly available in text or logs. We present EviDiag, a posterior-guided evidence-seeking agent that turns diagnosis into an auditable process of examination and belief revision. EviDiag recovers a context-conditioned healthy trajectory from a pre-fault anchor to ground subsequent examinations, and organizes diagnosis into a within-case inquiry loop and an across-case evolution loop. The within-case inquiry loop uses the fault posterior and finite-horizon information value to guide an LLM planner in selecting the next examination, whose outcome revises the fault belief via Bayesian updating. The across-case loop distills resolved state-action-outcome transitions into retrievable experience and recalibrates evidence likelihoods to improve future inquiry. Across microservice and industrial-process benchmarks, EviDiag attains 1.06× and 1.90× the top-1 diagnostic accuracy of the best baseline, respectively, while producing auditable evidence-to-posterior traces.

1 INTRODUCTION

From computing infrastructures to electrochemical assets, fault diagnosis from multivariate time series is essential for reliable and safe operation of complex dynamical systems. Existing methods learn temporal regularities through forecasting, reconstruction, representation learning, or generative modeling to detect anomalies and trigger alarms (Zhang et al., 2026; Chen & Zhao, 2024). Beyond anomaly detection, fault isolation and root-cause analysis (RCA) identify fault types, affected components, or plausible sources (Han et al., 2025; Nagalapatti et al., 2025). Yet most conventional time-series diagnostic pipelines infer a terminal score, label, or attribution from a pre-collected window, without explicitly modeling evidence acquisition or evaluation.

Large language models (LLMs) extend diagnosis beyond numerical outputs by interpreting and organizing heterogeneous observations in natural language(Yue et al., 2026). Domain-adapted and multi-modal models connect engineering observations with fault semantics, while LLM-based RCA systems aggregate telemetry, logs, traces, and historical incidents to generate diagnoses or explanations (Lin et al., 2025; Chen et al., 2024; Xu et al., 2025). These approaches improve diagnostic expressiveness and interpretability, but largely reason over an already assembled evidence set rather than actively determining what evidence to acquire.

Recent agentic reasoning methods begin to move beyond this fixed-evidence setting. Chain-of-thought (CoT) (Kojima et al., 2022) decomposes diagnostic inference into intermediate steps, while ReAct-style and tool-using agents (Yao et al., 2022) actively query telemetry, invoke analytical tools, or request additional observations. Interactive clinical agents (Schmidgall et al., 2026; Nori et al., 2025) further demonstrate sequential inquiry by requesting missing symptoms, examinations, or test results before diagnosis. These developments make diagnosis increasingly interactive, but interaction alone does not make the process explicitly evidence-guided. Even when tools return objective measurements, an action's diagnostic value, e.g., why to select it, what its possible outcomes imply

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.

for competing hypotheses, and how the observed outcome should revise the diagnosis, is often left to language-mediated reasoning rather than an explicitly quantified representation.

Explicit evidence seeking in time-series diagnosis raises two further challenges. First, fault evidence is often embedded in context-dependent temporal dynamics rather than exposed as an explicit event. For example, in an industrial chemical process, a change in reactor pressure may indicate a fault or reflect a normal shift in operating conditions(Zhao, 2022). Interpreting such changes requires a context-matched healthy reference, yet a reference updated with post-alarm observations may gradually absorb persistent fault dynamics. Second, the usefulness of diagnostic evidence changes as the diagnosis evolves. Evidence that strongly distinguishes the remaining fault hypotheses at one stage may become less informative after some candidates are ruled out. Multi-step diagnosis must therefore adapt what evidence to acquire as the set of plausible faults narrows. These challenges are coupled: reliable evidence seeking requires a trustworthy reference, while the value of new evidence depends on the evolving diagnostic state.

Driven by these issues, we develop EviDiag, a dual-loop evidence-seeking agent that organizes time-series diagnosis around explicit evidence acquisition and belief evolution. EviDiag first recovers a context-matched healthy trajectory from trusted pre-fault observations to construct health-relative diagnostic evidence. Given the current fault posterior, candidate examinations are evaluated for their expected diagnostic value, and an LLM planner selects the next examination. The resulting tool outcome drives an explicit Bayesian update of the posterior, forming an inner evidence-seeking loop in which each diagnostic step is grounded in acquired evidence rather than language reasoning alone. Beyond individual cases, an outer evolution loop retrieves useful examination experience and recalibrates how different evidence supports different faults. EviDiag thereby unifies time-series evidence construction, posterior-guided inquiry, and cross-case evolution in a traceable diagnostic process. Our contributions are threefold:

- **Evidence-seeking diagnostic agent.** We propose EviDiag, an LLM-based diagnostic agent that separates language planning from evidence generation and fault-belief updating. Candidate examinations are explicitly evaluated against the current diagnosis, while tool-grounded outcomes numerically revise the competing fault hypotheses, yielding an auditable alternative to purely language-driven multi-step reasoning.
- **Anchored time-series evidence construction.** We develop Anchored Normative Trajectory Recovery to convert context-dependent temporal changes into health-relative diagnostic evidence. We further provide an error decomposition showing that anchoring avoids the history-induced shift associated with post-alarm rolling updates.
- **Dual-loop diagnostic evolution.** We couple an inner evidence-seeking loop for each diagnosis with an outer evolution loop across resolved cases. The outer loop reuses informative examination experience and recalibrates evidence likelihoods, enabling future inquiry to improve without fine-tuning the LLM planner.

2 Related Work

2.1 Time-Series Anomaly Detection and Reference Modeling

Time-series anomaly detection commonly models normal temporal dynamics, learning regular patterns from multivariate time series and identifying departures from them as anomalies (Zhang et al., 2026; Sankararaman et al., 2022). Yet, these methods are primarily designed to determine whether the observed trajectory departs from learned normal behavior, rather than to maintain a fault-excluding reference for interpreting evidence after an alarm. In particular, continually adapting the reference with post-alarm observations may gradually absorb persistent faulty dynamics. Beyond standard normal-reference modeling, related work has explored alternative system evolutions: DoFlow models observational, interventional, and counterfactual time-series forecasts with explicit likelihoods (Wu et al., 2026), while RecAD uses causal recourse to localize or mitigate anomalous states in static and temporal settings (Han et al., 2023).

However, these alternative trajectories are mainly developed for prediction, localization, or mitigation, rather than as a persistent healthy reference for sequential diagnostic inquiry. EviDiag instead

recovers a context-matched normative trajectory from trusted pre-fault observations and uses it to construct health-relative evidence throughout diagnosis.

2.2 LLM Reasoning and Diagnostic Agents

LLMs have increasingly been used to support diagnostic reasoning and generate natural-language explanations. Early LLM-based approaches mainly reason over provided signals, telemetry, logs, retrieved knowledge, or incident context to produce diagnostic conclusions (Zhang et al., 2024; Lin et al., 2025; Chen et al., 2024; Xu et al., 2025). More recent methods extend this paradigm from reasoning over fixed context to iterative interaction. RCAgent alternates reasoning with telemetry collection and tool execution for cloud root-cause analysis (Wang et al., 2024), while TS-Agent repeatedly applies atomic statistical operators to raw time series and records the resulting insights in an explicit evidence log (Liu et al., 2025). Multi-agent systems further distribute diagnosis across specialized roles, coordinating them through scheduling, hierarchical planning, or context-controlled consensus (Wang et al., 2025; Gao et al., 2026; Kang et al., 2026; Hou et al., 2026). Language agents can also reuse past experience through verbal or episodic memory to improve subsequent decisions (Shinn et al., 2023), while recent RCA benchmarks move beyond terminal predictions to evaluate causal paths and action trajectories (Fang et al., 2026).

Despite this progress, existing LLM agents rarely make explicit how much each examination and its outcome changes the relative support for competing fault hypotheses. EviDiag explicitly represents this diagnostic transition and uses it to guide subsequent inquiry.

2.3 Sequential Diagnosis and Active Evidence Acquisition

Sequential diagnosis has a long history in domains such as clinical diagnosis and engineered-system troubleshooting, where tests or measurements are selected sequentially to discriminate competing hypotheses. Existing approaches use probabilistic, model-based, or learning-based formulations to select informative examinations under uncertain outcomes and adapt the inquiry process (Bellala et al., 2012; Yong & Ozay, 2018; Rodler, 2023; Chen et al., 2022; Hou et al., 2023). More recently, interactive LLM-based diagnosis has brought this paradigm to clinical reasoning, where agents actively seek missing symptoms or tests before diagnosis (Schmidgall et al., 2026; Nori et al., 2025; Long & Wei, 2026). UoT (Hu et al., 2024) further introduces information gain to guide the inquiry process, yet the diagnostic loop remains largely LLM-mediated. EviDiag instead structures inquiry around an explicit Bayesian state, from examination valuation to belief revision.

More broadly, these methods typically seek evidence that can be directly elicited, such as symptoms or test results. EviDiag instead closes the loop from raw multivariate time series to health-relative evidence construction, sequential examination, and posterior evolution, enabling end-to-end evidence-seeking diagnosis over time-series data.

3 EviDiag: Posterior-Guided Evidence Seeking

3.1 Problem Formulation and Diagnostic Motivation

At physical time t, the system provides a multivariate observation $\boldsymbol{y}_t \in \mathbb{R}^D$ together with operating context $\boldsymbol{c}_t \in \mathbb{R}^Q$, such as a setpoint or anticipated load. When anomalies are detected, EviDiag raises an alarm at t_0 and extracts a diagnostic time-series segment around it. The goal is to identify the underlying fault from a finite candidate set $\mathbb{F}$ and produce an auditable diagnosis trace.

Rather than committing to a diagnosis after the alarm, EviDiag refines it over inquiry rounds $r = 1, \ldots, R$. At round r, the agent selects an executable examination a_r from the feasible set $\mathbb{A}_r$ to test a specific temporal pattern, such as a Step rise, Ramp drop, or Oscillation. The corresponding tool returns $o_r \in \{\mathrm{Hit}, \mathrm{Miss}\}$, indicating whether the pattern is supported by the observed time series. Let $b_r(f)$ denote the current belief in fault hypothesis $f \in \mathbb{F}$, with $b_1(f) = \pi(f)$. Each outcome updates this belief, and the final diagnosis is $\hat{f} = \arg\max_{f \in \mathbb{F}} b_{R+1}(f)$.

This sequential diagnosis is hampered by two challenges: constructing a trustworthy, fault-excluding healthy reference when post-alarm observations may already be fault-affected, and selecting informative examinations as diagnostic beliefs, accumulated evidence, and feasible actions evolve. Ev-

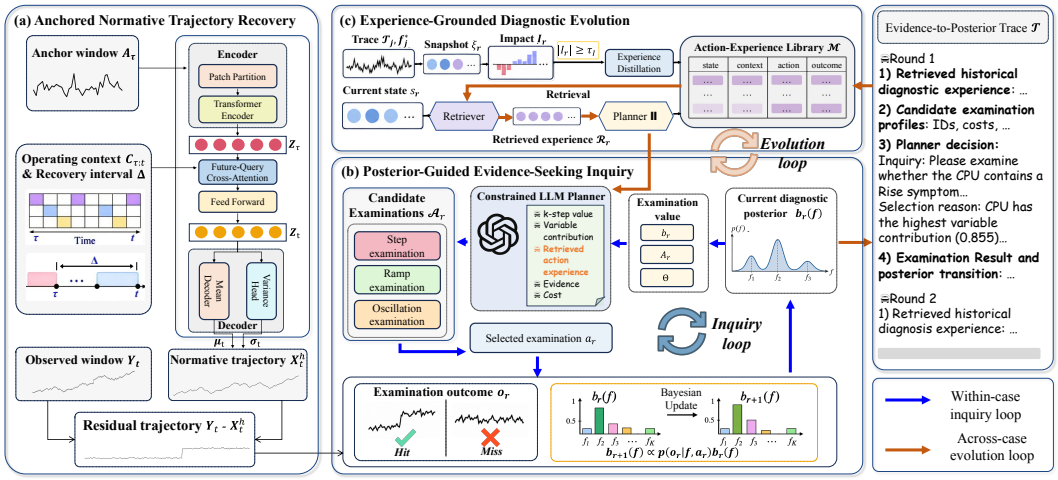


Figure 1: EviDiag couples anchored normative recovery with posterior-guided inquiry and experience-grounded diagnostic evolution. Diagnostic tools return *Hit*/*Miss* outcomes, and numerical updates advance the posterior across inquiry rounds.

iDiag addresses these challenges through anchored normative trajectory recovery and a dual-loop framework that couples posterior-guided inquiry within each case with experience-grounded evolution across cases, whose details are introduced below.

3.2 Framework Overview

Figure 1 illustrates the EviDiag framework. EviDiag constructs fault evidence from the residual between the observed trajectory and an anchored, context-conditioned normative trajectory X_t. At each inquiry round, the LLM planner selects an examination a_r based on the current posterior b_r, examination values and costs, and retrieved experience. The selected tool returns o_r, which updates the posterior to b_{r+1} through the Bayesian evidence model. Across resolved cases, consequential examination transitions are stored and retrieved to guide future inquiry, while confirmed Hit/Miss outcomes periodically recalibrate the evidence likelihoods. Unlike ReAct-style reasoning, the LLM here is confined to examination planning, while diagnostic tools and the Bayesian model determine examination outcomes and posterior updates, keeping the process traceable and auditable.

3.3 Anchored Normative Trajectory Recovery

Health-relative diagnostic evidence requires a trustworthy estimate of how the system would have evolved under normal operation. Because post-alarm observations may already be fault-affected, EviDiag recovers each target window from the latest admissible healthy anchor rather than conditioning on the measurements under diagnosis. Let $\boldsymbol{A}_\tau$ be an anchor window ending at $\tau < t_0$, $t \geq t_0$ the start of a target window, and $\Delta = t - \tau$ the recovery interval. Let $\boldsymbol{C}_{\tau:t}$ denote the operating context from the anchor through the target window. The healthy trajectory $\boldsymbol{X}_t^{\mathrm{h}}$ can be predicted as

$$\boldsymbol{X}_t^{\mathrm{h}} \sim p_\Theta\big(\boldsymbol{X}_t^{\mathrm{h}} \mid \boldsymbol{A}_\tau, \boldsymbol{C}_{\tau:t}, \Delta\big), \qquad p_\Theta = \mathcal{N}\big(\boldsymbol{\mu}_t, \mathrm{diag}(\boldsymbol{\sigma}_t^2)\big). \tag{1}$$

The model is trained on healthy anchor–target pairs with randomly sampled recovery intervals, allowing one gap-conditioned model to cover the supported horizons. At deployment, $\widehat{\boldsymbol{X}}_t = \boldsymbol{\mu}_t$ serves as the normative reference, while $\boldsymbol{\sigma}_t^2$ parameterizes predictive uncertainty. Before an alarm, the admissible anchor advances behind a safety interval; once confirmed, the alarm freezes the anchor so that all subsequent references use only trusted pre-alarm history and available context. Architecture, training, pair construction, and the online protocol are detailed in Appendix A.

Anchoring keeps the diagnostic contrast tied to healthy-trajectory recovery error, whereas a rolling reference is additionally affected by post-alarm history. Proposition 1 formalizes this distinction. Define the fault-induced deviation $\boldsymbol{D}_t := \boldsymbol{Y}_t - \boldsymbol{X}_t^{\mathrm{h}}$ and ideal diagnostic contrast $d_t := |\boldsymbol{D}_t| > 0$.

For the anchored reference, let

$$e_t^{\mathrm{A}} := \|\widehat{\boldsymbol{X}}_t - \boldsymbol{X}_t^{\mathrm{h}}\|, \qquad r_t^{\mathrm{A}} := \|\boldsymbol{Y}_t - \widehat{\boldsymbol{X}}_t\|.$$

Proposition 1 (Diagnostic contrast under anchored and rolling references)**.** *For any anchored recovery $\widehat{\boldsymbol{X}}_t$ satisfying $\|\widehat{\boldsymbol{X}}_t - \boldsymbol{X}_t^{\mathrm{h}}\| \le \epsilon_t$,*

$$e_t^{\mathrm{A}} \le \epsilon_t, \qquad \left|r_t^{\mathrm{A}} - d_t\right| \le \epsilon_t. \tag{2}$$

For comparison, consider the same rolling predictor under matched healthy and post-alarm histories, with outputs $\widetilde{\boldsymbol{X}}_t^0$ and $\widetilde{\boldsymbol{X}}_t^\delta$. Define

$$\boldsymbol{U}_t = \widetilde{\boldsymbol{X}}_t^0 - \boldsymbol{X}_t^{\mathrm{h}}, \qquad \boldsymbol{P}_t = \widetilde{\boldsymbol{X}}_t^\delta - \widetilde{\boldsymbol{X}}_t^0, \qquad \eta_t = \|\boldsymbol{U}_t\|,$$

where $\boldsymbol{P}_t$ is the output shift induced by post-alarm history. With

$$e_t^{\mathrm{R}} := \|\widetilde{\boldsymbol{X}}_t^\delta - \boldsymbol{X}_t^{\mathrm{h}}\|, \qquad r_t^{\mathrm{R}} := \|\boldsymbol{Y}_t - \widetilde{\boldsymbol{X}}_t^\delta\|,$$

the rolling quantities satisfy

$$\left|e_t^{\mathrm{R}} - \|\boldsymbol{P}_t\|\right| \le \eta_t, \qquad \left|r_t^{\mathrm{R}} - \|\boldsymbol{D}_t - \boldsymbol{P}_t\|\right| \le \eta_t. \tag{3}$$

Thus, the anchored residual differs from the ideal diagnostic contrast by at most ϵ_t, whereas the rolling residual additionally depends on the history-induced shift $\boldsymbol{P}_t$ and reflects $\boldsymbol{D}_t - \boldsymbol{P}_t$. When $\boldsymbol{P}_t$ points toward $\boldsymbol{D}_t$, the rolling reference can absorb part of the fault contrast, although the bound itself requires no such directional alignment. The proof and further analysis are provided in Appendix A.4.

Having established a fault-excluding normative reference, EviDiag uses the observation-to-reference residual as the basis for executable evidence acquisition. We design a set of temporal examination operators to identify fault-related patterns in the residual, including level shifts, steps, ramps, and oscillations. Each examination returns a discrete *Hit* or *Miss* outcome, which is subsequently used by the posterior-guided inquiry module. Appendix B provides the examination statistics, threshold calibration, temporal aggregation, and the corresponding robustness analysis.

3.4 Posterior-Guided Evidence-Seeking Inquiry

Using the health-relative examinations above, Posterior-Guided Evidence-Seeking Inquiry performs diagnosis through multi-round evidence acquisition. Each examination a returns an outcome $o \in \{\mathrm{Hit}, \mathrm{Miss}\}$ from the deviation relative to the anchored normative trajectory. To quantify how this outcome supports competing fault hypotheses, let $\theta_{f,a} = P(\mathrm{Hit} \mid f, a)$ and $P(\mathrm{Miss} \mid f, a) = 1 - \theta_{f,a}$. These likelihoods are initialized from expert-supported fault–examination relations and modeled with a naive-Bayes factorization over nonrepeated examinations (Appendix C). Once o_r is observed, the posterior is updated as

$$b_{r+1}(f) = \mathcal{B}(b_r, a_r, o_r)(f) \propto b_r(f) P(o_r \mid f, a_r). \tag{4}$$

At each inquiry round, EviDiag first removes previously executed examinations and evaluates the remaining candidates from the current posterior. For $\mathcal{O}_a^{\mathrm{plan}} = \{\mathrm{Hit}, \mathrm{Miss}\}$ and $p(o \mid b, a) = \sum_f b(f) P(o \mid f, a)$, the expected one-step information gain is

$$\mathrm{IG}(b, a) = H(b) - \sum_{o \in \mathcal{O}_a^{\mathrm{plan}}} p(o \mid b, a) H(\mathcal{B}(b, a, o)), \tag{5}$$

where $H(b) = -\sum_f b(f) \log b(f)$. Different outcomes can make different follow-up examinations informative, we further evaluate their outcome-dependent continuation. With $\mathbb{A}_{-a} = \mathbb{A} \setminus \{a\}$,

$$\begin{aligned} Q_k(b, \mathbb{A}, a) &= \mathrm{IG}(b, a) + \gamma \sum_{o \in \mathcal{O}_a^{\mathrm{plan}}} p(o \mid b, a) V_{k-1}(\mathcal{B}(b, a, o), \mathbb{A}_{-a}), \\ V_k(b, \mathbb{A}) &= \max_{a \in \mathbb{A}} Q_k(b, \mathbb{A}, a), \qquad V_0(b, \mathbb{A}) = 0, \qquad V_k(b, \varnothing) = 0. \end{aligned} \tag{6}$$

Thus, Q_k combines immediate information gain with outcome-dependent continuation value, yielding the expected reduction in terminal posterior entropy under our planning setting.

Importantly, Q_K is an information-value input rather than the complete action-selection rule. The LLM planner receives the current posterior, observed outcomes, candidate Q_K values, examination costs and variable contributions to select one examination from the feasible set. Unlike free-form tool selection, the LLM does not predict the examination outcome or modify the evidence likelihoods or posterior. The selected tool produces the actual outcome, Eq. 4 updates the posterior, and the next round replans from the updated diagnostic state. Appendix D gives the lookahead derivation, and Appendix C.3 specifies the planner interface.

3.5 Experience-Grounded Diagnostic Evolution

Experience-Grounded Diagnostic Evolution updates two forms of historical knowledge after a case is independently resolved: diagnostically consequential transitions are immediately added to an action-experience library $\mathcal{M}$ for retrieval in subsequent cases, whereas evidence likelihoods are recalibrated only after sufficient resolved cases have accumulated for stable estimation. At round r, define the diagnostic state $s_r = (b_r, \mathcal{H}_r, \mathbb{A}_r, \rho_r)$, where $\mathcal{H}_r$ is the executed action–outcome history, $\mathbb{A}_r$ is the current feasible examination set, and ρ_r denotes the health-relative variable contributions. Each inquiry round produces a transition (s_r, a_r, o_r, s_{r+1}). Once the case is independently resolved, the expert-confirmed fault label $f^\star$ is attached to form the snapshot $\xi_r = (s_r, a_r, o_r, s_{r+1}; f^\star)$.

The snapshot impact measures how the examination changes the posterior margin of the confirmed fault over its strongest competitor:

$$I_r = \left[b_{r+1}(f^\star) - \max_{f \in \mathbb{F},\, f \neq f^\star} b_{r+1}(f) \right] - \left[b_r(f^\star) - \max_{f \in \mathbb{F},\, f \neq f^\star} b_r(f) \right]. \tag{7}$$

Thus, I_r records whether and by how much the examination changes the separation between the confirmed fault and its strongest alternative. Snapshots satisfying $|I_r| \geq \tau_I$ are retained as state-indexed action experience, preserving diagnostically consequential transitions in either direction. An LLM distills each retained transition into a compact lesson, while its numerical state, tool outcome, posterior change, and provenance remain attached. For a subsequent case, retrieved lessons can guide examination selection but cannot generate a current outcome or directly update the posterior.

Separately, after a case is resolved, all eligible examinations are evaluated from its recorded health-relative residuals. Their *Hit*/*Miss* outcomes accumulate across historical cases and periodically recalibrate the fault-conditioned evidence likelihoods through a Beta–Bernoulli update (Appendix C).

Algorithm 1 gives the complete chronology of within-case inquiry and subsequent historical updates. Historical information can affect future action selection and likelihood estimation, but it neither supplies a current tool outcome nor directly overwrites b_r. Appendix C specifies likelihood initialization, eligibility, experience retrieval, and the experimental history-pool protocol.

Algorithm 1 Within-case inquiry and across-case evolution

Require: likelihood table Θ, action-experience library $\mathcal{M}$, calibration buffer $\mathcal{C}$, impact threshold τ_I, batch interval B
Ensure: diagnostic trace $\mathcal{T}_j$; updated $\mathcal{M}$ and periodically updated Θ

for each case j in chronological order **do**
  Initialize N, b_1, history $\mathcal{H}_1$, and feasible set $\mathbb{A}_1$.
  while the diagnostic stopping rule is unmet **do**
    Score feasible actions by Q_K and retrieve $\mathcal{R}_r$ from $\mathcal{M}$.
    Select a_r with the LLM planner and execute it to obtain o_r.
    Update $b_{r+1} = \mathcal{B}(b_r, a_r, o_r)$; append the transition to $\mathcal{T}_j$.
  end while; obtain the independently resolved target $f_j^\star$.
  Compute each I_r; add snapshots with $|I_r| \geq \tau_I$ to $\mathcal{M}$.
  Retrospectively evaluate all eligible examinations and append their outcomes to $\mathcal{C}$.
  if B cases accumulated **then** update cumulative counts and calibrate Θ by Eq. 39; clear $\mathcal{C}$.
end for

4 Experiments

We evaluate EviDiag on two benchmarks spanning microservice systems and industrial processes. The experiments assess end-to-end diagnostic performance, followed by ablations and mechanism analyses of anchoring, posterior guidance, and historical experience.

Datasets and metrics. We use RCAEval (Pham et al., 2025), a microservice root-cause diagnosis benchmark, and the Extended Tennessee Eastman Process (E-TEP) (Reinartz et al., 2021), an industrial-process benchmark. For RCAEval, all methods are evaluated on the same 60-case

Table 1: End-to-end diagnostic ranking results (%). All metrics are higher-is-better.

Family	Method	RCAEval					E-TEP				
		R@1	R@3	R@5	MRR	F1	R@1	R@3	R@5	MRR	F1
Probabilistic	Passive NB	50.00	**86.67**	95.00	65.28	52.29	22.00	60.00	74.00	39.00	13.56
	Active NB-IG	33.33	78.33	93.33	51.94	29.51	20.00	60.00	72.00	37.00	16.85
LLM-based	RAG-based	53.33	**86.67**	**96.67**	67.50	53.87	30.00	38.00	56.00	33.33	26.11
	CoT	55.00	**86.67**	93.33	68.61	54.71	16.00	30.00	50.00	21.33	14.04
	ReAct	33.33	65.00	90.00	46.94	27.34	18.00	40.00	72.00	27.00	8.64
	FaultExplainer	23.33	40.00	41.67	31.11	23.11	40.00	70.00	82.00	51.67	37.06
	OpenRCA	45.00	81.67	**96.67**	60.83	44.79	40.00	70.00	80.00	52.33	39.23
Ours	**EviDiag**	**58.33**	81.67	88.33	**70.00**	**60.40**	**76.00**	**90.00**	**92.00**	**83.00**	**75.58**

R@k: Recall@k; MRR: MRR@3; F1: Macro-F1. Bold indicates the best result in each column.

Online Boutique (OB)/Sock Shop (SS) test set, while E-TEP uses 50 held-out test runs. We evaluate diagnostic ranking using Recall@k ($k \in \{1, 3, 5\}$) and MRR@3, together with Macro-F1 for class-balanced fault identification. All metrics are computed end-to-end: missed alarms and absent diagnostic outputs receive zero credit. Dataset construction, preprocessing, and metric definitions are provided in Appendix E.3.

Baselines. We compare against two baseline families: traditional diagnosis using passive and information-gain-guided Naive Bayes (Dey & Stori, 2005; Zheng et al., 2012); and five LLM-based approaches, namely FaultExplainer (Khan et al., 2025), retrieval-augmented generation (RAG)-based diagnosis (Lewis et al., 2020), zero-shot chain-of-thought prompting (CoT) (Kojima et al., 2022), ReAct (Yao et al., 2022), and OpenRCA (Xu et al., 2025). Detailed configurations and implementation sources are provided in Appendix E.4.

4.1 Implementation Details

Unless otherwise stated, EviDiag uses DeepSeek-V4.1-Flash and two-step look-ahead ($K = 2$, $\gamma = 1$). Additional cross-backbone evaluations are reported in Appendix E.6.2 to assess robustness to the choice of LLM planner. Action experience and evidence likelihoods are constructed from a fixed validation pool and frozen during held-out testing. Dataset-specific settings and remaining implementation details are provided in Appendix A, Appendix C, and Appendix E.

4.2 Diagnostic Results

Table 1 summarizes the end-to-end diagnostic results on both benchmarks. On RCAEval, EviDiag achieves the best R@1 (58.33%), MRR@3 (70.00%), and Macro-F1 (60.40%). Although RAG and OpenRCA attain higher recall at broader cutoffs, EviDiag more often ranks the correct fault first, suggesting an advantage in diagnosis prioritization. The performance margin widens on E-TEP. EviDiag reaches 76.00% R@1 and 75.58% Macro-F1, while maintaining above 90% recall at ranks 3 and 5. The best baseline reaches 40.00% R@1 and 39.23% Macro-F1. Paired significance analyses are further provided in Appendix E.6.1. These results indicate that EviDiag maintains its diagnostic advantage across both microservice and industrial-process settings.

4.3 Ablation Studies

Table 2 evaluates major EviDiag components under the same test cases and inquiry budgets as Table 1. Removing anchored recovery reduces R@1 by 15.00 points on RCAEval and 18.00 on E-TEP, supporting the value of a fault-excluding normative reference, while removing historical experience reduces E-TEP R@1 from 76.00% to 40.00%. Removing LLM inquiry or information-gain guidance also degrades top-1 performance. On RCAEval, the anchored-recovery, historical-experience, and LLM-inquiry ablations all yield 43.33% R@1 (26/60 correct), but differ in MRR, Macro-F1, and higher-rank recall, indicating distinct case-level ranking behavior.

Table 2: Component ablation results (%). Test cases and inquiry budgets follow Table 1.

	RCAEval					E-TEP				
Variant	R@1	R@3	R@5	MRR	F1	R@1	R@3	R@5	MRR	F1
Full EviDiag	**58.33**	**81.67**	88.33	**70.00**	**60.40**	**76.00**	**90.00**	**92.00**	**83.00**	**75.58**
w/o Anchored Recovery	43.33	80.00	91.67	60.56	46.28	58.00	88.00	90.00	71.33	56.56
w/o Historical Experience	43.33	80.00	**93.33**	60.00	42.93	40.00	64.00	70.00	50.67	27.69
w/o LLM Inquiry (Q_K)	43.33	**81.67**	91.67	61.11	43.96	38.00	66.00	72.00	51.33	28.77
w/o Information-Gain Values	40.00	71.67	91.67	54.44	40.27	68.00	86.00	**92.00**	75.67	63.71

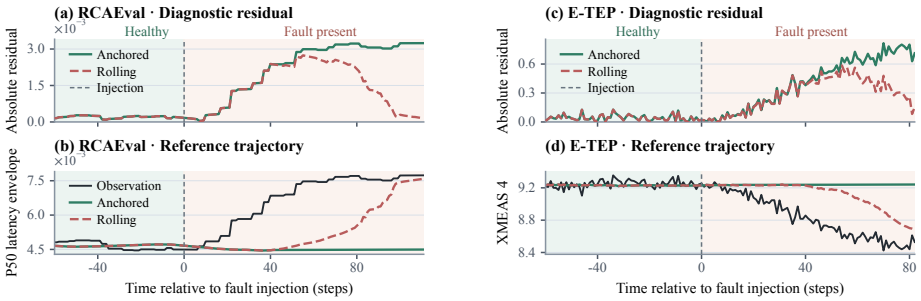


Figure 2: Reference tracking in two illustrative cases: RCAEval `RCAEvalss_user_cpu_3` ((a) residuals; (b) P50 latency trajectories) and E-TEP `idv01_run4088546448` ((c) residuals; (d) `XMEAS_4` trajectories). Shading marks healthy/fault segments; dashed lines indicate fault injection.

4.4 Mechanism Analysis

4.4.1 Anchoring and Reference Contamination

Figure 2 compares anchored and rolling recovery in illustrative RCAEval and E-TEP cases. Within each dataset, both use the same predictor architecture. After the alarm, the rolling reference continues updating from post-alarm observations and progressively approaches the faulty trajectory, whereas anchored recovery propagates from the last admissible pre-alarm state. In RCAEval, the rolling P50-latency residual decreases to 3.73×10^{-4} over the final 20 steps, compared with 3.17×10^{-3} under anchoring. E-TEP shows the same qualitative behavior on XMEAS 4: rolling recovery follows the post-fault response, while anchoring preserves observation-to-reference separation. These examples are consistent with the mechanism described in our Proposition 1.

4.4.2 Posterior Guidance across Inquiry Rounds

We compare EviDiag with sequential tool calls, max-Q_K selection without an LLM, and an LLM policy without information-gain or Q_K values. All policies start from the same post-initialization posterior and use the same calibrated evidence model, and the LLM-based variants also share the same historical experience. We track fault-classification accuracy over 14 inquiry rounds on the common 60-case RCAEval test set to examine the selection behavior of involved configurations.

Figure 3 shows that EviDiag improves more rapidly during the early inquiry rounds than sequential tool calls and the LLM without information-value guidance, reaching higher R@1 with fewer examinations. It also outperforms pure max-Q_K in the early rounds, suggesting that the LLM can flexibly combine numerical planning values with diagnostic context for more effective inquiry. The policies become comparable at later rounds, indicating that the main benefit lies in faster diagnostic refinement under limited inquiry budgets.

4.4.3 Historical Diagnostic Evidence

We examine how accumulated historical experience affects future inquiry while fixing the LLM planner, case order, and five-round budget. Using nested fractions from 0% to 100% of the 36-case validation pool, both the action-experience library and evidence-likelihood ta-

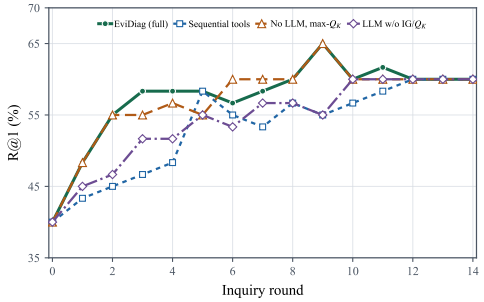


Figure 3. R@1 metrics after initialization (round 0) and each inquiry round.

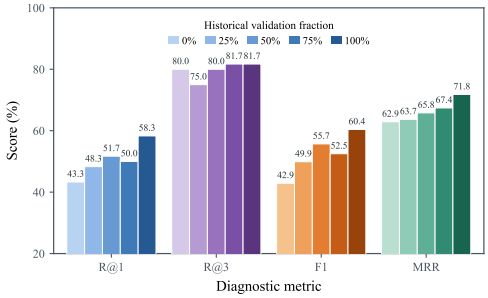


Figure 4. Effect of historical validation-pool size on diagnostic performance.

ble are rebuilt from each subset before evaluation on the same 60 RCAEval test cases. Figure 4 shows that the full historical pool achieves the best result on all four metrics. In particular, R@1 increases from 43.33% without historical evidence to 58.33% with the full pool, with performance showing an overall upward trend as historical cases increase. This sweep mimics practical case accumulation, where state-matched action experience and likelihood calibration progressively improve EviDiag's diagnostic performance.

4.5 Auditable Diagnostic Traces

Figure 5 illustrates how EviDiag exposes the diagnostic process rather than only its final prediction. The upper panel compares observed trajectories with their anchored healthy references, while the inquiry trace links each LLM-selected examination to its tool outcome and the resulting Bayesian posterior update. In the illustrated RCAEval case, successive *Miss* and *Hit* outcomes suppress competing hypotheses and raise the posterior of Socket to 97.9%. Because every belief change is tied to a specific examination and tool-grounded outcome, the trace makes each diagnostic step directly inspectable and allows questionable decisions or belief updates to be localized to the corresponding evidence. Each inquiry round further records the fault posterior, examination values, the LLM decision and rationale, etc., providing a complete view of the diagnostic process. Complete RCAEval traces and an additional E-TEP case are provided in Appendix E.6.

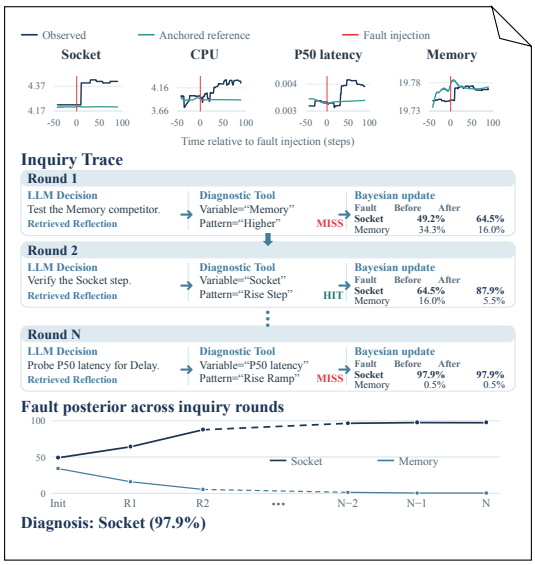


Figure 5: Example of an auditable diagnostic trace generated by EviDiag.

5 Conclusion

Reliable diagnosis of dynamical systems requires a trustworthy reference, explicit evidence acquisition, and traceable belief revision beyond one-shot fault prediction. EviDiag addresses these requirements through anchored normative recovery, posterior-guided inquiry, and experience-grounded diagnostic evolution. Anchoring preserves diagnostic contrast against post-alarm reference drift, while tool-grounded outcomes drive explicit Bayesian belief updates. Across RCAEval and E-TEP, EviDiag improves top-1 fault identification while producing auditable evidence-to-posterior traces. More broadly, EviDiag suggests an evidence-grounded agent paradigm for sequential decision-making under uncertainty, where LLM planning is separated from verifiable tool execution and explicit state updating, making each decision step directly auditable.

AI USE STATEMENT

Generative AI tools were used only in supporting roles for language polishing, literature discovery, and verification of mathematical derivations. For writing, the authors independently prepared the manuscript content, while AI tools were used to improve grammar, clarity, and academic expression rather than to draft substantive sections. For literature discovery, the authors conducted the literature search, screening, and selection independently; AI tools were used only as a supplementary check for potentially missed or recent relevant work. For the theoretical analysis, the motivation, formulation, and derivations were developed by the authors, while AI tools provided supplementary mathematical background and assisted in checking the consistency and validity of the derivations. All AI-assisted mathematical content was independently verified and re-derived by the authors before inclusion.

The research questions, core ideas, methodology, experimental design, result interpretation, and conclusions were independently developed by the authors. The authors take full responsibility for the accuracy, originality, and final content of the manuscript.

REFERENCES

Gowtham Bellala, Jason Stanley, Clayton Scott, and Suresh K Bhavnani. Active diagnosis via auc maximization: An efficient approach for multiple fault identification in large scale, noisy networks. *arXiv preprint arXiv:1202.3701*, 2012.

Jiawei Chen and Chunhui Zhao. Addressing spatial-temporal heterogeneity: General mixed time series analysis via latent continuity recovery and alignment. *Advances in Neural Information Processing Systems*, 37:17910–17946, 2024.

Junying Chen, Dongfang Li, Qingcai Chen, Wenxiu Zhou, and Xin Liu. Diaformer: Automatic diagnosis via symptoms sequence generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 4432–4440, 2022.

Yinfang Chen, Huaibing Xie, Minghua Ma, Yu Kang, Xin Gao, Liu Shi, Yunjie Cao, Xuedong Gao, Hao Fan, Ming Wen, et al. Automatic root cause analysis via large language models for cloud incidents. In *Proceedings of the Nineteenth European Conference on Computer Systems*, pp. 674–688, 2024.

Story Dey and JA Stori. A bayesian network approach to root cause diagnosis of process variations. *International Journal of Machine Tools and Manufacture*, 45(1):75–91, 2005.

Aoyang Fang, Yifan Yang, Jin'ao Shang, Qisheng Lu, Junjielung Xu, Rui Wang, Songhan Zhang, Yuzhong Zhang, Boxi Yu, and Pinjia He. Openrca 2.0: From outcome labels to causal process supervision. *arXiv preprint arXiv:2606.27154*, 2026.

Yufei Gao, Zhengong Cai, and Bowei Yang. Rcaflow: A workflow-informed hierarchical planning multi-agent system for root cause analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 300–308, 2026.

Xiao Han, Lu Zhang, Yongkai Wu, and Shuhan Yuan. Algorithmic recourse in abnormal multivariate time series. *arXiv preprint arXiv:2309.16896*, 2023.

Xiao Han, Saima Absar, Lu Zhang, and Shuhan Yuan. Root cause analysis of anomalies in multivariate time series through granger causal discovery. In *International Conference on Learning Representations*, volume 2025, pp. 45137–45153, 2025.

Difei Hou, Jiaqi Yue, and Chunhui Zhao. Words towards explainability: Caption label-free learning via dual loop agentic time series captioning. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=NezhkAYLK8.

Zhenyu Hou, Yukuo Cen, Ziding Liu, Dongxue Wu, Baoyan Wang, Xuanhe Li, Lei Hong, and Jie Tang. Mtdiag: an effective multi-task framework for automatic diagnosis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 14241–14248, 2023.

Zhiyuan Hu, Chumin Liu, Xidong Feng, Yilun Zhao, See-Kiong Ng, Anh T Luu, Junxian He, Pang W Koh, and Bryan Hooi. Uncertainty of thoughts: Uncertainty-aware planning enhances information seeking in llms. *Advances in Neural Information Processing Systems*, 37:24181–24215, 2024.

Junchi Kang, Xunhui Zhang, Yuanzhao Zhai, Dawei Feng, Bo Ding, and Huaimin Wang. Context-rca: A context-controlled multi-agent framework for root cause analysis in microservice systems. In *2026 IEEE International Conference on Joint Cloud Computing (JCC)*, pp. 327–332. IEEE, 2026.

Abdullah Khan, Rahul Nahar, Hao Chen, Gonzalo E Constante Flores, and Can Li. Faultexplainer: Leveraging large language models for interpretable fault detection and diagnosis. *Computers & Chemical Engineering*, 199:109152, 2025.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213, 2022.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33: 9459–9474, 2020.

Lin Lin, Sihao Zhang, Song Fu, and Yikun Liu. Fd-llm: Large language model for fault diagnosis of complex equipment. *Advanced Engineering Informatics*, 65:103208, 2025.

Penghang Liu, Elizabeth Fons, Annita Vapsi, Mohsen Ghassemi, Svitlana Vyetrenko, Daniel Borrajo, Vamsi K Potluru, and Manuela Veloso. Ts-agent: Understanding and reasoning over raw time series via iterative insight gathering. *arXiv preprint arXiv:2510.07432*, 2025.

Zhuohan Long and Zhongyu Wei. Strong reasoning isn't enough: Evaluating evidence elicitation in interactive diagnosis. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 27559–27575, 2026.

Lokesh Nagalapatti, Ashutosh Srivastava, Sunita Sarawagi, and Amit Sharma. Robust root cause diagnosis using in-distribution interventions. In *International Conference on Learning Representations*, volume 2025, pp. 25738–25765, 2025.

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

Luan Pham, Hongyu Zhang, Huong Ha, Flora Salim, and Xiuzhen Zhang. Rcaeval: A benchmark for root cause analysis of microservice systems with telemetry data. In *Companion Proceedings of the ACM on Web Conference 2025*, pp. 777–780, 2025.

Christopher Reinartz, Murat Kulahci, and Ole Ravn. An extended tennessee eastman simulation dataset for fault-detection and decision support systems. *Computers & chemical engineering*, 149:107281, 2021.

Patrick Rodler. Sequential model-based diagnosis by systematic search. *Artificial Intelligence*, 323: 103988, 2023.

Abishek Sankararaman, Balakrishnan Narayanaswamy, Vikramank Y Singh, and Zhao Song. Fitness:(fine tune on new and similar samples) to detect anomalies in streams with drift and outliers. In *International Conference on Machine Learning*, pp. 19153–19177. PMLR, 2022.

Samuel Schmidgall, Rojin Ziaei, Carl Harris, Ji Woong Kim, Eduardo Pontes Reis, Jeffrey Jopling, and Michael Moor. Agentclinic: a multimodal benchmark for tool-using clinical ai agents. *npj Digital Medicine*, 2026.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. *Advances in neural information processing systems*, 36:8634–8652, 2023.

Yixin Wang, Yingxin Su, Bingli Zhang, Zishan Bai, Guozhong Zhang, Junzhao Jiang, Xinyu Wang, Zhang Chengbiao, Yifan Wang, Xiang Luo, et al. Multi-agent collaborative framework for intelligent it operations: An aoi system with context-aware compression and dynamic task scheduling. In *ICLR 2026 Workshop on Memory for LLM-Based Agentic Systems*, 2025.

Zefan Wang, Zichuan Liu, Yingying Zhang, Aoxiao Zhong, Jihong Wang, Fengbin Yin, Lunting Fan, Lingfei Wu, and Qingsong Wen. Rcagent: Cloud root cause analysis by autonomous agents with tool-augmented large language models. In *Proceedings of the 33rd ACM international conference on information and knowledge management*, pp. 4966–4974, 2024.

Dongze Wu, Feng Qiu, and Yao Xie. Doflow: Flow-based generative models for interventional and counterfactual forecasting on time series. In *International Conference on Learning Representations*, volume 2026, pp. 149949–149980, 2026.

Junjielong Xu, Qinan Zhang, Zhiqing Zhong, Shilin He, Chaoyun Zhang, Qingwei Lin, Dan Pei, Pinjia He, Dongmei Zhang, and Qi Zhang. Openrca: Can large language models locate the root cause of software failures? In *International Conference on Learning Representations*, volume 2025, pp. 84562–84590, 2025.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*, 2022.

Sze Zhang Yong and Necmiye Ozay. Discrete state estimation with persistent sensor faults and non-persistent noise via noisy bayesian active diagnosis. In *2018 Annual American Control Conference (ACC)*, pp. 313–320. IEEE, 2018.

Jiaqi Yue, Chunhui Zhao, and Cheng Xu. An on-the-fly signals-to-semantics storytelling framework for generating explainable industrial maintenance decisions. *IEEE Transactions on Automation Science and Engineering*, 2026.

Fan Zhang, Sinchee Chin, Jing-Hao Xue, and Wenming Yang. Icdiffad: Implicit conditioning diffusion model for time series anomaly detection. In *International Conference on Learning Representations*, volume 2026, pp. 54492–54510, 2026.

Xiyuan Zhang, Ranak Roy Chowdhury, Rajesh K Gupta, and Jingbo Shang. Large language models for time series: A survey. *arXiv preprint arXiv:2402.01801*, 2024.

Chunhui Zhao. Perspectives on nonstationary process monitoring in the era of industrial artificial intelligence. *Journal of Process Control*, 116:255–272, 2022.

Alice X Zheng, Irina Rish, and Alina Beygelzimer. Efficient test selection in active diagnosis via entropy approximation. *arXiv preprint arXiv:1207.1418*, 2012.

A Anchored Normative Trajectory Recovery: Implementation and Analysis

The main text defines the normative recovery distribution and its role in diagnosis. This section specifies how its training pairs, probabilistic outputs, online anchors, and alarm thresholds are constructed. The subsequent analysis explains why excluding post-alarm observations preserves diagnostic contrast.

A.1 Patch representation and healthy anchor–target pairs

We first train a deterministic patch autoencoder on healthy trajectories. For the current implementation, both the history and reconstruction lengths are 32; patches have length four and stride two, producing 15 overlapping temporal tokens. The encoder retains this token geometry, and the decoder reconstructs the trajectory by overlap-add. It is optimized by mean squared reconstruction error and frozen before the probabilistic recovery stage.

Let a healthy trajectory be $\{(\boldsymbol{y}_t, \boldsymbol{c}_t)\}_{t=1}^{T}$, and let $\Delta_{\max}$ be the largest recovery interval supported by a dataset. For every admissible anchor ending at τ, we independently sample

$$\Delta \sim \mathrm{Unif}\{1, \ldots, \Delta_{\max}\} \tag{8}$$

and pair the anchor $\boldsymbol{A}_\tau = \boldsymbol{y}_{\tau-P+1:\tau}$ with the healthy target $\boldsymbol{X}^{\mathrm{h}}_{\tau+\Delta:\tau+\Delta+L-1}$. The current experiments draw one interval per admissible anchor in each constructed training set. The transition context between anchor and target and the available target-horizon context accompany every pair; target observations never enter the recovery input. Sampling the interval rather than fitting one fixed horizon trains a single model for both the rolling pre-alarm state and the expanding post-alarm horizon.

A.2 Gap-conditioned probabilistic propagation

The frozen encoder produces $\boldsymbol{Z}_\tau = E_\phi(\boldsymbol{A}_\tau)$. Future-position queries, parameterized by their continuous offset from the anchor, cross-attend to these historical tokens and are conditioned on the intervening and target-horizon context:

$$\boldsymbol{Z}_{t|\tau} = G_\psi(\boldsymbol{Z}_\tau, \Delta, \boldsymbol{C}_{\tau:t})\,. \tag{9}$$

No target-value token or teacher-forcing observation is supplied. The frozen decoder maps this representation to the conditional mean, while a separate head maps the same representation to one log variance per target time and variable:

$$\boldsymbol{\mu}_t = D_\omega(\boldsymbol{Z}_{t|\tau}, \boldsymbol{C}_{\tau:t}), \qquad \log \boldsymbol{\sigma}_t^2 = H_\nu(\boldsymbol{Z}_{t|\tau}). \tag{10}$$

The log variance is smoothly bounded between dataset-independent standard-deviation limits using a sigmoid parameterization. For a healthy target, the joint objective is

$$\mathcal{L}_{\mathrm{ANTR}} = \frac{1}{2LD} \sum_{\ell=1}^{L} \sum_{j=1}^{D} \left[\frac{(x^{\mathrm{h}}_{t+\ell-1,j} - \mu_{t,\ell,j})^2}{\sigma^2_{t,\ell,j}} + \log \sigma^2_{t,\ell,j} \right]. \tag{11}$$

Here, $\boldsymbol{\sigma}_t^2$ is the heteroscedastic residual scale in the likelihood: it weights target- and variable-specific errors, while the log-scale term prevents unbounded inflation. We jointly optimize G_ψ and H_ν by Eq. 11, select the checkpoint by validation negative log-likelihood, and keep the encoder and decoder frozen. Operational alarm and examination thresholds are calibrated separately from empirical healthy trajectories rather than treating the predicted scale as a calibrated decision threshold. Algorithm 2 summarizes the two training stages.

Algorithm 2. Training Anchored Normative Trajectory Recovery

1 Fit E_ϕ, D_ω on healthy windows by patch reconstruction.
2 Freeze E_ϕ, D_ω and construct random-gap pairs by Eq. 8.
3 **for** each anchor–target minibatch **do**
4 Encode the anchor and propagate future queries by Eq. 9.
5 Decode $\boldsymbol{\mu}_t$ and predict bounded $\log \boldsymbol{\sigma}_t^2$ by Eq. 10.
6 Update G_ψ, H_ν by the Gaussian objective in Eq. 11.
7 **end for**; select by validation NLL and freeze all parameters.

A.3 Online anchoring and pointwise monitoring

At physical time t before an alarm, the anchor and target advance together while maintaining a dataset-specific safety interval $g_{\min}$. The model predicts a complete length-L distribution, but monitoring retains the first offset $(\boldsymbol{\mu}_{t,1}, \boldsymbol{\sigma}_{t,1})$ so that the score at t uses no future observation. Once an alarm is confirmed, the last admissible anchor is frozen and subsequent targets are recovered at increasing intervals up to $\Delta_{\max}$. Generated values are never written back as observations.

For variable j, the pointwise alarm score is

$$S_{t,j} = (y_{t,j} - \mu_{t,1,j})^2\,. \tag{12}$$

Healthy calibration scores are transformed by $\log(1 + S)$, extreme upper-tail values are removed by a robust median/MAD rule, and a Gaussian kernel density estimate supplies the chosen upper

quantile. Its bandwidth is one tenth of the retained score range, and a small positive floor prevents a degenerate threshold when healthy residuals are nearly zero. An alarm is confirmed only after the required number of consecutive threshold exceedances. The quantile, safety interval, warm-up period, confirmation length, and maximum recovery interval are reported per dataset in Appendix E.1. RCAEv a l additionally ignores a short, nearly constant pulse in its dedicated Error channel during detection; this dataset-specific exemption ends when either its duration or magnitude limit is exceeded.

A.4 Paired rolling histories

We first specify how post-alarm observations affect a rolling predictor. Let $\mathbf{h}_j$ denote the history available when predicting the reference at time j. It may be an explicit sliding window, a recurrent hidden state, or any finite-dimensional representation of the observations used by the predictor. Write its update and output maps as F_j and G_j. Starting from the same pre-alarm state and using the same context $\boldsymbol{C}_j$, consider the paired paths

$$\begin{aligned} \mathbf{h}^{\delta}_{j+1} &= F_j\big(\mathbf{h}^{\delta}_j, \boldsymbol{X}^{\mathrm{h}}_j + \boldsymbol{D}_j, \boldsymbol{C}_j\big)\,, \quad & \widetilde{\boldsymbol{X}}^{\delta}_j &= G_j\big(\mathbf{h}^{\delta}_j, \boldsymbol{C}_j\big)\,, \\ \mathbf{h}^{0}_{j+1} &= F_j\big(\mathbf{h}^{0}_j, \boldsymbol{X}^{\mathrm{h}}_j, \boldsymbol{C}_j\big)\,, & \widetilde{\boldsymbol{X}}^{0}_j &= G_j\big(\mathbf{h}^{0}_j, \boldsymbol{C}_j\big)\,, \end{aligned} \tag{13}$$

with $\mathbf{h}^{\delta}_{t_0} = \mathbf{h}^{0}_{t_0}$. The superscript δ denotes the path driven by post-alarm observations $\boldsymbol{X}^{\mathrm{h}}_j + \boldsymbol{D}_j$, while 0 denotes its matched healthy path. All other inputs are shared. Define the history-state difference and the resulting reference difference as

$$\Delta\mathbf{h}_j := \mathbf{h}^{\delta}_j - \mathbf{h}^{0}_j, \qquad \boldsymbol{P}_j := \widetilde{\boldsymbol{X}}^{\delta}_j - \widetilde{\boldsymbol{X}}^{0}_j. \tag{14}$$

This pairing isolates the effect of feeding post-alarm observations into the rolling history. In particular, $\boldsymbol{P}_j$ is defined by two executions of the same predictor, not by a scalar update or a direction assumption.

A.5 Exact propagation through a differentiable predictor

Assume that the line segment joining the paired inputs lies in the domains of F_j and G_j, and that these maps are continuously differentiable along the segment. For the state update, define

$$\varphi_j(s) = F_j\big(\mathbf{h}^{0}_j + s\Delta\mathbf{h}_j, \boldsymbol{X}^{\mathrm{h}}_j + s\boldsymbol{D}_j, \boldsymbol{C}_j\big)\,, \qquad 0 \le s \le 1. \tag{15}$$

Its endpoints are $\varphi_j(0) = \mathbf{h}^{0}_{j+1}$ and $\varphi_j(1) = \mathbf{h}^{\delta}_{j+1}$. The fundamental theorem of calculus and the chain rule give

$$\begin{aligned} \Delta\mathbf{h}_{j+1} &= \int_0^1 \varphi_j'(s)\,\mathrm{d}s \\ &= \mathbf{A}_j\Delta\mathbf{h}_j + \mathbf{B}_j\boldsymbol{D}_j, \end{aligned} \tag{16}$$

where the path-averaged Jacobians are

$$\begin{aligned} \mathbf{A}_j &:= \int_0^1 J_h F_j\big(\mathbf{h}^{0}_j + s\Delta\mathbf{h}_j, \boldsymbol{X}^{\mathrm{h}}_j + s\boldsymbol{D}_j, \boldsymbol{C}_j\big)\,\mathrm{d}s, \\ \mathbf{B}_j &:= \int_0^1 J_y F_j\big(\mathbf{h}^{0}_j + s\Delta\mathbf{h}_j, \boldsymbol{X}^{\mathrm{h}}_j + s\boldsymbol{D}_j, \boldsymbol{C}_j\big)\,\mathrm{d}s. \end{aligned} \tag{17}$$

Here, $J_h F_j$ and $J_y F_j$ are the Jacobians with respect to the history state and the newly incorporated observation. Eq. 16 is exact: $\mathbf{A}_j$ and $\mathbf{B}_j$ are integrated along the segment between the paired inputs, so no Taylor remainder is discarded.

Applying the same argument to the output map gives

$$\mathbf{K}_t := \int_0^1 J_h G_t\big(\mathbf{h}^{0}_t + s\Delta\mathbf{h}_t, \boldsymbol{C}_t\big)\,\mathrm{d}s, \qquad \boldsymbol{P}_t = \mathbf{K}_t\Delta\mathbf{h}_t. \tag{18}$$

Since $\Delta\mathbf{h}_{t_0} = \mathbf{0}$, repeated substitution in Eq. 16 yields

$$\Delta\mathbf{h}_t = \sum_{i=t_0}^{t-1} \left(\mathbf{A}_{t-1}\mathbf{A}_{t-2}\cdots\mathbf{A}_{i+1}\right)\mathbf{B}_i\boldsymbol{D}_i, \tag{19}$$

where the product is the identity map when $i = t-1$. Combining Eqs. 18 and 19 gives the reference change induced by the complete post-alarm input path:

$$\boxed{\boldsymbol{P}_t = \mathbf{K}_t \sum_{i=t_0}^{t-1} \left(\mathbf{A}_{t-1}\mathbf{A}_{t-2}\cdots\mathbf{A}_{i+1}\right)\mathbf{B}_i\boldsymbol{D}_i}. \tag{20}$$

Each term has a direct temporal interpretation. The factor $\mathbf{B}_i\boldsymbol{D}_i$ writes the fault-affected observation at time i into the history state; the subsequent $\mathbf{A}_j$ factors propagate that state difference through later rolling updates; $\mathbf{K}_t$ maps the accumulated state difference to the reference predicted at time t.

Let

$$a_j := \|\mathbf{A}_j\|_{\mathrm{op}}, \qquad b_i := \|\mathbf{B}_i\|_{\mathrm{op}}, \qquad \kappa_t := \|\mathbf{K}_t\|_{\mathrm{op}}.$$

Taking norms in Eq. 20 gives

$$\|\boldsymbol{P}_t\| \le \kappa_t \sum_{i=t_0}^{t-1} b_i \|\boldsymbol{D}_i\| \prod_{j=i+1}^{t-1} a_j. \tag{21}$$

The matrices in Eq. 20 may rotate or cancel different fault contributions, and the predictor may attenuate affected inputs. The general recurrent construction therefore identifies the propagation path and bounds its magnitude; it does not imply that $\|\boldsymbol{P}_t\|$ grows monotonically or that $\boldsymbol{P}_t$ is aligned with $\boldsymbol{D}_t$.

A.6 Proof of the fitting bounds

Proof of Proposition 1. Define the anchored recovery error

$$\boldsymbol{E}_t^{\mathrm{A}} := \widehat{\boldsymbol{X}}_t - \boldsymbol{X}_t^{\mathrm{h}}.$$

The health-reference error is $e_t^{\mathrm{A}} = \|\boldsymbol{E}_t^{\mathrm{A}}\| \le \epsilon_t$. Since $\boldsymbol{Y}_t = \boldsymbol{X}_t^{\mathrm{h}} + \boldsymbol{D}_t$, the anchored diagnostic residual is

$$r_t^{\mathrm{A}} = \|\boldsymbol{Y}_t - \widehat{\boldsymbol{X}}_t\| = \|\boldsymbol{D}_t - \boldsymbol{E}_t^{\mathrm{A}}\|. \tag{22}$$

The reverse triangle inequality gives

$$\left|r_t^{\mathrm{A}} - d_t\right| \le \|\boldsymbol{E}_t^{\mathrm{A}}\| \le \epsilon_t, \tag{23}$$

which proves Eq. 2.

For the paired rolling predictor, define its error on the matched healthy path as

$$\boldsymbol{U}_t := \widetilde{\boldsymbol{X}}_t^{0} - \boldsymbol{X}_t^{\mathrm{h}}, \qquad \eta_t := \|\boldsymbol{U}_t\|.$$

Eq. 14 gives the exact decomposition

$$\widetilde{\boldsymbol{X}}_t^{\delta} = \widetilde{\boldsymbol{X}}_t^{0} + \boldsymbol{P}_t = \boldsymbol{X}_t^{\mathrm{h}} + \boldsymbol{U}_t + \boldsymbol{P}_t. \tag{24}$$

Consequently, its two fitting quantities are exactly

$$e_t^{\mathrm{R}} = \|\boldsymbol{U}_t + \boldsymbol{P}_t\|, \qquad r_t^{\mathrm{R}} = \|\boldsymbol{D}_t - \boldsymbol{U}_t - \boldsymbol{P}_t\|. \tag{25}$$

Applying the reverse triangle inequality gives

$$\left|e_t^{\mathrm{R}} - \|\boldsymbol{P}_t\|\right| \le \|\boldsymbol{U}_t\| = \eta_t, \qquad \left|r_t^{\mathrm{R}} - \|\boldsymbol{D}_t - \boldsymbol{P}_t\|\right| \le \|\boldsymbol{U}_t\| = \eta_t, \tag{26}$$

which proves Eq. 3. □

Eq. 24 separates two sources of rolling-reference error. The term $\boldsymbol{U}_t$ is the predictor's error on the matched healthy path. The term $\boldsymbol{P}_t$, characterized by Eq. 20, is the additional output change caused by feeding post-alarm observations into the rolling history.

Corollary 2 (Vanishing-fit contrast ratio)**.** *Let $\Gamma_t = e_t/r_t$ whenever $r_t > 0$. For fixed $\boldsymbol{D}_t$ and $\boldsymbol{P}_t$ with $\boldsymbol{D}_t \neq \boldsymbol{P}_t$, Eq. 2–equation 3 imply*

$$\epsilon_t \to 0 : \ \Gamma_t^{\mathrm{A}} \to 0, \qquad \eta_t \to 0 : \ \Gamma_t^{\mathrm{R}} \to \frac{\|\boldsymbol{P}_t\|}{\|\boldsymbol{D}_t - \boldsymbol{P}_t\|}. \tag{27}$$

If a particular rolling predictor satisfies the direction-preserving response $\boldsymbol{P}_t = \alpha_t \boldsymbol{D}_t$ with $0 \le \alpha_t < 1$, then $\Gamma_t^{\mathrm{R}} \to \alpha_t/(1-\alpha_t)$; this limit is nondecreasing whenever α_t is nondecreasing. Monotone growth of d_t alone does not imply this response property.

A.7 Proof of the contrast ratio and examination dead band

Proof of Corollary 2. When $d_t > \epsilon_t$, Eq. 2 implies

$$0 \le \Gamma_t^{\mathrm{A}} \le \frac{\epsilon_t}{d_t - \epsilon_t}, \tag{28}$$

so $\Gamma_t^{\mathrm{A}} \to 0$ as $\epsilon_t \to 0$. Eq. 3 gives

$$e_t^{\mathrm{R}} - \|\boldsymbol{P}_t\| \to 0, \qquad r_t^{\mathrm{R}} - \|\boldsymbol{D}_t - \boldsymbol{P}_t\| \to 0 \tag{29}$$

as $\eta_t \to 0$. Because $\boldsymbol{D}_t \neq \boldsymbol{P}_t$, the limiting denominator is positive, and continuity of division yields the rolling ratio in Eq. 27. If $\boldsymbol{P}_t = \alpha_t \boldsymbol{D}_t$ with $0 \le \alpha_t < 1$, homogeneity of the norm gives

$$\frac{\|\boldsymbol{P}_t\|}{\|\boldsymbol{D}_t - \boldsymbol{P}_t\|} = \frac{\alpha_t d_t}{(1-\alpha_t) d_t} = \frac{\alpha_t}{1-\alpha_t}. \tag{30}$$

The derivative of $\alpha/(1-\alpha)$ is $1/(1-\alpha)^2 > 0$ on $[0, 1)$, proving the final monotonicity statement. □

Finally, the ideal and recovered residual inputs to an examination differ by

$$(\boldsymbol{Y}_t - \widehat{\boldsymbol{X}}_t) - (\boldsymbol{Y}_t - \boldsymbol{X}_t^{\mathrm{h}}) = -\boldsymbol{E}_t^{\mathrm{A}}. \tag{31}$$

For the examination statistic ψ_a, the L_a-Lipschitz condition is

$$|\psi_a(\mathbf{V}) - \psi_a(\mathbf{W})| \le L_a \|\mathbf{V} - \mathbf{W}\| \tag{32}$$

for residual windows $\mathbf{V}$ and $\mathbf{W}$ in its calibrated domain. Setting $\mathbf{V} = \boldsymbol{Y}_t - \widehat{\boldsymbol{X}}_t$ and $\mathbf{W} = \boldsymbol{Y}_t - \boldsymbol{X}_t^{\mathrm{h}}$ yields

$$\left|\widehat{s}_{a,t} - s_{a,t}^{\star}\right| \le L_a \|\boldsymbol{E}_t^{\mathrm{A}}\| \le L_a \epsilon_t, \tag{33}$$

If $\widehat{s}_{a,t} > \delta_a + L_a \epsilon_t$, then $s_{a,t}^{\star} \ge \widehat{s}_{a,t} - L_a \epsilon_t > \delta_a$. Likewise, $\widehat{s}_{a,t} < \delta_a - L_a \epsilon_t$ implies $s_{a,t}^{\star} \le \widehat{s}_{a,t} + L_a \epsilon_t < \delta_a$. These two implications establish the stated dead-band decisions.

B Executable Examination Statistics and Binary Evidence

This section specifies how a selected examination is executed against the recovered normative trajectory. Eq. 33 supplies the common error-propagation principle; the remaining choices are the statistic used by each examination, its calibrated threshold, and its temporal persistence. The LLM neither observes raw trajectories nor assigns examination outcomes.

B.1 Health-relative effects and recovery-aware calibration

For variable j, Anchored Normative Trajectory Recovery supplies a causal pointwise normative mean $\boldsymbol{\mu}_j$. Its difference from the observation defines the health-relative effect

$$\mathbf{e}_{j,w} = \mathbf{y}_{j,w} - \boldsymbol{\mu}_{j,w} \tag{34}$$

over a length-W window w. To reproduce the pre- and post-alarm anchor geometry without using fault labels, we place a fixed number of pseudo alarms on each healthy validation trajectory, freeze the corresponding anchors, and collect complete effect windows. For system group s and variable j, their recovery radii are

$$\epsilon_{s,j}^{(W)} = \mathrm{Quantile}_q(\{\|\mathbf{e}_{j,w}\|_2\}_{\mathrm{healthy}}), \qquad \epsilon_{s,j}^{(2W)} = \mathrm{Quantile}_q(\{\|\mathbf{e}_{j,w:w+2W}\|_2\}_{\mathrm{healthy}}), \tag{35}$$

where $q = 0.99$ in the current experiments. For every non-oscillation examination, we also fit a one-dimensional Gaussian kernel density estimate to its healthy score and take its q-quantile $\tau_{s,j,a}$. The deployed threshold is

$$\delta_{s,j,a} = \max\{L_a \epsilon_{s,j}, \tau_{s,j,a}\}, \tag{36}$$

using $\epsilon^{(2W)}$ for step examinations and $\epsilon^{(W)}$ otherwise. Thus, an effect must exceed both the propagated recovery-error radius and the empirical healthy-score range. Availability checks are performed before planning; every examination admitted to the action set has a valid normative reference and calibrated threshold.

Table 3: Executable examination statistics. For the first six rows, a window supports the named phenomenon when its target-oriented score exceeds the calibrated threshold in Eq. 36. The current implementation uses $W = 8$. Oscillation uses the morphology gates described below.

Examination	Target-oriented score ψ_a	Window	L_a	Temporal requirement
Higher	$\bar{e}$	W	$1/\sqrt{W}$	K_a consecutive windows
Lower	$-\bar{e}$	W	$1/\sqrt{W}$	K_a consecutive windows
Rise Step	$d(\mathbf{e})$	$2W$	$\sqrt{2/W}$	one window pair
Drop Step	$-d(\mathbf{e})$	$2W$	$\sqrt{2/W}$	one window pair
Rise Ramp	$b(\mathbf{e})$	W	$\|\mathbf{i} - \bar{i}\|_2^{-1}$	K_a consecutive windows
Drop Ramp	$-b(\mathbf{e})$	W	$\|\mathbf{i} - \bar{i}\|_2^{-1}$	K_a consecutive windows
Oscillation	detrended energy and shape gates	evidence interval	—	complete morphology test

B.2 Target-oriented examination statistics

For a scalar length-W effect path $\mathbf{e} = (e_0, \ldots, e_{W-1})$, define its mean $\bar{e} = W^{-1} \sum_i e_i$ and least-squares slope

$$b(\mathbf{e}) = \frac{\sum_{i=0}^{W-1} (i - \bar{i})(e_i - \bar{e})}{\sum_{i=0}^{W-1} (i - \bar{i})^2}, \qquad \bar{i} = \frac{W-1}{2}. \tag{37}$$

For adjacent length-W windows $\mathbf{e}^{\mathrm{L}}$ and $\mathbf{e}^{\mathrm{R}}$, let $d(\mathbf{e}) = \overline{\mathbf{e}}^{\mathrm{R}} - \overline{\mathbf{e}}^{\mathrm{L}}$. Table 3 lists the seven examination families. Every score is target-oriented: a larger value supports the phenomenon named in the first column.

The level and ramp examinations require $K_a = 3$ consecutive overlapping windows in the current implementation; step examinations require one adjacent-window pair. Oscillation is treated separately because a large step or ramp can also have high local variance. We smooth the post-alarm effect, remove its least-squares trend, and compare its residual RMS with the corresponding pre-alarm RMS. A *Hit* additionally requires repeated derivative-direction reversals, balanced upward and downward derivatives, and excursions on both sides of the healthy level. These shape gates prevent monotone changes from being mislabeled as oscillation.

B.3 Binary outcome aggregation

Let $\mathcal{W}_a$ be the ordered set of valid windows for examination a and let K_a be its persistence requirement. The executor returns

$$O_a = \begin{cases} \text{Hit}, & \exists\, K_a \text{ consecutive } w:\ \psi_a(\mathbf{e}_{j,w}) > \delta_{s,j,a}, \\ \text{Miss}, & \text{otherwise}, \end{cases} \tag{38}$$

with the step and oscillation specializations described above. The scan records the first supporting interval, score, threshold components, and observed persistence for replay. Hence *Miss* means that the named phenomenon was not supported under its executable calibrated rule; it does not assert the opposite phenomenon.

C Bayesian Evidence Model and Experience-Grounded Diagnostic Evolution

C.1 Initial evidence likelihoods

For each benchmark, domain documents and fault descriptions are retrieved to identify candidate fault–symptom relations. Experts retain relations that can be expressed by an available variable–examination pair. This retrieval step defines the action vocabulary; it does not produce a diagnosis or directly modify a case posterior. The same two-state initialization rule is used across benchmarks. An explicitly supported relation receives $\mathrm{Beta}(8, 2)$, whose mean is 0.8 and whose prior sample size is 10. A pair without an explicit relation receives the neutral $\mathrm{Beta}(1, 1)$ prior, whose mean is 0.5.

These values encode a fixed expert-supported relation and an unspecified relation, respectively; they are prior states rather than data-fitted optima. Their posterior means form the evidence model used by Eq. 4 before historical outcome calibration.

C.2 Batched likelihood calibration

After a historical case has a resolved target $f^\star$, every executable binary result can calibrate the corresponding row of the evidence model. Because each examination is a deterministic statistic of the recorded health-relative residuals, all available variable–examination pairs can be evaluated retrospectively for a historical case. A result is eligible when the required samples are present, the tool's validity conditions hold, and the case target has been resolved independently of the inquiry trace.

For each fault–examination pair (f, a), let $\theta_{f,a} = P(\mathrm{Hit} \mid f, a)$ and assign the prior $\theta_{f,a} \sim \mathrm{Beta}(\alpha_{f,a}, \beta_{f,a})$. After accumulating $n^{\mathrm{H}}_{f,a}$ eligible ***Hit*** outcomes and $n^{\mathrm{M}}_{f,a}$ eligible ***Miss*** outcomes, the posterior is

$$\theta_{f,a} \mid \mathcal{C} \sim \mathrm{Beta}\big(\alpha_{f,a} + n^{\mathrm{H}}_{f,a}, \beta_{f,a} + n^{\mathrm{M}}_{f,a}\big)\,.$$

We use its posterior mean,

$$\widehat{\theta}_{f,a} = \frac{\alpha_{f,a} + n^{\mathrm{H}}_{f,a}}{\alpha_{f,a} + \beta_{f,a} + n^{\mathrm{H}}_{f,a} + n^{\mathrm{M}}_{f,a}}, \tag{39}$$

as the calibrated evidence likelihood. For each pair (f, a), we accumulate $n^{\mathrm{H}}_{f,a}$ and $n^{\mathrm{M}}_{f,a}$ across eligible historical cases and apply Eq. 39. This population is independent of which examinations the online planner selected and of the impact threshold used for the action-experience library, preventing the action-selection policy from determining which evidence probabilities can be estimated.

Calibration uses historical validation labels only. The resulting likelihood table is frozen before evaluation on held-out test cases, and no test label enters the update. In a continuing deployment, the same sufficient statistics can be extended when another resolved case becomes historical, but the update remains batched rather than being triggered by every individual outcome.

C.3 Constrained planner interface and prompt schema

The prompt exposes only decision variables available at the current inquiry round. Its instantiated form is stored with the raw model response in the diagnostic trace. The schema is:

Block	**Prompt content**
Current state	Ranked diagnostic posterior, executed examinations with observed *Hit*/*Miss* outcomes, and the remaining feasible action identifiers.
Candidate action	Examination meaning, involved variables, linked active hypotheses, immediate information gain, Q_K, cost, and the best continuation under each modeled outcome.
Signal context	Health-relative variable contributions, explicitly described as evidence concentration rather than causal fault attribution.
Historical context	Retrieved case lessons labeled as historical strategy experience that cannot substitute for current evidence.
Decision rules	Select one current feasible action, do not repeat an action, do not predict its outcome, and do not modify likelihoods or the posterior.
Output contract	A JSON object containing only `selected_action_id` and a concise `selection_reason`.

The planner's decision is restricted to the current feasible set. The selected tool then produces the binary outcome, and Eq. 4 performs the posterior update; neither tool execution nor probabilistic inference is delegated to the language model.

C.4 Action experience and experimental history protocol

For a resolved case, each state–action–outcome transition is scored by Eq. 7. Transitions satisfying $|I_r| \geq \tau_I$ are retained together with their numerical before/after state, diagnostic target, and tool

provenance. An LLM compresses this record into a case summary, a transferable lesson, and conditional guidance; the underlying transition remains attached so that the memory is auditable. Retrieval compares the current posterior, executed and remaining examinations, and variable-contribution profile with historical states, then supplies the most similar lessons to the planner. Retrieved text can influence action selection but cannot create an outcome or update the posterior.

The framework supports chronological insertion of a resolved case before the next episode, as Algorithm 1 shows. The reported experiments emulate a system that already has an accumulated history: the experience library and likelihood table are first built from designated historical validation cases, then frozen for all held-out test episodes. This protocol avoids test-order effects and label leakage while evaluating the benefit of different amounts of prior diagnostic experience.

D Derivation of the Finite-Horizon Inquiry Value

We derive Eq. 6 for a general finite outcome set $\mathcal{O}_a$. The current EviDiag planner specializes this set to $\mathcal{O}_a^{\text{plan}} = \{\text{Hit}, \text{Miss}\}$. Let b be the current posterior over $\mathbb{F}$, and let $\mathbb{A}$ contain the examinations that remain available. For $a \in \mathbb{A}$, write $\mathbb{A}_{-a} := \mathbb{A} \setminus \{a\}$ for the remaining examinations. Marginalizing the diagnostic-hypothesis-conditional likelihood of a candidate examination gives

$$p(o \mid b, a) = \sum_{f \in \mathbb{F}} b(f) P(o \mid f, a). \tag{40}$$

For any outcome with $p(o \mid b, a) > 0$, the corresponding belief is

$$\mathcal{B}(b, a, o)(f) = \frac{b(f) P(o \mid f, a)}{p(o \mid b, a)}. \tag{41}$$

The expected entropy after executing a is therefore $\sum_{o \in \mathcal{O}_a} p(o \mid b, a) H(\mathcal{B}(b, a, o))$. Subtracting it from the current entropy gives

$$\text{IG}(b, a) = H(b) - \sum_{o \in \mathcal{O}_a} p(o \mid b, a) H(\mathcal{B}(b, a, o)) \,. \tag{42}$$

A policy π maps each observed inquiry history to the next available examination. Starting from $(b, \mathbb{A})$ with at most k examinations remaining, define its cumulative information value as

$$J_k^\pi(b, \mathbb{A}) = \mathbb{E}_\pi \left[\sum_{\ell=0}^{k-1} \gamma^\ell \, \text{IG}(b_\ell, a_\ell) \,\middle|\, b_0 = b, \ \mathbb{A}_0 = \mathbb{A} \right]. \tag{43}$$

If the available set becomes empty before k examinations have been performed, the policy stops and accrues no further value; its terminal belief is held fixed over the unused indices.

We first fix the initial examination to a. For $k \geq 2$, the sum inside Eq. 43 separates as

$$\begin{aligned} \sum_{\ell=0}^{k-1} \gamma^\ell \, \text{IG}(b_\ell, a_\ell) = \ & \text{IG}(b, a) \\ & + \gamma \sum_{j=0}^{k-2} \gamma^j \, \text{IG}(b_{j+1}, a_{j+1}). \end{aligned} \tag{44}$$

The first term is determined by the current state and the fixed examination. The second term starts from the state reached after observing its result. If the result is o, that state is

$$S'_o = (\mathcal{B}(b, a, o), \mathbb{A}_{-a}) \,. \tag{45}$$

The continuation policy can depend on this observed result; denote it by π_o. Applying the law of total expectation over the first result yields

$$J_k^{(a, \{\pi_o\})}(b, \mathbb{A}) = \text{IG}(b, a) + \gamma \sum_{o \in \mathcal{O}_a} p(o \mid b, a) J_{k-1}^{\pi_o}(\mathcal{B}(b, a, o), \mathbb{A}_{-a}) \,. \tag{46}$$

Define the largest cumulative information value attainable from a state by

$$V_j(b', \mathbb{A}') = \max_{\pi'} J_j^{\pi'}(b', \mathbb{A}'). \tag{47}$$

Once o is observed, only its associated branch is followed. The continuation policies π_o can consequently be chosen separately across outcomes. Maximizing Eq. 46 over the collection $\{\pi_o : o \in \mathcal{O}_a\}$ gives

$$\begin{aligned} Q_k(b, \mathbb{A}, a) &= \max_{\{\pi_o\}} J_k^{(a,\{\pi_o\})}(b, \mathbb{A}) \\ &= \mathrm{IG}(b, a) + \gamma \sum_{o \in \mathcal{O}_a} p(o \mid b, a) \max_{\pi_o} J_{k-1}^{\pi_o}(\mathcal{B}(b, a, o), \mathbb{A}_{-a}) \\ &= \mathrm{IG}(b, a) + \gamma \sum_{o \in \mathcal{O}_a} p(o \mid b, a) V_{k-1}(\mathcal{B}(b, a, o), \mathbb{A}_{-a}) \,. \end{aligned} \tag{48}$$

The first examination has not yet been fixed when evaluating the state value, so maximizing once more over $a \in \mathbb{A}$ gives

$$V_k(b, \mathbb{A}) = \max_{a \in \mathbb{A}} Q_k(b, \mathbb{A}, a), \qquad V_0(b, \mathbb{A}) = V_k(b, \varnothing) = 0. \tag{49}$$

For $k = 1$, the continuation value is zero and $Q_1(b, \mathbb{A}, a) = \mathrm{IG}(b, a)$. Assuming V_{k-1} gives the largest value for the remaining $k - 1$ examinations, Eq. 48 evaluates every possible first result followed by its best continuation, and the outer maximization selects the best first examination. Starting from the $k = 1$ case, this argument establishes the recursion for every finite k.

For the two-step horizon used by EviDiag, substituting $V_1(b', \mathbb{A}') = \max_{a' \in \mathbb{A}'} \mathrm{IG}(b', a')$ gives the explicit form

$$Q_2(b, \mathbb{A}, a) = \mathrm{IG}(b, a) + \gamma \sum_{o \in \mathcal{O}_a^{\mathrm{plan}}} p(o \mid b, a) \max_{a' \in \mathbb{A}_{-a}} \mathrm{IG}(\mathcal{B}(b, a, o), a') \,. \tag{50}$$

If $\mathbb{A}_{-a}$ is empty, the continuation term in Eq. 50 is zero by definition. Thus, the first examination is scored by its immediate information gain and by the best second examination available under each possible first outcome.

When $\gamma = 1$, the cumulative objective also has a terminal-entropy form. From Eq. 42,

$$\mathrm{IG}(b_\ell, a_\ell) = H(b_\ell) - \mathbb{E}[H(b_{\ell+1}) \mid b_\ell, a_\ell] \,. \tag{51}$$

Taking expectations and summing over ℓ cancels the intermediate entropy terms:

$$\begin{aligned} J_k^\pi(b, \mathbb{A}) &= \sum_{\ell=0}^{k-1} \left(\mathbb{E}_\pi[H(b_\ell)] - \mathbb{E}_\pi[H(b_{\ell+1})] \right) \\ &= H(b) - \mathbb{E}_\pi[H(b_k)]. \end{aligned} \tag{52}$$

Maximizing cumulative information gain at $\gamma = 1$ is therefore equivalent, under the planning likelihood model, to minimizing the expected fault-belief entropy after the final planned examination.

Eq. 48 applies to any finite modeled outcome set. The current EviDiag planner and executor both use the binary set $\{\mathrm{Hit}, \mathrm{Miss}\}$. The resulting Q_k remains an information-value input to the constrained LLM planner; examination cost, variable contribution, and retrieved experience remain separate inputs to the final action selection.

E Experimental Details

E.1 Datasets

We evaluate EviDiag on two multivariate time-series benchmarks. Complete trajectories, rather than windows sampled across splits, are assigned to training, validation, and test sets. Training data are used to fit the normative-recovery model; validation data support checkpoint selection, healthy-score calibration, and construction of the historical evidence available at test time. Test cases never enter the historical pool.

Table 4: Online recovery and alarm settings. $\Delta_{\max}$ is the largest anchor-to-target interval, $g_{\min}$ is the pre-alarm safety interval, and K_{alarm} is the number of consecutive pointwise exceedances required to confirm an alarm.

$\Delta_{\max}$	$g_{\min}$	Warm-up	K_{alarm}	Pseudo alarms/trajectory
100	30	100	10	5

RCAEval. RCAEval (Pham et al., 2025) contains traces from Online Boutique (OB), Sock Shop (SS), and Train Ticket (TT), with injected CPU, memory, disk, socket, delay, and loss faults. Its 270 one-second-resolution cases each contain a 720-point healthy prefix followed by fault injection. Following the fixed local protocol, the first two repetitions form a development pool and the third is held out for testing, yielding 144/36/90 train/validation/test cases (48/12/30 per system). The diagnostic comparison uses the common 60 OB/SS test cases. The complete 36-case validation split forms the historical pool. Each observation contains eight system-level aggregates—CPU, memory, disk I/O, socket, workload, error, and P50/P90 latency—and a three-dimensional one-hot system indicator provides operating context. Preprocessing statistics are fitted only on healthy prefixes in the training split.

E-TEP. For the Extended Tennessee Eastman Process benchmark (Reinartz et al., 2021), we retain ten fault categories and the first 30 source-order runs of each category. Every run contains 2,000 observations over 52 channels (41 measured and 11 manipulated variables); samples 1–599 form the healthy prefix, and fault injection begins at zero-based index 599. With seed 20260920, complete runs in every category are divided 20/5/5 into train/validation/test sets, producing 200/50/50 cases. All 52 variables are retained. The 50 validation cases form the historical pool, and Table 1 evaluates all 50 held-out test runs. Because the retained data shares aligned operating mode (Mode 1), E-TEP uses a constant context token and is not presented as mode-conditioned recovery.

E.1.1 Normative-recovery training

Both benchmarks use an observed window and recovery horizon of 32, with patch length four and stride two. The encoder has two Transformer layers, four attention heads, latent width 64, and feed-forward width 128; the future-query propagation module uses two layers with the same hidden width. The deterministic reconstruction stage is trained for 100 epochs with AdamW, batch size 256, weight decay 10^{-5}, and learning rates 5×10^{-5} on RCAEval and 2×10^{-4} on E-TEP.

The probabilistic propagation stage freezes the reconstruction model and optimizes the interval-conditioned mean and diagonal variance by Gaussian negative log-likelihood. It is trained for at most 200 epochs with learning rate 2×10^{-4}; each admissible anchor contributes one interval sampled uniformly from $[1, 100]$. Standard deviations are constrained to $[0.05, 10]$, and the checkpoint with the lowest validation NLL is retained. Online recovery uses 32 samples from the conditional predictive distribution. Experiments were conducted on a Windows 11 workstation with an Intel Core i5-13500H CPU and 16 GB RAM, using Python 3.9.21 and PyTorch 2.7.1. The reported recovery results use a single fixed-seed run, rather than averages over repeated training runs.

E.1.2 Online recovery and alarm settings

Only parameters that affect the experimental protocol are reported here. Monitoring scores are calibrated from healthy data at the 0.99 quantile. Their robust normalization uses a clipping threshold of 15 and a 200-point density grid. Diagnostic examinations use windows of length $W = 8$; persistent-pattern examinations require three consecutive exceedances, and their Hit/Miss thresholds use the same 0.99 healthy quantile. Table 4 lists the remaining online settings. For RCAEval, short isolated pulses in the Error channel are excluded from alarm confirmation when their absolute magnitude is at most 0.2 and their duration is at most 15 samples; this rule is fixed before test evaluation.

E.2 Historical Evidence and Inquiry Protocol

Historical evidence is constructed only from the fixed validation split: 36 complete cases for RCAEval and 50 for E-TEP. After replaying each resolved validation case, every executable examination

is evaluated retrospectively from its health-relative residuals, including examinations not selected by the original inquiry. Eligible outcomes are accumulated with the same priors, eligibility rules, and calibration configuration used by the local pipeline. Likelihood calibration is enabled when the historical validation pool contains at least ten completed cases. Once enabled, every eligible fault–examination cell is updated using its corresponding historical outcomes and the specified Beta prior. The resulting calibrated likelihood table and complete action-experience library are frozen before test evaluation; neither is updated from a test label. In particular, all E-TEP results in Table 1 use this calibrated table and the complete 50-case validation library.

An experience snapshot is retained when its posterior impact is at least 0.05. At each inquiry round, the retriever returns the three closest eligible snapshots. The planner evaluates the five highest-posterior fault hypotheses with two-step lookahead and discount factor one, while restricting examinations to the available tool set. It may inspect the three most contributive variables on RCAEval and ten on E-TEP. We use fixed budgets of 5 inquiry rounds for RCAEval and 15 for E-TEP, with deterministic decoding.

E.3 Evaluation Protocols and Metrics

Let N be the number of test episodes and r_i the rank of the ground-truth fault in episode i. We report Recall@k and truncated mean reciprocal rank at the ranking depths available in the corresponding result table:

$$\mathrm{R@}k = \frac{1}{N}\sum_{i=1}^{N}\mathbf{1}[r_i \le k], \qquad \mathrm{MRR@}K = \frac{1}{N}\sum_{i=1}^{N}\frac{\mathbf{1}[r_i \le K]}{r_i}. \tag{53}$$

An unreturned label, missed alarm, or absent diagnostic output contributes zero. Because each episode has one ground-truth class, R@1 is also top-1 accuracy. Macro-F1 is the unweighted mean of the class-wise F1 scores, with zero assigned to a class whose precision–recall denominator is zero. All reported values are multiplied by 100 and rounded to two decimal places. For the budget curves, the posterior after at most 5 examinations on RCAEval or 15 on E-TEP determines the prediction; an episode that terminates earlier retains its last posterior.

The RCAEval end-to-end evaluation contains the same 60 OB/SS test cases for every method. Its two missed alarms remain in the denominator and receive zero credit. The E-TEP evaluation contains all 50 test runs and has no missed alarms, so its metric denominator is 50 for every reported EviDiag result. Both evaluations use the fixed split and validation-derived history described above.

E.4 Baselines

We use the following six baselines, covering neural time-series models and LLM-based diagnostic reasoning:

- **Passive NB** (Dey & Stori, 2005): This baseline applies Naive Bayes to the complete set of executable binary examinations. It shares the fault prior, examination candidates, diagnostic tools, and calibrated fault–examination likelihoods with EviDiag. Unlike sequential diagnosis methods, Passive NB does not select examinations adaptively. Instead, it executes all available examinations and aggregates their *Hit*/*Miss* outcomes under the conditional-independence assumption to directly compute the final fault posterior and ranking.
- **Active NB-IG** (Zheng et al., 2012): This baseline shares the fault prior, examination space, diagnostic tools, and calibrated fault–examination likelihoods with EviDiag. At each inquiry round, it selects the remaining examination with the largest exact one-step expected reduction in posterior entropy, executes the corresponding tool, and updates the Naive Bayes posterior. The process terminates when the inquiry budget is reached, the action set is exhausted, or no remaining examination has positive information gain. We use the same inquiry budgets as EviDiag: at most five examinations on RCAEval and 15 examinations on E-TEP.
- **FaultExplainer** (Khan et al., 2025): FaultExplainer adopts statistical detection and LLM diagnosis workflow to the benchmark fault classes. A PCA model fitted on healthy training observations detects deviations and identifies contributing variables; the LLM then combines this evidence with domain fault descriptions to produce class scores and a diagnostic

explanation. PCA retains 95% of training variance, and detection requires exceedance of the Hotelling T^2 or squared-prediction-error threshold. The diagnostic prompt includes knowledge-base symptom descriptions and matched executable evidence (the same as our corresponding settings). We use the same LLM backbone with our method, DeepSeek-V4.1-Flash with temperature zero, and restrict observations to the shared diagnosis endpoint. The final scores rank the ten candidate faults. This local adaptation follows the statistical and diagnostic workflow of the official implementation.

- **CoT (Zero-shot Chain-of-Thought)** (Kojima et al., 2022): This baseline prompts the LLM to reason over case-level time-series statistics and return a diagnosis without worked examples. The input summarizes eight aggregated variables on RCAEval and 52 process variables on E-TEP, including healthy-reference statistics and changes during the observation interval. Compared to our method which can stably output auditable results, CoT can solely provide candidate fault labels with post-hoc explanations. We use DeepSeek-V4.1-Flash with temperature zero, the same with our method to implement CoT. The returned class scores determine the final ranking over six RCAEval faults or ten E-TEP faults.
- **ReAct** (Yao et al., 2022): This baseline interleaves LLM reasoning, tool selection, and returned observations to construct a diagnosis. At each step, the agent can request additional evidence or terminate with fault-class scores and an explanation. The recorded E-TEP configuration provides tools for querying healthy references, identifying anomalous variables and checking temporal patterns. It also uses DeepSeek-V4.1-Flash with temperature zero for diagnosis reasoning.
- **OpenRCA-style RCA agent** (Xu et al., 2025): We adapt OpenRCA's iterative Controller–Executor design to the common time-series input. The Controller issues one dataframe-analysis instruction per step; the Executor translates it into Python, executes it on the observed variables, and returns the numerical result for subsequent reasoning. The agent receives the pre-diagnosis history and processed observations only through the end of the diagnosis interval: eight aggregated variables on RCAEval and 52 process variables on E-TEP. We also use DeepSeek-V4.1-Flash with temperature zero and at most 5 (RCAEval) or 15 (E-TEP) rounds of / Controller–Executor steps in the recorded local configuration. The final call ranks every candidate fault class, and a missed alarm is retained as an end-to-end error without invoking the agent.

We standardize the task-level evaluation across methods: the training/validation/test partitions, held-out test cases, candidate fault labels, and evaluation metrics are fixed. The recorded LLM configurations use the same decoding temperature ($T = 0$). In the controlled RCAEval inquiry comparison, we additionally fix the initialization, calibrated likelihoods, and examination interface, and compare policies at the same inquiry rounds. Within this evaluation framework, each baseline follows its corresponding implementation for preprocessing, tool use, and inference, with dataset and interface adaptations documented above. The main-table comparison retains method-specific model configurations and query limits, also reported above; matched-round policy comparisons provide the controlled evidence for the inquiry mechanism.

LLM-based compared methods like ReAct and OpenRCA operate within individual diagnostic episodes: their observations, tool outputs, and intermediate reasoning states are retained only for the current case, and their native protocols do not define cross-case adaptation. In contrast, EviDiag introduces an outer evolution loop that learns from independently resolved historical cases by calibrating the evidence likelihoods and retrieving state-matched action experience. Accordingly, the main comparison evaluates the end-to-end diagnostic performance of each framework under its native information-use protocol, including its ability to exploit historical diagnostic experience when supported by the method. This comparison should not be interpreted as isolating the effect of the within-case inquiry policy alone. To isolate inquiry-policy effects, the controlled RCAEval study compares alternative action-selection policies using the same post-initialization posterior, calibrated evidence model, examination interface, and inquiry rounds (See Section 4.4.2). The LLM-based variants also share the same historical experience.

E.5 LLM Inquiry Settings

The following templates reproduce the current online inquiry prompts. Angle-bracketed fields are populated from the current diagnostic state; posterior, executed-tool, variable-contribution, and

candidate-tool entries repeat as needed. The planner receives one system message and one user message per round, with temperature zero and JSON-object output. The system message and main user-message structure are shared by RCAEval and E-TEP. The constrained interface is described in Appendix C.3.

System message.

```
Select exactly one legal current diagnostic action. The conditional
    next-action recommendations in the prompt describe future rounds
    after a possible Hit or Miss; they are not additional actions for
    the current round. Do not invent posterior values, tool results, or
    action IDs. Return JSON with selected_action_id and selection_reason.
```

User message template.

```
SEQUENTIAL FAULT-DIAGNOSIS TASK
You are the decision assistant in a sequential fault-diagnosis process.
    A healthy-recovery model has already produced the healthy reference
    X_minus. Diagnostic tools examine the innovation Y - X_minus to test
    concrete fault phenomena.
Your role in this round is only to choose ONE remaining diagnostic tool.
    You do not execute the tool, predict or invent its result, or update
    the fault posterior. After your choice, the numerical tool executor
    will return Hit, Miss, or Indeterminate, and the Bayesian module
    will perform the update.

CURRENT DIAGNOSTIC STATE
1. Current fault posterior (ranked):
  <rank>. <fault>: <posterior_percent>

2. Tools already executed and their observed results:
  <index>. <executed_action_id> -> <observed_result>

3. Remaining tools:
  <candidate_count> executable candidates remain. Their complete IDs and
    meanings are listed under CANDIDATE TOOLS below.

4. Variable contributions to the current fault-related anomaly evidence
    (ranked):
  <rank>. <variable>: <contribution_percent>
  Definition: each percentage is the variable's normalized share of the
    median absolute innovation |Y - X_minus| in the diagnostic buffer.
    It indicates where abnormal residual evidence is concentrated; it is
    not causal attribution to a fault class.

HOW TO READ THE NUMERICAL PLANNING INFORMATION
- IG is the expected immediate reduction in fault-posterior uncertainty
    from executing the candidate now.
- Q_K is the K-step look-ahead value: current IG plus expected future
    diagnostic value from later rounds.
- Cost is reported separately and is not subtracted inside Q_K.
- The Hit/Miss branch suggestions identify a possible next tool for each
    future outcome. They are planning guidance only; neither outcome is
    known before the current tool is actually executed.

CANDIDATE TOOLS
[<rank>] Action ID: <action_id>
    Examination: <examination>
    Variables: <variable>
    Knowledge-base-linked active faults: <fault>
    Relevant variable-evidence shares: <variable>=<contribution_percent>
    Diagnostic cost: <cost>
    Immediate IG: <IG>
    K-step value Q_K: <Q_K>
```

```
    If the actual result is Hit, a possible next tool is:
    <next_action_if_hit>
    If the actual result is Miss, a possible next tool is:
    <next_action_if_miss>

RETRIEVED REFLECTION MEMORY
The following text comes from historical diagnosed cases. Use it only as
    decision-strategy experience, not as evidence about the current case:
<retrieved_historical_experience>

DECISION RULES
1. Choose the single current tool that best distinguishes the plausible
    competing faults; do not merely seek confirmation of the
    highest-probability fault.
2. Consider Q_K as multi-step diagnostic value, IG as immediate value,
    and diagnostic cost separately.
3. Use variable contribution to judge whether a candidate examines where
    current anomaly evidence is concentrated, but do not treat it as
    proof of any particular fault.
4. If multiple candidate actions have the same or very similar
    information value, prefer an action on a variable that has not yet
    been examined. If an unexamined variable has clearly lower
    information value or substantially lower fault-related variable
    contribution, choose the higher-value action instead.
5. Respect completed-tool outcomes and never repeat an executed tool.
6. Use branch suggestions to compare possible diagnostic sequences;
    select only the current action, not a future branch action.
7. When retrieved reflection memory is available, use similar Good
    Actions as positive strategy experience and similar Bad Actions as
    warnings. Trust this experience more in the first inquiry rounds,
    when current evidence is sparse; as current Hit/Miss evidence
    accumulates, give the current posterior, Q_K, and variable evidence
    greater weight. Historical advice never overrides action legality.

REQUIRED OUTPUT
Return exactly one JSON object with only these two fields:
{"selected_action_id": "<exact candidate Action ID>",
    "selection_reason": "<concise reason for choosing it over the other
    current candidates>"}
Do not add Markdown, extra fields, an assumed tool result, or a
    posterior update.
```

When no tool has been executed, the executed-tool block reads "None. No diagnostic tool has been executed yet." Missing variable contributions are rendered as "None available." If no historical experience is retrieved, the reflection-memory block reads "None for this round." The historical-memory field contains the retrieved text, rather than a newly generated observation about the current case.

E.6 Additional Experimental Results

E.6.1 Paired significance analysis

We assess whether the observed performance differences are supported at the episode level. The analysis uses the final case ranking saved for every method and the same paired test cohort: 60 RCAEval episodes and 50 E-TEP episodes. For method m, let

$$h_{i,k}^{(m)} = \mathbf{1}\left[r_i^{(m)} \leq k\right] \tag{54}$$

denote whether the ground-truth fault of episode i appears within the first k positions. We compare EviDiag with each baseline separately while retaining the pairing induced by the common episode identifier.

Recall@k. For each $k \in \{1, 3, 5\}$, we form the two discordant counts n_{10} and n_{01}, where n_{10} counts episodes solved by EviDiag but not by the baseline and n_{01} counts the reverse. Because both benchmarks are small and the discordant counts can be sparse, we use the two-sided exact McNemar test rather than its asymptotic chi-square approximation:

$$p_{\mathrm{McN}} = \min\left\{1,\ 2 \sum_{j=0}^{\min(n_{10},n_{01})} \binom{n_{10}+n_{01}}{j} 2^{-(n_{10}+n_{01})}\right\}. \tag{55}$$

This test evaluates whether the two methods have equal paired success probabilities at the selected ranking depth.

MRR@3 and Macro-F1. The per-episode reciprocal-rank contribution is

$$z_i^{(m)} = \frac{\mathbf{1}[r_i^{(m)} \le 3]}{r_i^{(m)}}. \tag{56}$$

For MRR@3, we apply an exact paired sign-flip randomization test to the differences $z_i^{(\mathrm{EviDiag})} - z_i^{(m)}$. Zero differences are removed, and all sign assignments of the remaining differences are aggregated exactly. Macro-F1 is nonlinear and cannot be decomposed into independent per-episode scores. We therefore use a paired label-swap randomization test: for every episode, the two methods' predictions are exchanged with probability $1/2$, Macro-F1 is recomputed for both methods, and the absolute difference is compared with the observed value. We use 200,000 randomizations and the plus-one correction

$$\widehat{p}_{\mathrm{F1}} = \frac{1 + \#\{|\Delta_b| \ge |\Delta_{\mathrm{obs}}|\}}{200{,}001}. \tag{57}$$

Confidence intervals and multiplicity. All effect sizes are reported as percentage-point differences, $\Delta = \mathrm{EviDiag} - \mathrm{baseline}$. We construct percentile 95% confidence intervals from 50,000 class-stratified paired bootstrap samples. Each bootstrap replicate resamples episodes with replacement within every ground-truth fault class and applies the same sampled indices to both methods. This preserves the class composition needed by Macro-F1 and retains the paired comparison. Within each dataset, Holm's step-down procedure adjusts the complete family of 35 tests formed by seven baselines and five metrics. We use $q < 0.05$ as the family-wise significance criterion. These tests quantify uncertainty over the observed test episodes conditional on the saved runs; they do not estimate variation across new model runs or random seeds.

Results. EviDiag achieves the best R@1, MRR@3, and Macro-F1 on both benchmarks. On RCAEval, it obtains 58.33% R@1, 70.00% MRR@3, and 60.40% Macro-F1. After Holm correction over all 35 comparisons, EviDiag remains significantly better than FaultExplainer on all five metrics and significantly improves Macro-F1 over Active NB-IG and ReAct. Its differences from Passive NB, RAG-based, CoT, and OpenRCA are not statistically significant on this benchmark. The advantage is substantially stronger on E-TEP. EviDiag reaches 76.00% R@1, 83.00% MRR@3, and 75.58% Macro-F1, improving R@1 by 36–60 percentage points over the baselines. All comparisons on these three core metrics remain significant after Holm correction. The R@3 improvement is also significant against every baseline except OpenRCA, while the R@5 improvement remains significant against RAG-based and CoT.

Analysis. The significance pattern shows that EviDiag's main advantage lies in producing an accurate and well-ordered final diagnosis. The consistent gains in R@1 and MRR@3 indicate that posterior-guided inquiry moves the correct fault toward the top of the ranking rather than merely retaining it in a broad candidate set. The simultaneous Macro-F1 gains further show that this improvement extends across fault classes instead of being driven by a small subset of easy cases. This advantage is especially clear on E-TEP, where the large effect sizes remain significant against every baseline under the strict family-wise correction. RCAEval contains only 60 episodes, and the R@1 difference between EviDiag and CoT corresponds to two additional correct diagnoses, which limits statistical power despite EviDiag obtaining the best numerical R@1. The weaker separation at R@5 is also expected because a five-label prediction already covers most RCAEval classes and half of the E-TEP label space. Overall, the results provide the strongest support for EviDiag at the most decision-relevant ranking positions, particularly on the more challenging E-TEP benchmark.

Table 5: Paired significance analysis on RCAEval ($N = 60$). Each metric header gives the EviDiag score. Within a cell, the three lines report the baseline score, Δ with its class-stratified paired-bootstrap 95% confidence interval, and raw p/Holm-adjusted q. A superscript $*$ marks $q < 0.05$. Scores, differences, and interval endpoints are percentages or percentage points.

Baseline	R@1 EviDiag 58.33	R@3 EviDiag 81.67	R@5 EviDiag 88.33	MRR@3 EviDiag 70.00	Macro-F1 EviDiag 60.40
	50.00 +8.33 [−3.33, 20.00]	86.67 −5.00 [−15.00, 5.00]	95.00 −6.67 [−13.33, 0.00]	65.28 +4.72 [−5.00, 13.89]	52.29 +8.10 [−4.43, 21.30]
Passive NB	.3323/1.000	.5811/1.000	.2188/1.000	.4265/1.000	.2395/1.000
	33.33 +25.00 [11.67, 38.33]	78.33 +3.33 [−8.33, 15.00]	93.33 −5.00 [−13.33, 3.33]	51.94 +18.06 [7.78, 28.33]	29.51 +30.89 [17.39, 43.74]
Active NB-IG	.00408/.1101	.8145/1.000	.4531/1.000	.00794/.1985	.000450/.0131*
	53.33 +5.00 [−10.00, 20.00]	86.67 −5.00 [−16.67, 5.00]	96.67 −8.33 [−16.67, 0.00]	67.50 +2.50 [−9.44, 14.17]	53.87 +6.52 [−9.14, 22.40]
RAG-based	.6900/1.000	.5811/1.000	.1250/1.000	.7313/1.000	.4329/1.000
	55.00 +3.33 [−10.00, 16.67]	86.67 −5.00 [−15.00, 5.00]	93.33 −5.00 [−13.33, 3.33]	68.61 +1.39 [−9.17, 11.94]	54.71 +5.69 [−8.55, 19.63]
CoT	.8318/1.000	.5488/1.000	.5078/1.000	.8497/1.000	.4731/1.000
	33.33 +25.00 [13.33, 36.67]	65.00 +16.67 [3.33, 30.00]	90.00 −1.67 [−13.33, 8.33]	46.94 +23.06 [11.67, 34.44]	27.34 +33.05 [20.66, 45.14]
ReAct	.00592/.1540	.0525/1.000	1.000/1.000	.00239/.0670	.000340/.0102*
	23.33 +35.00 [21.67, 48.33]	40.00 +41.67 [30.00, 51.67]	41.67 +46.67 [35.00, 58.33]	31.11 +38.89 [28.61, 48.89]	23.11 +37.28 [23.27, 50.75]
FaultExplainer	.000324/.0100*	1.09e−5/.000361*	1.94e−6/(6.80e−5)*	6.77e−6/.000230*	.000280/.00896*
	45.00 +13.33 [−1.67, 28.33]	81.67 0.00 [−8.33, 8.33]	96.67 −8.33 [−15.00, −1.67]	60.83 +9.17 [−1.39, 19.17]	44.79 +15.60 [0.19, 31.01]
OpenRCA	.1686/1.000	1.000/1.000	.0625/1.000	.1232/1.000	.0676/1.000

Table 6: Paired significance analysis on E-TEP ($N = 50$), using the same notation and correction family as Table 5. EviDiag's R@1, MRR@3, and Macro-F1 improvements remain significant against every baseline after Holm correction.

Baseline	R@1 EviDiag 76.00	R@3 EviDiag 90.00	R@5 EviDiag 92.00	MRR@3 EviDiag 83.00	Macro-F1 EviDiag 75.58
	22.00 +54.00 [44.00, 64.00]	60.00 +30.00 [22.00, 38.00]	74.00 +18.00 [12.00, 24.00]	39.00 +44.00 [36.00, 51.33]	13.56 +62.03 [49.49, 73.39]
Passive NB	1.49e−8/(4.47e−7)*	.000729/.00656*	.0225/.1064	2.01e−9/(6.63e−8)*	5.00e−6/.000110*
	20.00 +56.00 [44.00, 66.00]	60.00 +30.00 [22.00, 36.00]	72.00 +20.00 [12.00, 26.00]	37.00 +46.00 [38.00, 53.67]	16.85 +58.74 [46.07, 70.95]
Active NB-IG	7.45e−9/(2.38e−7)*	.000729/.00656*	.0129/.0776	1.54e−9/(5.24e−8)*	5.00e−6/.000110*
	30.00 +46.00 [36.00, 56.00]	38.00 +52.00 [44.00, 60.00]	56.00 +36.00 [26.00, 44.00]	33.33 +49.67 [42.00, 57.33]	26.11 +49.47 [37.87, 60.17]
RAG-based	1.55e−6/(3.72e−5)*	2.16e−7/(5.40e−6)*	.000277/.00333*	3.21e−8/(8.68e−7)*	5.00e−6/.000110*
	16.00 +60.00 [48.00, 72.00]	30.00 +60.00 [50.00, 70.00]	50.00 +42.00 [30.00, 54.00]	21.33 +61.67 [51.67, 71.67]	14.04 +61.54 [49.42, 73.46]
CoT	1.54e−8/(4.47e−7)*	1.54e−8/(4.47e−7)*	1.94e−5/.000330*	2.76e−10/(9.68e−9)*	5.00e−6/.000110*
	18.00 +58.00 [46.00, 70.00]	40.00 +50.00 [40.00, 60.00]	72.00 +20.00 [8.00, 32.00]	27.00 +56.00 [46.67, 64.67]	8.64 +66.94 [53.61, 77.76]
ReAct	1.31e−7/(3.40e−6)*	4.65e−6/.000107*	.0213/.1064	1.23e−8/(3.81e−7)*	5.00e−6/.000110*
	40.00 +36.00 [22.00, 50.00]	70.00 +20.00 [10.00, 30.00]	82.00 +10.00 [2.00, 18.00]	51.67 +31.33 [20.00, 42.00]	37.06 +38.52 [23.29, 52.50]
FaultExplainer	.000277/.00333*	.00635/.0444*	.1250/.2188	2.89e−5/.000463*	.000355/.00355*
	40.00 +36.00 [22.00, 50.00]	70.00 +20.00 [8.00, 32.00]	80.00 +12.00 [2.00, 22.00]	52.33 +30.67 [18.33, 42.33]	39.23 +36.36 [21.95, 51.83]
OpenRCA	.000121/.00157*	.0213/.1064	.1094/.2188	5.68e−5/.000795*	5.00e−5/.000750*

E.6.2 Robustness to the LLM Planner

We examine EviDiag's sensitivity to the choice of LLM planner using DeepSeek-V4.1-Flash, Qwen3.8-27B, and GLM-5.3. Table 7 summarizes their diagnostic performance within EviDiag's evidence-seeking framework.

On RCAEval, all three planners yield similar diagnostic performance. R@1 ranges from 55.00% to 58.33%, a spread corresponding to two of the 60 test cases. R@3 ranges from 80.00% to 83.33%, while R@5 remains between 88.33% and 90.00%. MRR and Macro-F1 vary by 3.06 and 3.63

Table 7: EviDiag results with different LLM planners.

Dataset	LLM planner	R@1	R@3	R@5	MRR	Macro-F1
RCAEval	DeepSeek-V4.1-Flash	58.33	81.67	88.33	70.00	60.40
	Qwen3.8-27B	55.00	80.00	88.33	67.22	57.28
	GLM-5.3	58.33	83.33	90.00	70.28	60.91
E-TEP	DeepSeek-V4.1-Flash	76.00	90.00	92.00	83.00	75.58
	Qwen3.8-27B	82.00	90.00	92.00	85.67	82.14
	GLM-5.3	78.00	90.00	92.00	83.67	78.22

percentage points, respectively. Thus, changing the planner preserves both fault-ranking quality and class-balanced identification performance on this benchmark.

On E-TEP, the completed DeepSeek and Qwen runs both obtain 90.00% R@3 and 92.00% R@5. Qwen improves R@1 from 76.00% to 82.00% and Macro-F1 from 75.58% to 82.14%, with MRR increasing from 83.00% to 85.67%. The planner therefore affects which fault is ultimately prioritized, while broader fault coverage remains stable. The GLM-5.3 results on E-TEP are pending and are excluded from this comparison.

These results support robustness to planner choice in the evaluated settings, with consistency most evident across the three RCAEval planners and in the higher-rank E-TEP recalls. This pattern is consistent with EviDiag's separation of examination selection from belief updating: the LLM selects among feasible examinations using posterior-based information values and retrieved experience, whereas tool-generated outcomes and evidence likelihoods determine the numerical update. Different planners can choose different examination sequences, but each realized belief transition follows the same explicit update rule. This shared structure provides a basis for stable diagnostic performance while retaining flexibility in action selection; the observed consistency concerns aggregate performance rather than identical case-level decisions.

E.6.3 Auditability Analysis

EviDiag makes diagnostic inquiry auditable by recording the examination selected at each round, its tool-generated outcome, and the corresponding posterior before and after the update. These records connect changes in diagnostic performance to an inspectable sequence of evidence acquisition and belief revision.

Figure 6 shows how diagnosis improves as inquiry proceeds on the common 60-case RCAEval OB/SS test set and all 50 E-TEP test cases. Round 0 denotes the posterior after direction initialization. Across this interval, RCAEval accuracy increases from 40.00% to 60.00%, while E-TEP accuracy increases from 56.00% to 76.00%. Macro-F1 also increases, from 39.33% to 62.76% on RCAEval and from 51.91% to 75.58% on E-TEP. Thus, the additional examinations yield measurable gains in fault identification over initialization. The curves retain the observed intermediate fluctuations: the improvement is an overall trend, rather than a guaranteed increase at every round.

Figure 7 illustrates how this audit trail exposes the formation and revision of a diagnosis through two complete cases. In RCAEval, a *Miss* on the CPU *Higher* examination reduces the CPU posterior from 22.3% to 9.5%. Subsequent *Rise Ramp* hits on P90 and P50 latency increase the Delay posterior from 40.5% to 61.1% and then 82.7%. The E-TEP case reveals a different trajectory: the first oscillation *Hit* on XMEAS_9 strengthens IDV11 to 83.9%, but the subsequent *Miss* outcomes on XMV_6 and XMEAS_10 shift support toward IDV14, whose posterior rises from 9.3% to 37.3% and then 74.4%. The trace therefore exposes both the accumulation of supporting evidence and the correction of an initially favored hypothesis. A *Miss* is informative when the tested pattern is more strongly expected under a competing fault.

This correspondence between an examination, its observed outcome, and its posterior effect provides a concrete basis for auditing. An examiner can identify the variable and pattern tested at a disputed round, inspect the recorded tool output, and recompute the belief transition from the preceding posterior and the evidence likelihoods using Eq. 4. A questionable increase in confidence can thus

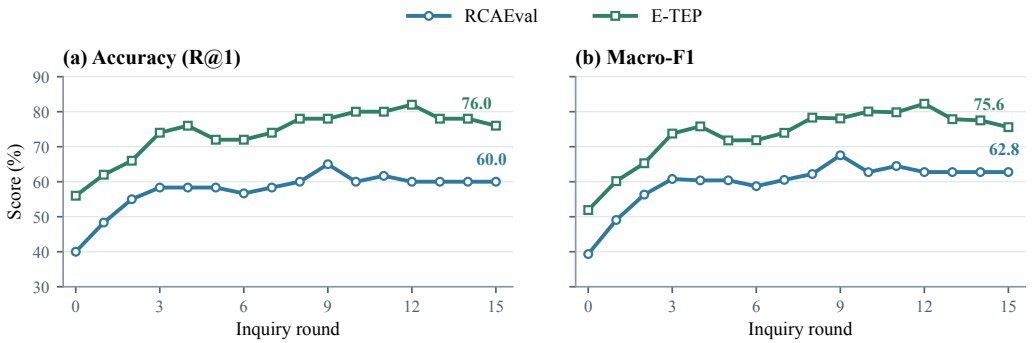


Figure 6: Diagnostic performance across inquiry rounds on RCAEval and E-TEP: (a) Accuracy (R@1) and (b) Macro-F1. The E-TEP curve uses the same 15-round full-method run as the result tables.

be investigated at the point where it arose: in the choice of examination, in the extracted outcome, or in the likelihood assigned to that outcome. Retaining the intermediate states also makes it possible to determine whether later evidence reinforced or reversed that decision, rather than judging the diagnostic process solely from its final label.

Retrieved reflection has a distinct role in this record. It supplies contextual guidance about which examination may distinguish the remaining hypotheses; the numerical update is carried out only after the selected tool returns an outcome. Conditional on the same prior state, examination, outcome, and fixed likelihood table, the posterior transition is unchanged by the wording of the retrieved reflection or the LLM's explanation. Reflection can consequently influence the diagnostic trajectory through examination selection, while each realized belief change remains attributable to explicit numerical evidence. Recording these two roles separately makes the language-guided search for evidence and the evidence-based revision of belief independently inspectable.

Failure-case audit. Auditability also makes an incorrect diagnosis amenable to inspection. We examine the RCAEval case `re2ss_catalogue_socket_3`, whose ground-truth fault is Socket. The archived full-method run uses the LLM planner, retrieved reflection, and calibrated evidence likelihoods with a maximum budget of 15 inquiry rounds. After three direction-initialization checks, the agent executes 13 LLM-selected examinations without fallback and stops because no actions remain. It predicts Delay with posterior 66.63%, while Socket retains 10.83%. Figure 8 reconstructs the posterior trajectory and the two Socket examination statistics from the saved numerical records; Table 8 lists every initialization check and inquiry update.

Where the diagnosis diverges. The first loss of Socket-specific support occurs at Q2: the Socket/*Higher* examination returns *Miss*, reducing $P(\text{Socket})$ from 31.25% to 18.88%. Socket recovers to the highest posterior at Q5, but its *Rise Step* examination at Q6 also returns *Miss*, reducing its probability from 36.50% to 24.72%. This is not an irreversible cascade: Socket becomes the leading hypothesis again at Q9. The final switch to Delay occurs at Q10, when a P50-latency *Rise Step* hit increases $P(\text{Delay})$ from 22.33% to 44.79%. The subsequent P90-latency ramp and step hits raise it to 66.49%, while Socket falls to 10.81%. The trace identifies a specific combination of missed Socket evidence and accumulating latency evidence behind the wrong final label.

From a belief change to its numerical cause. The saved tool statistics explain why both Socket checks returned *Miss*. For *Higher*, the maximum statistic is 0.1210, below the operative threshold of 0.2159. For *Rise Step*, the maximum is 0.1037: it exceeds the empirical threshold of 0.0980 but remains below the recovery-error dead zone of 0.3250, which sets the operative threshold. Thus, the visible Socket deviation does not become a positive diagnostic outcome under the recorded criteria. At Q2, the calibrated Miss likelihood is 0.375 for Socket and 0.875 for CPU, so this outcome decreases the Socket-to-CPU posterior odds. Replaying all 16 executed examinations reproduces their Hit/Miss results, and recomputing their Bayesian transitions with the saved calibrated likelihoods reproduces every recorded posterior to numerical precision. The adverse belief changes therefore



Figure 7: Complete diagnostic inquiry traces for additional cases from (a) RCAEval and (b) E-TEP. The observed trajectories and anchored healthy references are shown above the inquiry rounds, which record the LLM decision, retrieved reflection, executed diagnostic tool, tool output, and Bayesian posterior update.

follow the implemented evidence rules; the trace localizes the issue to evidence sensitivity and its diagnostic interpretation, without establishing a numerical-update error.

Actionable implications. This case motivates checking, on validation cases, whether the Socket-specific dead zones and examination windows suppress diagnostically useful deviations, and whether the calibrated Miss likelihoods adequately represent that loss of sensitivity. Any revised threshold must be evaluated jointly for recovered detections and false positives. The later P50/P90 ramp and step hits also motivate testing dependence-aware aggregation of overlapping latency evidence. Because this run exhausts the action set, simply reordering the same examinations would leave the terminal posterior unchanged under the fixed initialized hypothesis set, likelihoods, and outcomes. Improvements therefore require examining the evidence extraction or update assumptions, rather than relying on a different verbal explanation. These are testable modifications suggested by the audited failure, not improvements demonstrated by this trace. Retrieved reflection

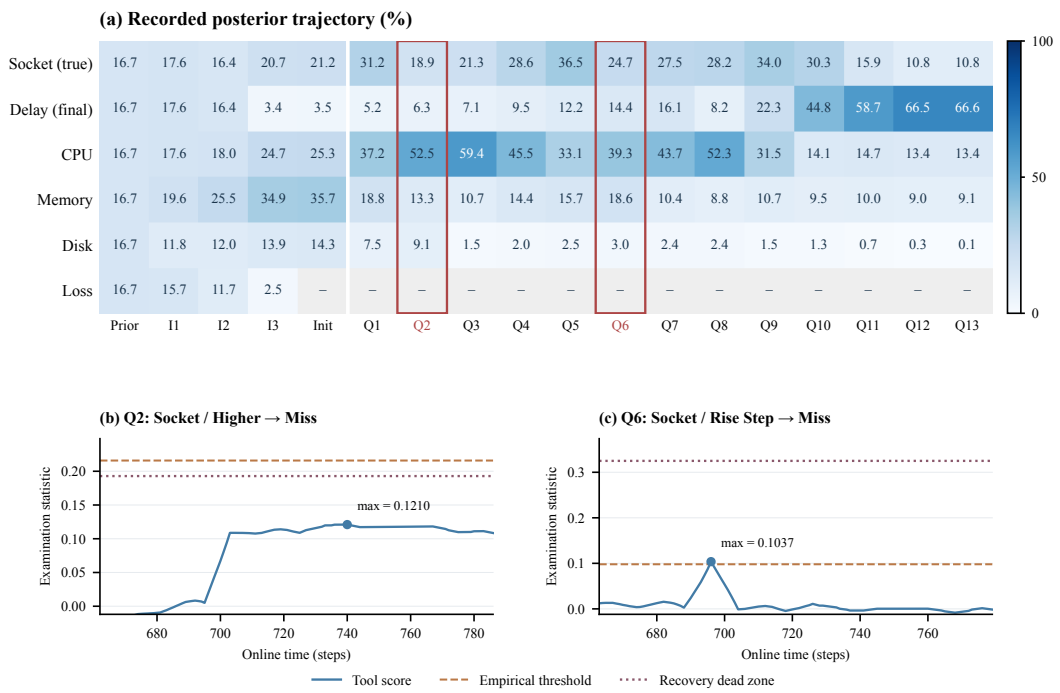


Figure 8: Audit of a failed RCAEval Socket diagnosis. (a) Recorded posterior probabilities; outlined columns mark the Socket misses at Q2 and Q6. I1–I3 denote initialization checks, and Init denotes top-five hypothesis retention and renormalization; dashes mark the excluded Loss hypothesis. (b,c) Socket examination statistics and their empirical and recovery-error thresholds. The operative threshold is the larger of the two bounds.

Table 8: Complete recorded trace for the failed RCAEval Socket case. Probabilities are percentages before and after each update; MAP is the highest-posterior hypothesis after the update. I1–I3 are initialization checks; Init retains and renormalizes the top five hypotheses. Bold stage labels mark the two Socket misses and the later Delay-dominant updates.

Stage	Examination	Output	$P(\text{Socket})$	$P(\text{Delay})$	MAP after
Prior	Uniform prior	–	16.67	16.67	Tie (16.67)
I1	P50 latency / Higher	Hit	16.67 → 17.65	16.67 → 17.65	Memory (19.61)
I2	P90 latency / Higher	Hit	17.65 → 16.39	17.65 → 16.39	Memory (25.50)
I3	CPU / Higher	Hit	16.39 → 20.69	16.39 → 3.45	Memory (34.86)
Init	Retain top five; normalize	–	20.69 → 21.21	3.45 → 3.53	Memory (35.74)
Q1	Memory / Higher	Miss	21.21 → 31.25	3.53 → 5.21	CPU (37.24)
Q2	Socket / Higher	Miss	31.25 → 18.88	5.21 → 6.29	CPU (52.50)
Q3	Disk I/O / Higher	Miss	18.88 → 21.35	6.29 → 7.12	CPU (59.36)
Q4	CPU / Rise Ramp	Miss	21.35 → 28.63	7.12 → 9.54	CPU (45.50)
Q5	CPU / Rise Step	Miss	28.63 → 36.50	9.54 → 12.17	Socket (36.50)
Q6	Socket / Rise Step	Miss	36.50 → 24.72	12.17 → 14.42	CPU (39.28)
Q7	Memory / Rise Step	Miss	24.72 → 27.52	14.42 → 16.05	CPU (43.72)
Q8	Workload / Lower	Miss	27.52 → 28.21	16.05 → 8.23	CPU (52.30)
Q9	P50 latency / Rise Ramp	Hit	28.21 → 34.02	8.23 → 22.33	Socket (34.02)
Q10	P50 latency / Rise Step	Hit	34.02 → 30.34	22.33 → 44.79	Delay (44.79)
Q11	P90 latency / Rise Ramp	Hit	30.34 → 15.90	44.79 → 58.71	Delay (58.71)
Q12	P90 latency / Rise Step	Hit	15.90 → 10.81	58.71 → 66.49	Delay (66.49)
Q13	Disk I/O / Rise Ramp	Miss	10.81 → 10.83	66.49 → 66.63	Delay (66.63)

remains visible as guidance for selecting checks, while the associated tool statistics, thresholds, and likelihoods identify the numerical basis of each realized belief change.

Table 9: Posterior calibration on RCAEval and E-TEP. N_p/N is posterior coverage among test episodes. ECE uses five equal-frequency bins and reports an episode-bootstrap 95% confidence interval from 10,000 resamples. Brier is the multiclass score summed over classes. Lower values are better, and bold marks the best point estimate within each dataset.

Dataset	Method	N_p/N	$\mathrm{ECE}_5 \downarrow$	95% CI	Brier $\downarrow$
RCAEval	Passive NB	58/60	**0.100**	[0.070, 0.251]	0.649
RCAEval	Active NB-IG	58/60	0.167	[0.102, 0.254]	0.705
RCAEval	EviDiag	58/60	0.211	[0.102, 0.309]	**0.614**
E-TEP	Passive NB	50/50	0.540	[0.429, 0.649]	1.138
E-TEP	Active NB-IG	50/50	0.172	[0.137, 0.290]	0.852
E-TEP	EviDiag	50/50	**0.126**	[0.049, 0.223]	**0.346**

E.6.4 Posterior calibration

We evaluate calibration for the three methods whose saved outputs contain numerical final posteriors: Passive NB, Active NB-IG, and EviDiag. Methods that return only a label ranking are not assigned synthetic probabilities. For EviDiag, the posterior is embedded into the full dataset label space by assigning probability zero to hypotheses removed by hard candidate pruning. The retained probabilities already sum to one. RCAEval contains two missed-alarm episodes without a diagnostic posterior, so calibration is evaluated on the 58 episodes with posterior outputs and posterior coverage is reported explicitly. All three methods provide posteriors for the 50 E-TEP episodes.

We report the five-bin adaptive expected calibration error (ECE) and the multiclass Brier score. For the N_p test episodes with posterior outputs, let $p_i(f)$ denote the predicted probability of fault f, y_i the ground-truth fault, $c_i = \max_f p_i(f)$ the confidence, and $a_i = \mathbf{1}[\arg\max_f p_i(f) = y_i]$ the top-label correctness. We partition the episodes into five equal-frequency confidence bins $\{\mathcal{B}_m\}_{m=1}^{5}$ and compute

$$\mathrm{ECE}_5 = \sum_{m=1}^{5} \frac{|\mathcal{B}_m|}{N_p} \left| \frac{1}{|\mathcal{B}_m|} \sum_{i \in \mathcal{B}_m} a_i - \frac{1}{|\mathcal{B}_m|} \sum_{i \in \mathcal{B}_m} c_i \right|. \tag{58}$$

The multiclass Brier score is

$$\mathrm{Brier} = \frac{1}{N_p} \sum_{i=1}^{N_p} \sum_{f \in \mathcal{F}} \left(p_i(f) - \mathbf{1}[y_i = f] \right)^2. \tag{59}$$

Both metrics are lower-is-better. The ECE confidence interval is the percentile interval from 10,000 episode-level bootstrap samples. Each bootstrap replicate resamples eligible episodes with replacement and reconstructs the five equal-frequency bins. Brier scores use the dense posterior over all fault classes, including zero probability for pruned hypotheses. Both metrics are lower-is-better.

Results. EviDiag attains the lowest Brier score on both benchmarks, reducing the observed squared probabilistic error from 0.649 to 0.614 relative to the stronger NB result on RCAEval and from 0.852 to 0.346 on E-TEP. Its top-label calibration is dataset-dependent. Passive NB has the lowest RCAEval ECE point estimate, whereas EviDiag has the lowest E-TEP estimate. The bootstrap intervals overlap on RCAEval and between EviDiag and Active NB-IG on E-TEP, so the ECE results do not establish a uniform separation between these methods. Together, the two metrics show that EviDiag produces the most accurate probability vectors under the Brier score, while its agreement between top-label confidence and empirical accuracy is strongest on E-TEP and remains less well aligned on RCAEval.