DynaAudit: Auditing Hidden Supervision-Channel Drift

Anonymous authors
Paper under double-blind review

Abstract

Weak-supervision systems are often monitored through features and source outputs, yet the mapping from those observables to the latent label can change after deployment. We study *hidden supervision-channel drift*, where distinct supervision mechanisms induce the same declared unlabeled monitor transcript while implying different gold-label posteriors. This creates a fundamental blind spot: under exact transcript equivalence, purely unlabeled detection has power no greater than its false-alarm rate, and unlabeled estimation incurs irreducible error. We characterize *support auditability*—whether trusted labels available on the deployment support can distinguish candidate mechanisms—and, for a finite correctly specified candidate family, derive an anytime finite-sample certificate for fully adaptive auditing based on accumulated pairwise Bhattacharyya information. The guarantee is acquisition-policy agnostic, separating statistical validity from acquisition speed. These results motivate DynaAudit, which separates representative sentinel detection, mechanism-aware diagnosis, model criticism and abstention, and inclusion-aware reuse of adaptively selected labels. Controlled and temporal experiments show that targeted audits concentrate diagnostic information when post-alarm labels are scarce, while matched-budget random auditing can catch up as total trusted-label budgets grow. In natural temporal data, adaptation improves prediction over a frozen model, while the targeted policy shows no established advantage over random auditing and most mechanism diagnoses are inconclusive. Overall, DynaAudit characterizes when unlabeled monitoring is fundamentally insufficient and what additional trusted evidence is required for auditable adaptation.

1 Introduction

Weak supervision replaces expensive labels with imperfect signals such as heuristics, crowds, pretrained models, and automated evaluators (Ratner et al., 2017; Bach et al., 2017). Deployment changes can alter source reliability, class prevalence, coverage, or error dependence. Throughout, a *mechanism* means a statistical supervision-channel hypothesis, not necessarily a causal root cause. Monitoring source outputs is useful when such changes alter the observed stream, but it leaves a more basic question unanswered: *does the same source configuration still mean the same thing about the gold label?*

The answer can be no even with unlimited unlabeled data. All hidden-drift impossibility statements are relative to the declared monitor information set; adding monitored variables can change the equivalence class. With two binary sources and a balanced prior, let World A have conditionally independent sources of accuracy 0.8. Let World B have source accuracy 0.7 with positive error dependence. Both induce

$$(P(++), P(--), P(+-), P(-+)) = (.34, .34, .16, .16),$$

yet $P_A(Y{=}+1 \mid ++) = 16/17 \approx .941$ and $P_B(Y{=}+1 \mid ++) = 27/34 \approx .794$. Moreover, disagreement states have posterior $1/2$ in both worlds: they maximize predictive label entropy while providing zero mechanism information. Predictive uncertainty and mechanism uncertainty are therefore different objects.

Recent drift-aware weak supervision asks how to re-estimate changing source accuracies from historical data (Mazzetto et al., 2025), while recent work also uses the broader term supervision

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

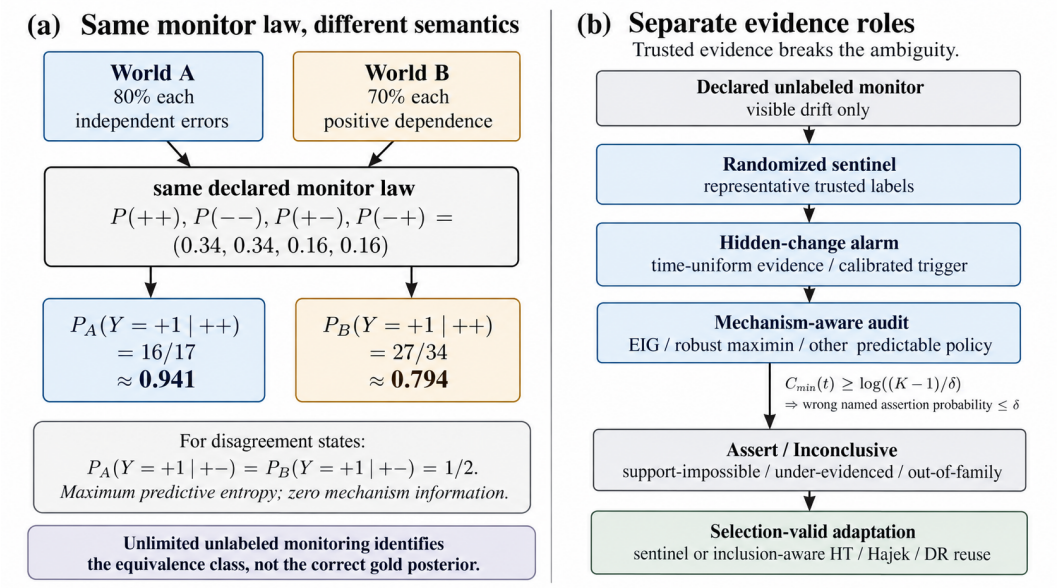


Figure 1: **Hidden supervision drift and DynaAudit.** Distinct supervision mechanisms can have the same declared monitor law but different gold semantics. DynaAudit therefore separates representative hidden-change evidence, targeted mechanism auditing, explicit inconclusive outcomes, and population-valid reuse.

drift for changing weak-label relationships under distribution shift (Shoeibi et al., 2026). We ask an earlier identifiability question: can reliability, dependence, or prior changes be distinguished from the declared monitor at all?

Trusted labels can break monitor equivalence, but their roles differ. Representative labels are suited to detecting hidden change and estimating population quantities; post-alarm labels can instead target states that discriminate candidate mechanisms. Those targeted labels are selected data and can bias population adaptation if reused naively. DynaAudit therefore separates *detection*, *diagnosis*, *model criticism*, and *adaptation* (Fig. 1).

Contributions. **(1) Hidden-drift identifiability:** we formalize equivalence relative to the complete declared monitor transcript and give unlabeled estimation and detection limits. **(2) Adaptive auditability:** we characterize when trusted labels can separate mechanisms on feasible support and derive an anytime certificate for fully adaptive auditing, plus an information lower bound. **(3) Auditable adaptation:** we instantiate the theory with randomized sentinels, policy-agnostic mechanism acquisition, explicit rejection, and selection-valid reuse; controlled, real-label temporal, and natural studies establish both capability and failure boundaries.

2 Hidden Drift and Support Auditability

Let $O_t = \phi(X_t, \Lambda_t)$ be the complete state exposed to the unlabeled monitor, and let Z denote the full monitored transcript over the horizon of interest. Under mechanism θ, write Q_θ for the law of Z. Mechanisms are *monitor-equivalent* when $Q_{\theta_0} = Q_{\theta_1}$; hidden drift occurs when they are monitor-equivalent but disagree on a gold-relevant target $T(\theta)$.

Theorem 1 (Monitor-relative information barrier)**.** *If $\Delta = d(T(\theta_0), T(\theta_1))$, then every estimator based only on Z obeys*

$$\inf_{\widehat{T}} \max_{i \in \{0,1\}} \mathbb{E}_{\theta_i} d(\widehat{T}, T(\theta_i)) \geq \frac{\Delta}{2}\{1 - \mathrm{TV}(Q_{\theta_0}, Q_{\theta_1})\}. \tag{1}$$

If an unlabeled alarm event E has $Q_{\theta_0}(E) \leq \alpha$, then $Q_{\theta_1}(E) \leq \alpha + \mathrm{TV}(Q_{\theta_0}, Q_{\theta_1})$.

Table 1: Candidate supervision-change blocks used to make diagnosis interpretable. The theory applies to any finite family of predictive channels.

Hypothesis	Changed block	Interpretation
H_0	none	historical channel remains adequate
H_π	prior/intercept	prevalence or stratum-prior change
H_A	source main effects	reliability/confusion change
H_D	source interactions	conditional error-dependence change
H_{AD}	main + interactions	joint reliability/dependence change
H_U	outside family	absolute-fit rejection / misspecification

The result is a two-point argument: integrate the triangle inequality against the overlap of Q_{θ_0} and Q_{θ_1}; the detection claim is the variational definition of total variation (proofs in Appendix A). Exact transcript equivalence therefore makes unlabeled detection power no larger than the false-alarm probability. In Fig. 1, $\Delta = 5/34$ on the $++$ posterior, giving minimax absolute error at least $5/68$.

Trusted labels help only where their conditional laws differ. For candidate mechanisms h, k at auditable context o, define

$$b_{h,k}(o) = -\log \sum_y \sqrt{p_h(y \mid o) p_k(y \mid o)}.$$

For feasible audit allocations $q \in \mathcal{Q}$, let

$$\Gamma_B(\mathcal{Q}) = \sup_{q \in \mathcal{Q}} \min_{h \neq k} \sum_o q(o) b_{h,k}(o). \tag{2}$$

With finite candidates and full-simplex repeated-draw designs, $\Gamma_B > 0$ iff every pair differs on at least one supported context; finite-pool capacities are encoded in $\mathcal{Q}$. Under the stationary item-wise audit model used by our exact-hidden benchmarks, if a monitor-equivalent pair also satisfies $p_h(\cdot \mid o) = p_k(\cdot \mid o)$ for every admissible context, no adaptive audit can distinguish it; the appendix states the stronger conditional-process assumption needed beyond the i.i.d. case. We call monitor-equivalent pairs with positive supported separation *hidden but auditable*.

3 DynaAudit and adaptive recovery

DynaAudit operationalizes the distinction above. A low-rate randomized *sentinel* stream supplies representative trusted evidence even when the unlabeled monitor is quiet. After an alarm, a *mechanism audit* targets supported distinctions among a prespecified candidate family. A *critic* may return inconclusive when the family is support-impossible, under-evidenced, or fails absolute fit. Finally, representative sentinels can be used directly for adaptation, whereas targeted audits are reused only with design-aware inclusion probabilities. The architecture is acquisition-policy agnostic. It also distinguishes three reasons not to name a mechanism: *support-impossible* (some pair cannot be separated on feasible support), *under-evidenced* (the family is separable but the evidence threshold has not been reached), and *out-of-family* (absolute fit is inadequate). These are statistically different failure modes and should not be collapsed into forced classification.

The central certificate allows fully label-adaptive querying. Immediately before audit label Y_t is revealed, let $\mathcal{G}_t$ contain the entire history and the chosen audit context. Candidate h specifies predictive law $p_{h,t}(\cdot \mid \mathcal{G}_t)$. For $h \neq k$, define

$$L_t^{k/h} = \prod_{s \leq t} \frac{p_{k,s}(Y_s \mid \mathcal{G}_s)}{p_{h,s}(Y_s \mid \mathcal{G}_s)}, \quad C_{h,k}(t) = \sum_{s \leq t} b_{h,k,s}. \tag{3}$$

Theorem 2 (Anytime adaptive audit certificate)**.** *Assume the true candidate h is fixed during the audit and its predictive channel is correct with common positive support. Then*

$$M_t^{k/h} = \sqrt{L_t^{k/h}} \exp\{C_{h,k}(t)\}$$

is a nonnegative martingale under h. Hence

$$\Pr_h\Big(\exists t : L_t^{k/h} \geq 1,\ C_{h,k}(t) \geq c\Big) \leq e^{-c}. \tag{4}$$

For K equally weighted candidates, let $\widehat{H}_t$ be the conditional-likelihood leader and stop when

$$\min_{j \neq \widehat{H}_t} C_{\widehat{H}_t, j}(t) \geq \log \frac{K-1}{\delta}. \tag{5}$$

The probability of ever stopping with an incorrect named mechanism is at most δ.

The key identity is $\mathbb{E}_h[\sqrt{p_k(Y_t \mid \mathcal{G}_t)/p_h(Y_t \mid \mathcal{G}_t)} \mid \mathcal{G}_t] = e^{-b_{h,k,t}}$; Ville's inequality and a union bound give the result. Thus the policy affects *how fast* information is accumulated, not validity of the certificate. At a fixed budget B,

$$\Pr_h(\widehat{H}_B \neq h) \leq (K-1)e^{-c} + \Pr_h\{C_{\min,h}(B) < c\}, \tag{6}$$

separating inference error after sufficient evidence from policy-dependent evidence shortfall. Conversely, any uniformly δ-correct auditor that terminates with a named candidate satisfies

$$\mathbb{E}_h \tau \geq \mathrm{kl}(1-\delta, \delta)/C_h^{\mathrm{KL},\star},$$

where $C_h^{\mathrm{KL},\star}$ is the best feasible worst-alternative directed-KL rate. For Bernoulli channels clipped to $[\eta, 1-\eta]$, the Bhattacharyya and directed-KL rates are constant-equivalent: pointwise $2B(P,Q) \leq D_{\mathrm{KL}}(P\|Q) \leq 2B(P,Q)/[\eta(1-\eta)]$. Thus the sufficient certificate and necessary information lower bound measure the same supported separation up to clipping-dependent constants. Full assumptions, prior-weight extensions, and a certified estimated-channel extension are in Appendix A.

We study two acquisition principles. With current posterior over candidates, EIG scores a context by

$$\mathrm{EIG}_t(o) = I(H; Y \mid o, \mathcal{F}_{t-1}), \qquad q_t(o) \propto N_t(o)\,\mathrm{EIG}_t(o), \tag{7}$$

where $N_t(o)$ is the number of currently available unlabeled deployment items whose declared audit state equals o. Robust maximin acquisition instead chooses a feasible design that protects the weakest plausible pair,

$$q_t^{\mathrm{rob}} \in \arg\max_{q \in \mathcal{Q}_t} \min_{h \neq k \in \mathcal{C}_t} \sum_o q(o) J_{h,k}(o), \tag{8}$$

with symmetric KL $J_{h,k}$ and credible set $\mathcal{C}_t$; the prespecified implementations use either a support-constrained LP or a randomized soft two-candidate variant as specified in the appendix. We do not claim either acquisition objective is universally optimal. Equation (6) instead suggests comparing how reliably policies accumulate $C_{\min}$.

4 Experiments

Design. We evaluate exact-hidden and near-hidden accuracy-versus-dependence drift in two-source and genuine three/five-source synthetic channels; a real-label controlled study with five out-of-fold weak learners on the breast-cancer benchmark; randomized-sentinel and selection-correction studies; three WRENCH weak-label pools with controlled conditional-label changes (Zhang et al., 2021); four future episodes of the UCI Occupancy time series (Candanedo & Feldheim, 2016); and natural temporal stress on the UCI Gas Sensor Array Drift data (Vergara et al., 2012; Fonollosa et al., 2015). Exact controlled pairs share the declared $P(\Lambda)$; near-hidden pairs have one-step TV 0.015. The primary two-source exact-hidden scope contains 12,000 world-runs; 2,000 multi-source exact and 1,500 near-hidden seeds; 1,000 real-label-controlled seeds; and four temporal episodes with 500 repetitions per world. Policy comparisons use shared worlds/seeds and paired bootstrap inference; temporal claims use episode-aware or cluster-aware resampling as appropriate. Full protocols and provenance are in Appendix E.

Acquisition terminology. We denote the robust maximin acquisition rule based on symmetric KL separation (Eq. (8)) as SR-MKL (Symmetric Robust Maximin KL).

Baseline harmonization. The original controlled EIG baseline used temperature-softened state sampling. A later independently specified temporal implementation used the parameter-free multiplicity-aware rule in Eq. (7). We therefore conduct a post-hoc baseline-strength audit on the immutable worlds, seeds, and budgets, holding Random and SR-MKL outputs fixed and rerunning only stronger EIG policies. Main-text comparisons use *multiplicity-aware EIG*; *greedy EIG + 10% exploration* is an aggressive stress comparator. This avoids preserving an implementation-sensitive historical advantage.

Table 2: Mechanism-identification accuracy (%) at each benchmark's primary audit budget. Strong EIG is competitive throughout; there is no universal acquisition winner.

Benchmark	B	Random	Mult. EIG	Greedy EIG	SR-MKL
2-source exact	10	78.11	94.95	95.22	94.70
2-source near	40	89.77	93.93	93.03	93.23
3/5-source exact	10	94.09	98.73	99.75	99.44
3/5-source near	10	92.85	97.97	98.87	98.92
Real-label controlled	10	72.40	98.10	98.85	98.20

Hidden drift is auditable; acquisition ranking is regime-dependent. Table 2 shows that mechanism-aware auditing sharply outperforms Random in the hard real-label controlled case while strong EIG and SR-MKL are nearly tied at the primary endpoints. In the multi-source exact setting, SR-MKL exceeds multiplicity-aware EIG by 0.71 points but is 0.31 points below greedy EIG + exploration (paired 95% CI $[-0.51, -0.11]$); near-hidden SR-MKL and greedy EIG differ by only 0.05 points (CI $[-0.35, 0.43]$). In the real-label controlled setting, SR-MKL minus multiplicity-aware EIG is $+0.10$ points (CI $[-0.65, 0.85]$). Replacing the legacy temperature-softened EIG with multiplicity-aware EIG improves accuracy by 5.33 points in two-source exact and 5.30 in real-label controlled, showing why acquisition comparisons must be implementation-robust.

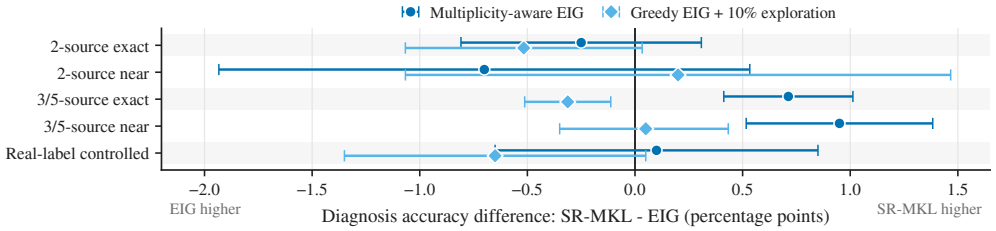


Figure 2: **Paired harmonized acquisition contrasts.** Positive values favor SR-MKL; negative values favor EIG. Circles compare multiplicity-aware EIG; diamonds compare greedy EIG + 10% exploration. Strong EIG removes the large legacy gaps and rules out a universal ranking. Error bars are paired bootstrap intervals from the baseline-strength audit.

A separate exact-hidden geometry stress test varies $K \in \{3, 5, 8\}$ and where pairwise information is concentrated. At $B = 10$, multiplicity-aware EIG has the larger macro-accuracy point estimate and larger mean $C_{\min}$ in all nine aggregate $K \times$geometry cells, while worst-candidate ordering changes across regimes (Appendix G). This supports the theoretical interpretation: policies matter through evidence accumulation, not through the validity of the certificate.

WRENCH weak-label support stress test. On IMDb, SMS, and YouTube, we preserve real WRENCH weak-label states and vary five simulated conditional gold-label channels; this is semi-synthetic drift, not an observed temporal change. At $B = 100$ diagnostic labels, greedy EIG reaches 91.4%, 71.0%, and 70.8% identification accuracy, respectively, versus SR-MKL's 86.6%, 70.6%, and 70.4%. The paired IMDb contrast (SR-MKL minus greedy EIG) is -4.8 points (95% bootstrap CI $[-7.8, -2.0]$); the other two intervals include zero. Strong EIG also changes the comparison with the original temperature-softened EIG substantially (Appendix H).

Model criticism is not cosmetic. In a prespecified out-of-family stress test, known-family mechanisms are accepted 95.03% of the time, whereas a deliberately unknown mechanism is rejected 100% of the time. Without rejection the false named-assertion rate on that unknown mechanism is 100%; with rejection it is 0%. The known correct-supported diagnosis rate is only 46.49% at this budget-20 stress point, so the result supports safe abstention rather than high open-world identification accuracy. Independent-critic power on WRENCH varies sharply by dataset and alternative (Appendix H).

5 DEPLOYMENT VALIDITY

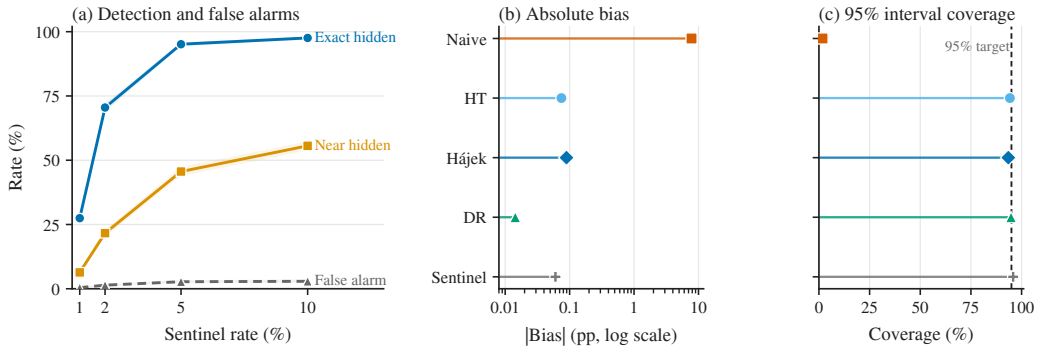


Figure 3: **Representative detection and selection-valid reuse.** (a) Detection power and false alarms versus sentinel rate, with 95% binomial intervals. (b) Absolute source-accuracy bias on a log scale. (c) Nominal 95% interval coverage; the dashed line marks 95%. Targeted agreement auditing severely biases naive population estimates, while inclusion-aware estimators restore validity.

Representative sentinels expose hidden change at a measurable cost. The controlled detector uses a likelihood-ratio mixture over audited change points. If R_j is the candidate change/no-change likelihood ratio at sentinel audit j, the prespecified evidence process is

$$E_r = \gamma^r + \sum_{s=1}^{r}(1-\gamma)\gamma^{s-1}\prod_{j=s}^{r} R_j, \qquad T = \inf\{r : E_r \geq 1/\alpha\},$$

with $\gamma = 0.98$ and $\alpha = 0.05$. Under the reference channel this is a nonnegative likelihood-ratio mixture/e-process, giving a time-uniform false-alarm interpretation.

In a 4,000-observation stream with change at $t = 1200$, the primary 5% sentinel rate detects exact-hidden abrupt drift in 95.13% of 3,000 runs (95% CI $[94.30, 95.88]$) with a 2.73% no-change false-alarm rate. Among detected runs, median delay is 857.5 arriving observations and 43 post-change sentinel labels; near-hidden power at the same rate is only 45.60%. Complete accounting matters: after adding pre-change sentinels and the fixed $B = 10$ post-alarm diagnostic budget, the median trusted-label cost through diagnosis is 114 (mean 119.13; 90th percentile 163). DynaAudit is therefore not a "ten-label system."

Targeted labels require design-aware reuse. Under deliberate agreement oversampling, naive selected-sample source-accuracy bias is $+0.07795$ and nominal 95% interval coverage collapses to 1.77%. Doubly robust correction reduces bias to $+0.00014$, RMSE to 0.03323, and restores coverage to 94.80%; a representative random-sentinel reference has lower RMSE 0.01977, as expected for pure estimation (Fig. 3). For a population target $\mu = \mathbb{E}[g(Y, O)]$, the reusable selected-label form is

$$\widehat{\mu}_{\mathrm{DR}} = \frac{1}{n}\sum_i \left[\widehat{m}_i + \frac{S_i}{\pi_i}\{g(Y_i, O_i) - \widehat{m}_i\}\right],$$

with cross-fitted outcome model $\widehat{m}_i$ and first-order inclusion probability π_i. For the prespecified with-replacement batch, first-order inclusion probability is $\pi_i = 1-(1-q_i)^B$, not the per-draw probability q_i; a 200,000-repetition audit finds maximum discrepancy 0.00157 between this formula and empirical inclusion frequency. Other sampling designs require their own valid inclusion probabilities or sequential estimators.

6 TEMPORAL EVIDENCE AND CAPABILITY BOUNDARIES

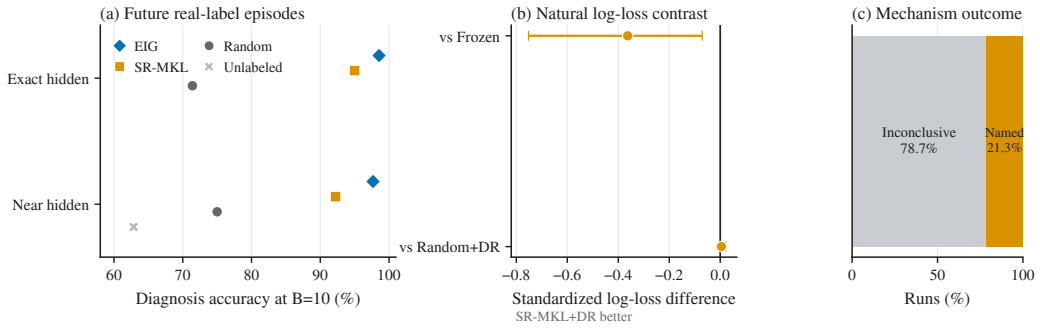


Figure 4: **Temporal capability boundary.** (a) Future real-label diagnosis at $B = 10$. (b) Natural gas crossed task–window 95% CIs for SR-MKL+DR log-loss contrasts; negative values favor SR-MKL+DR. (c) In the same natural setting, 78.7% of mechanism diagnoses are inconclusive.

Real-label temporal diagnosis. The occupancy benchmark uses 20,560 timestamped observations, freezes weak-source training before evaluation, selects four future weekdays by a prespecified rule, class-balances within difficulty strata, and controls only the weak-source failure mechanism. Exact equivalence is for declared $O = (S, \Lambda)$, not unrestricted raw X. At $B = 10$, exact-hidden diagnosis is 98.55% for EIG, 95.00% for SR-MKL (hierarchical CI $[94.25, 95.70]\%$), and 71.40% for Random. Near-hidden accuracy is 97.68% EIG, 92.23% SR-MKL (CI $[90.80, 93.63]\%$), 75.00% Random, and 62.85% unlabeled-only. The EIG advantage reinforces the policy-agnostic conclusion.

Natural temporal drift: adaptation can succeed without explanation. Across 45 degraded gas task–window units, SR-MKL+DR improves standardized log loss over the frozen channel by -0.363 (crossed task–window 95% CI $[-0.752, -0.071]$), but differs from Random+DR by $+0.00489$ (CI $[-0.00356, 0.02174]$), failing the prespecified 0.01 noninferiority criterion. Mechanism diagnosis is inconclusive in 78.72% of runs, with zero agreement with a retrospective single-dominant characterization; reduced candidate families do not repair this. Yet across 40 next-window transitions the downstream predictor trained from SR-MKL+DR supervision has mean log loss 0.807 versus 1.131 Frozen (difference -0.3236, task-cluster CI $[-0.6745, -0.0411]$) and is statistically indistinguishable from Random+DR (0.815). Thus

$$\boxed{\text{useful adaptation} \not\Rightarrow \text{defensible mechanism explanation.}}$$

7 DISCUSSION: WHAT THE SEPARATION BUYS

The theory and experiments support a narrower claim than "active querying helps." First, a declared unlabeled monitor can have an exact null space in which the correct fusion posterior changes, so detecting such shifts requires information beyond the declared unlabeled monitor; randomized sentinels provide a representative implementation that also supports population estimation. Second, when $\Gamma_B(\mathcal{Q}) > 0$, targeted trusted labels can resolve the mechanism efficiently, but no validity theorem makes one acquisition objective uniformly fastest; the harmonized EIG results make this empirical boundary explicit. Third, labels optimized for diagnosis are selected observations, so successful mechanism identification does not by itself license population adaptation. Finally, even selection-valid adaptation does not imply a defensible mechanism explanation: the natural gas experiment improves prediction while returning inconclusive for most mechanism queries. These distinctions motivate the architecture as much as any individual component.

8 RELATED WORK

Static weak-supervision methods estimate latent labels and dependencies from noisy sources (Ratner et al., 2017; Bach et al., 2017); partial-identification work asks what performance can be bounded

without gold labels (Polo et al., 2024). Most directly related, Mazzetto et al. adapt weak supervision under drift by selecting historical windows for conditionally independent sources with drifting accuracies (Mazzetto et al., 2025). Shoeibi et al. call changes in weak-label relationships across contexts *supervision drift* and study stable prediction under distribution shift (Shoeibi et al., 2026). DynaAudit does not claim that term; it asks a complementary identifiability question: reliability, dependence, or prior changes can lie in the same declared monitor-equivalence class, in which case no unlabeled history window or feature-stability diagnostic determines the correct gold posterior.

Mechanism auditing connects to active weak supervision (Biegel et al., 2021; Guan & Koudas, 2024), label-efficient model selection (Okanovic et al., 2025), and controlled sensing/active hypothesis testing (Nitinawarat et al., 2013; Garivier & Kaufmann, 2016). Recent analyses establish performance guarantees for information-maximizing active tests and finite-sample active hypothesis testing (Chattopadhyay et al., 2024; Lin et al., 2026). The architecture also draws on time-uniform evidence (Howard et al., 2021), and logged/selection-aware estimation (Yan et al., 2018; Horvitz & Thompson, 1952; Bang & Robins, 2005). We do not claim novelty for those generic tools. The contribution is the chain *monitor equivalence* $\rightarrow$ *unlabeled impossibility* $\rightarrow$ *support auditability* $\rightarrow$ *adaptive trusted recovery* $\rightarrow$ *selection-valid reuse* for evolving supervision channels, including an explicit model-rejection state. This complements self-healing ML's diagnosis-before-adaptation perspective (Rauba et al., 2024) by showing that useful adaptation can remain possible when a named mechanism is not defensible.

Two distinctions are central to our positioning. First, the novelty claim is not a new generic maximin or information-gain objective: Section 4 deliberately shows that a strong EIG implementation can match or beat robust maximin acquisition. Second, selection correction is not itself new; its role here is architectural. Mechanism-efficient labels are intentionally nonrepresentative, so a deployment pipeline that reuses them must carry the sampling design forward rather than silently switching statistical targets.

9 Limitations

The main certificate assumes a finite, correctly specified candidate family and a mechanism fixed during an audit episode; certified estimated-channel extensions are deferred to the appendix. When near-hidden diagnosis combines passive and trusted evidence, every modeled evidence increment must be included in the sequential likelihood rather than treating the trusted-label certificate as automatically covering the passive term. Primary theory/experiments are binary. Hiddenness is always relative to the declared transcript, and $\Gamma_B = 0$ is a fundamental support limitation rather than an acquisition failure. Sentinels incur real cost/delay and have limited power for smaller shifts. The occupancy and WRENCH mechanism studies are semi-synthetic; the latter retains real weak-label states but simulates drifted gold labels. The natural gas study has no causal mechanism ground truth. Strong EIG often matches or exceeds robust maximin acquisition; an independent WRENCH critic has low out-of-family power on SMS and YouTube for the tested alternative. Selection-valid adaptation requires correct inclusion probabilities and overlap.

10 Conclusion

Stability of the observed weak-source stream does not imply stability of its meaning for the latent label. DynaAudit makes that gap explicit: declare what the monitor sees; obtain trusted evidence when it can be blind, using randomized sentinels when representative evidence is also needed; acquire targeted labels only where candidate mechanisms separate; assert a mechanism only after sufficient evidence and adequate fit; and preserve the sampling design when selected labels are reused. The resulting goal is not universal root-cause naming or acquisition dominance, but a system that distinguishes *what is observable*, *what requires trusted evidence*, *what can be asserted safely*, and *what can be adapted validly*.

AI USE STATEMENT

In this work, generative AI tools were used for literature-search assistance and related-work discovery; development and critique of the conceptual framework; formulation, drafting, and adversarial checking of mathematical claims and proofs; feedback on experimental design; drafting and editing analysis/code notebooks; interpretation and auditing of saved experimental outputs; creation and editing of scientific figures; and manuscript structuring, drafting, editing, reference formatting, and LaTeX preparation. Experiments reported in the paper were executed in the authors' compute environment, and numerical claims were derived from saved experimental artifacts rather than generated by the language model. AI-assisted outputs are research and drafting assistance; the authors are responsible for independently verifying all mathematical claims, citations, code, analyses, and reported results before submission and take responsibility for the final content.

REPRODUCIBILITY STATEMENT

The controlled and temporal studies use prespecified protocol versions, shared seed namespaces for paired policy comparisons, saved trial-level outputs, and cryptographic manifests. The appendix states the assumptions and complete proofs for the adaptive audit results and documents controlled-world construction, sentinel accounting, inclusion-probability validation, temporal benchmark design, the post-hoc strong-baseline harmonization audit, and the semi-synthetic WRENCH support stress test. The accompanying anonymous supplement contains paper/source snapshots, protocols, executable notebooks and code packages, result summaries, verification reports, and hashes identifying the full raw-result archives used for the reported tables and figures.

ETHICS STATEMENT

The empirical study introduces no new human-subject data collection. It uses synthetic supervision mechanisms and public benchmark datasets. The work concerns statistical auditing and uncertainty in deployed supervision systems; where the candidate mechanism family is unsupported, the proposed architecture explicitly permits an inconclusive output rather than forcing a potentially misleading mechanism claim.

REFERENCES

Stephen H. Bach, Bryan He, Alexander Ratner, and Christopher Ré. Learning the structure of generative models without labeled data. In *Proceedings of the 34th International Conference on Machine Learning*, pp. 273–282, 2017.

Heejung Bang and James M. Robins. Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4):962–973, 2005. doi: 10.1111/j.1541-0420.2005.00377.x.

Sander Biegel, Rami El-Khatib, Lucas O. V. B. Oliveira, Max Baak, and Nanne Aben. Active weasul: Improving weak supervision with active learning, 2021. ICLR Workshop on Weakly Supervised Learning.

Luis M. Candanedo and Véronique Feldheim. Accurate occupancy detection of an office room from light, temperature, humidity and co2 measurements using statistical learning models. *Energy and Buildings*, 112:28–39, 2016. doi: 10.1016/j.enbuild.2015.11.071.

Aditya Chattopadhyay, Benjamin David Haeffele, Rene Vidal, and Donald Geman. Performance bounds for active binary testing with information maximization. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 6346–6371, 2024.

Jordi Fonollosa, S. Sheik, Ramón Huerta, and Santiago Marco. Reservoir computing compensates slow response of chemosensor arrays exposed to fast varying gas concentrations in continuous monitoring. *Sensors and Actuators B: Chemical*, 215:618–629, 2015. doi: 10.1016/j.snb.2015.03.028.

Aurélien Garivier and Emilie Kaufmann. Optimal best arm identification with fixed confidence. In *Proceedings of the 29th Conference on Learning Theory*, pp. 998–1027, 2016.

Nan Guan and Nick Koudas. Activedp: Bridging active learning and data programming. In *Proceedings of the 27th International Conference on Extending Database Technology*, pp. 515–521, 2024.

Daniel G. Horvitz and Donovan J. Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260):663–685, 1952. doi: 10.1080/01621459.1952.10483446.

Steven R. Howard, Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon. Time-uniform, nonparametric, nonasymptotic confidence sequences. *The Annals of Statistics*, 49(2):1055–1080, 2021. doi: 10.1214/20-AOS1991.

Ziyuan Lin, Hoang Ngoc Nguyen, Jie Xu, and Ivan Ruchkin. Finite-sample analysis of elimination in active hypothesis testing, 2026.

Alessio Mazzetto, Reza Esfandiarpoor, Abhilash Singirikonda, Eli Upfal, and Stephen H. Bach. An adaptive method for weak supervision with drifting data. In *Proceedings of the 28th International Conference on Artificial Intelligence and Statistics*, pp. 1513–1521, 2025.

Sirin Nitinawarat, George K. Atia, and Venugopal V. Veeravalli. Controlled sensing for multihypothesis testing. *IEEE Transactions on Automatic Control*, 58(10):2451–2464, 2013. doi: 10.1109/TAC.2013.2256216.

Patrik Okanovic, Andreas Kirsch, Jannes Kasper, Torsten Hoefler, Andreas Krause, and Nezihe Merve Gürel. All models are wrong, some are useful: Model selection with limited labels. In *Proceedings of the 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pp. 2035–2043, 2025.

Felipe M. Polo, Subha Maity, Mikhail Yurochkin, Moulinath Banerjee, and Yuekai Sun. Weak supervision performance evaluation via partial identification. In *Advances in Neural Information Processing Systems*, volume 37, pp. 135704–135740, 2024.

Alexander Ratner, Stephen H. Bach, Henry Ehrenberg, Jason Fries, Sen Wu, and Christopher Ré. Snorkel: Rapid training data creation with weak supervision. *Proceedings of the VLDB Endowment*, 11(3):269–282, 2017. doi: 10.14778/3157794.3157797.

Paul Rauba, Nabeel Seedat, Krzysztof Kacprzyk, and Mihaela van der Schaar. Self-healing machine learning: A framework for autonomous adaptation in real-world environments. In *Advances in Neural Information Processing Systems*, volume 37, pp. 42225–42267, 2024.

Mehrdad Shoeibi, Elias Hossain, Ivan Garibay, and Niloofar Yousefi. Learning stable predictors from weak supervision under distribution shift, 2026.

Alexander Vergara, Shankar Vembu, Tuba Ayhan, Margaret A. Ryan, Margie L. Homer, and Ramón Huerta. Chemical gas sensor drift compensation using classifier ensembles. *Sensors and Actuators B: Chemical*, 166–167:320–329, 2012. doi: 10.1016/j.snb.2012.01.074.

Songbai Yan, Kamalika Chaudhuri, and Tara Javidi. Active learning with logged data. In *Proceedings of the 35th International Conference on Machine Learning*, pp. 5521–5530, 2018.

Jieyu Zhang, Yue Yu, Yinghao Li, Yujing Wang, Yaming Yang, Mao Yang, and Alexander Ratner. WRENCH: A comprehensive benchmark for weak supervision. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2021. URL https://openreview.net/forum?id=Q9SKS5k8io.

A Full theory and proofs

A.1 Setup and scope

Let $H \in \mathcal{H} = \{1, \dots, K\}$, with $K \geq 2$, denote the active supervision mechanism. The deployment monitor observes

$$O = \phi(X, \Lambda),$$

where Λ denotes the weak-source outputs and X any features explicitly included in the declared monitor. The theory distinguishes two information channels:

(i) *passive monitor information*, available without trusted labels; and

(ii) *trusted audit information*, obtained by querying Y on selected deployment items.

The passive and active channels serve different statistical roles. The results below do not assume any particular acquisition objective such as expected information gain (EIG) or maximin separation.

For the adaptive results, it is useful to allow a general sequential predictive formulation. Immediately before trusted label Y_t is revealed, let $\mathcal{G}_t$ contain all information available to the auditor, including previous trusted labels, current pool/support information, passive monitor observations, and the randomized choice of the item/state to audit. Candidate h specifies a predictable density

$$p_{h,t}(y \mid \mathcal{G}_t)$$

with respect to a common dominating measure μ. After observing Y_t, set

$$\mathcal{F}_t = \sigma(\mathcal{G}_t, Y_t).$$

The next pre-label sigma-field $\mathcal{G}_{t+1}$ may contain arbitrary new passive information in addition to $\mathcal{F}_t$.

Assumption 3 (Correct candidate channel and common support)**.** *For the true mechanism $h \in \mathcal{H}$,*

$$Y_t \mid \mathcal{G}_t \sim p_{h,t}(\cdot \mid \mathcal{G}_t).$$

For each candidate pair h, k, the predictive densities have common positive support almost surely, so the likelihood ratios and Hellinger affinities below are well defined.

The stationary support-constrained model used in DynaAudit is the special case

$$p_{h,t}(y \mid \mathcal{G}_t) = p_h(y \mid A_t),$$

where $A_t \in \mathcal{S}_t$ is the auditable deployment state selected before observing Y_t. Here an "audit state" must include *all pre-label side information that the acquisition rule is allowed to use to discriminate among items*; otherwise selection within a nominal state can change the conditional label law and the reduced-state channel $p_h(y \mid A_t)$ is not the correct predictive object.

Remark 4 (What the theory does and does not cover)**.** The adaptive guarantees allow arbitrary label-adaptive acquisition, including EIG, SR-MKL, randomized exploration, credible-set-dependent querying, and changing empirical support. They require the final diagnostic likelihood to be computed from a correctly specified candidate predictive family, or from the certified approximation extension in Appendix D. They do not guarantee correctness under arbitrary candidate-family misspecification; that is the role of explicit model criticism and an unknown/inconclusive output.

A.2 Monitor-relative information barriers

Let Z denote the complete unlabeled information available to a specified deployment procedure. This may be a single monitored state, an n-sample transcript, or an entire finite-horizon monitored trajectory. Under mechanisms θ_0, θ_1, let Q_0, Q_1 denote the corresponding laws of Z.

Theorem 5 (Monitor-relative two-point lower bound)**.** *Let $T(\theta)$ take values in a metric space $(\mathcal{T}, d)$ and define*

$$\Delta = d(T(\theta_0), T(\theta_1)).$$

Then every estimator $\widehat{T} = \widehat{T}(Z)$ satisfies

$$\inf_{\widehat{T}} \max_{i \in \{0,1\}} \mathbb{E}_{\theta_i} d\Big(\widehat{T}, T(\theta_i)\Big) \;\geq\; \frac{\Delta}{2} \left[1 - \mathrm{TV}(Q_0, Q_1)\right].$$

In particular, if $Q_0 = Q_1$, the minimax risk is at least $\Delta/2$ no matter how much unlabeled data are contained in Z.

Proof. Write $T_i = T(\theta_i)$ and let q_0, q_1 be densities of Q_0, Q_1 with respect to a common dominating measure. For any realized estimate $a \in \mathcal{T}$, the triangle inequality gives

$$d(a, T_0) + d(a, T_1) \geq d(T_0, T_1) = \Delta.$$

Hence, for any estimator $\widehat{T}$,

$$\begin{aligned} R_0 + R_1 &= \int q_0(z) d(\widehat{T}(z), T_0)\, dz + \int q_1(z) d(\widehat{T}(z), T_1)\, dz \\ &\geq \int \min\{q_0(z), q_1(z)\} \left[d(\widehat{T}(z), T_0) + d(\widehat{T}(z), T_1) \right] dz \\ &\geq \Delta \int \min\{q_0(z), q_1(z)\}\, dz \\ &= \Delta \left[1 - \mathrm{TV}(Q_0, Q_1) \right]. \end{aligned}$$

Therefore $\max(R_0, R_1) \geq (R_0 + R_1)/2$, proving the claim. □

Corollary 6 (Canonical DynaAudit construction)**.** *In the two-source construction with identical weak-output law and*

$$P_A(Y = +1 \mid ++) = \frac{16}{17}, \qquad P_B(Y = +1 \mid ++) = \frac{27}{34},$$

any estimator based only on the unlabeled monitored weak-output transcript has worst-case expected absolute error at least

$$\frac{1}{2} \left| \frac{16}{17} - \frac{27}{34} \right| = \frac{5}{68}.$$

Proof. The monitored transcript law is identical under the two mechanisms, so $\mathrm{TV}(Q_A, Q_B) = 0$. Apply Theorem 5 with absolute-error loss and target $T(\theta) = P_\theta(Y = +1 \mid ++)$. The target separation is $5/34$. □

Corollary 7 (Power ceiling for unlabeled hidden-drift detection)**.** *Let E be any alarm event measurable with respect to the complete unlabeled monitor transcript Z. If*

$$Q_0(E) \leq \alpha,$$

then

$$Q_1(E) \leq \alpha + \mathrm{TV}(Q_0, Q_1).$$

For transcript-exact *hidden drift, $Q_0 = Q_1$, so the alarm probability under the changed mechanism cannot exceed the false-alarm probability:*

$$Q_1(E) \leq \alpha.$$

If the complete transcript laws are identical, the entire distribution of any unlabeled stopping time is identical under the two mechanisms. In the i.i.d. exact-hidden benchmarks, equality of the one-step monitored law implies equality of every finite-horizon transcript law; without such a process assumption, equality of one-step marginals alone is not enough.

Proof. By the variational characterization of total variation,

$$Q_1(E) - Q_0(E) \leq \mathrm{TV}(Q_0, Q_1).$$

The result follows from $Q_0(E) \leq \alpha$. □

Remark 8 (Near-hidden drift)**.** The total variation in Theorem 5 and Corollary 7 is the distance between the *complete monitor transcript laws* available to the procedure. A small one-observation total variation does not remain small automatically as the number of passive observations grows. The exact-hidden case is qualitatively different because transcript-level equality persists for every sample size.

A.3 SUPPORT AUDITABILITY

For the remainder of this section, consider stationary trusted-label channels $p_h(y \mid o)$ on a finite auditable support $\mathcal{S}$, in a repeated-draw/superpopulation model in which the declared audit contexts can be sampled according to a feasible design. Finite-pool capacities or no-replacement restrictions should instead be encoded in a constrained feasible design set as described below. Define the pairwise Bhattacharyya information

$$b_{h,k}(o) = -\log \int \sqrt{p_h(y \mid o) p_k(y \mid o)}\, d\mu(y).$$

It is nonnegative and equals zero exactly when the two conditional label laws coincide almost everywhere.

Definition 9 (Support-constrained auditability coefficient)**.** Define

$$\Gamma_B(\mathcal{S}) = \max_{q \in \Delta(\mathcal{S})} \min_{h \neq k} \sum_{o \in \mathcal{S}} q(o) b_{h,k}(o),$$

where $\Delta(\mathcal{S})$ is the probability simplex over the current auditable support.

Proposition 10 (Characterization of supported auditability)**.** *For finite $\mathcal{H}$ and $\mathcal{S}$,*

$$\Gamma_B(\mathcal{S}) > 0$$

if and only if every candidate pair is separated by at least one supported audit state:

$$\forall h \neq k, \quad \exists o \in \mathcal{S} \text{ such that } p_h(\cdot \mid o) \neq p_k(\cdot \mid o).$$

Proof. If $\Gamma_B(\mathcal{S}) > 0$, then for its maximizing design q every pair has strictly positive weighted Bhattacharyya information, hence each pair must differ on at least one state assigned positive mass.

Conversely, suppose every pair differs on at least one supported state. Let q be the uniform distribution on $\mathcal{S}$. For each fixed pair (h, k), at least one summand $b_{h,k}(o)$ is strictly positive and all are nonnegative, so

$$\sum_o q(o) b_{h,k}(o) > 0.$$

There are finitely many candidate pairs; therefore the minimum of these finitely many positive quantities is itself positive. Hence $\Gamma_B(\mathcal{S}) > 0$. □

Definition 11 (Hidden-but-auditable pair)**.** A pair (h, k) is *hidden but auditable* relative to the declared monitor and current support $\mathcal{S}$ when the relevant passive monitor transcript laws agree,

$$Q_h = Q_k,$$

(and, in the i.i.d. setting, it is enough that the one-step monitored laws agree, $P_h(O) = P_k(O)$), but

$$\max_{o \in \mathcal{S}} b_{h,k}(o) > 0.$$

Thus the passive monitor cannot distinguish the mechanisms, while a trusted label at an appropriate supported state can.

Proposition 12 (Support non-separation is an impossibility, not an acquisition failure)**.** *Work in the stationary audit model in which the passive continuation/context process has the same conditional law under h and k given every common observed history, and trusted labels obey $Y_t \mid A_t = o \sim p_j(\cdot \mid o)$ under candidate j. If*

$$p_h(\cdot \mid o) = p_k(\cdot \mid o) \qquad \forall o \in \mathcal{S}$$

for every state that an admissible audit policy can query, then every adaptive audit policy has the same complete observable-and-audited transcript law under h and k. Consequently, under equal prior probability on the pair, every diagnosis rule has pairwise error at least $1/2$, regardless of audit budget.

Proof. Conditional on any common realized history, the next passive/context observation has the same law under the two mechanisms by assumption, and the audit policy applies the same randomized action kernel because it depends only on that observed history and its own randomization. Every admissible selected state lies in $\mathcal{S}$, where the conditional trusted-label laws agree under h and k. Induction over passive and audited increments therefore gives equality of the full transcript laws. Equal transcript laws imply Bayes average error $1/2$ for testing the two equally likely mechanisms, and hence minimax pairwise error at least $1/2$. □

Remark 13 (Why marginal passive equivalence alone is not enough)**.** Equality of the marginal law of the full passive sequence is sufficient in the i.i.d. deployment model used by the exact-hidden benchmarks, where passive contexts are generated independently across items and the audit label channel is item-wise. In a general dependent process, a queried label could change the conditional law of future passive observations differently across mechanisms even when the passive sequence has the same marginal law before conditioning. The proposition therefore states the stronger conditional-continuation condition actually needed for transcript equality.

Remark 14 (Constrained design sets)**.** If deployment constraints restrict feasible long-run allocations to a subset $\mathcal{Q} \subseteq \Delta(\mathcal{S})$, replace Definition 9 by

$$\Gamma_B(\mathcal{Q}) = \sup_{q \in \mathcal{Q}} \min_{h \neq k} \sum_o q(o) b_{h,k}(o).$$

Then $\Gamma_B(\mathcal{Q}) > 0$ means there exists at least one feasible allocation that gives positive information against every pair. This is the appropriate formulation when capacities, no-replacement constraints, or operational restrictions matter.

A.4 Adaptive Bhattacharyya martingale

For candidate pair (h, k) define the sequential likelihood ratio

$$L_t^{k/h} = \prod_{s=1}^{t} \frac{p_{k,s}(Y_s \mid \mathcal{G}_s)}{p_{h,s}(Y_s \mid \mathcal{G}_s)}$$

and the predictable conditional Bhattacharyya information

$$b_{h,k,s} = -\log \int \sqrt{p_{h,s}(y \mid \mathcal{G}_s) p_{k,s}(y \mid \mathcal{G}_s)}\, d\mu(y).$$

Let

$$C_{h,k}(t) = \sum_{s=1}^{t} b_{h,k,s}.$$

Note that $b_{h,k,s} = b_{k,h,s}$ and therefore $C_{h,k}(t) = C_{k,h}(t)$.

Lemma 15 (Adaptive Hellinger martingale)**.** *Under true mechanism h and Assumption 3,*

$$M_t^{k/h} = \sqrt{L_t^{k/h}}\, \exp\{C_{h,k}(t)\}, \qquad M_0^{k/h} = 1,$$

is a nonnegative martingale with respect to $\{\mathcal{F}_t\}$.

Proof. Conditioning on $\mathcal{G}_t$,

$$\begin{aligned}
\mathbb{E}_h\left[\sqrt{\frac{p_{k,t}(Y_t \mid \mathcal{G}_t)}{p_{h,t}(Y_t \mid \mathcal{G}_t)}}\,\middle|\,\mathcal{G}_t\right] &= \int p_{h,t}(y \mid \mathcal{G}_t) \sqrt{\frac{p_{k,t}(y \mid \mathcal{G}_t)}{p_{h,t}(y \mid \mathcal{G}_t)}}\, d\mu(y) \\
&= \int \sqrt{p_{h,t}(y \mid \mathcal{G}_t) p_{k,t}(y \mid \mathcal{G}_t)}\, d\mu(y) \\
&= e^{-b_{h,k,t}}.
\end{aligned}$$

Since $b_{h,k,t}$ is $\mathcal{G}_t$-measurable,

$$\mathbb{E}_h[M_t^{k/h} \mid \mathcal{G}_t] = M_{t-1}^{k/h}.$$

Taking conditional expectation with respect to $\mathcal{F}_{t-1} \subseteq \mathcal{G}_t$ proves the martingale property. Nonnegativity is immediate. □

Remark 16 (Why label-adaptive acquisition is allowed)**.** The selected audit state may depend arbitrarily on previous trusted labels and on new passive observations, because it is already included in $\mathcal{G}_t$ before Y_t is revealed. The proof conditions on the realized predictable choice. No fixed-design assumption is needed.

B ANYTIME-VALID ADAPTIVE DIAGNOSIS

Theorem 17 (Pairwise anytime audit certificate)**.** *For any pair $h \neq k$, any $r > 0$, and any $c \geq 0$,*

$$\mathbb{P}_h\Big(\exists t \geq 1 :\ L_t^{k/h} \geq r \text{ and } C_{h,k}(t) \geq c\Big) \leq r^{-1/2} e^{-c}.$$

In particular,

$$\mathbb{P}_h\Big(\exists t \geq 1 :\ L_t^{k/h} \geq 1 \text{ and } C_{h,k}(t) \geq c\Big) \leq e^{-c}.$$

Proof. On the displayed event,

$$M_t^{k/h} = \sqrt{L_t^{k/h}}\, e^{C_{h,k}(t)} \geq \sqrt{r}\, e^{c}.$$

By Lemma 15, $M_t^{k/h}$ is a nonnegative martingale with initial value one. Ville's inequality therefore gives

$$\mathbb{P}_h\left(\sup_t M_t^{k/h} \geq \sqrt{r}\, e^{c}\right) \leq (\sqrt{r}\, e^{c})^{-1}.$$

□

For equal candidate priors, define the conditional-likelihood leader

$$\widehat{H}_t \in \arg\max_{j \in \mathcal{H}} \prod_{s=1}^{t} p_{j,s}(Y_s \mid \mathcal{G}_s),$$

with any deterministic measurable tie-breaking rule.

Theorem 18 (Anytime multihypothesis certificate)**.** *Fix $c > 0$ and define the stopping time*

$$\tau_c = \inf\left\{ t \geq 1 :\ \min_{j \neq \widehat{H}_t} C_{\widehat{H}_t, j}(t) \geq c \right\},$$

with $\inf \varnothing = \infty$. Under Assumption 3, for every true $h \in \mathcal{H}$,

$$\mathbb{P}_h\Big(\tau_c < \infty, \widehat{H}_{\tau_c} \neq h\Big) \leq (K-1)e^{-c}.$$

Consequently, with

$$c_\delta = \log \frac{K-1}{\delta},$$

the rule that asserts $\widehat{H}_{\tau_{c_\delta}}$ only when $\tau_{c_\delta} < \infty$ has error probability at most δ for every candidate mechanism.

Proof. Suppose the true mechanism is h and the stopping rule makes an incorrect assertion $k \neq h$ at time $t = \tau_c$. Since k is the conditional-likelihood leader,

$$L_t^{k/h} = \frac{\prod_{s \leq t} p_{k,s}(Y_s \mid \mathcal{G}_s)}{\prod_{s \leq t} p_{h,s}(Y_s \mid \mathcal{G}_s)} \geq 1.$$

The stopping condition also implies

$$C_{k,h}(t) \geq c.$$

By symmetry of Bhattacharyya information, $C_{k,h}(t) = C_{h,k}(t)$. Thus an incorrect assertion of k implies the pairwise event in Theorem 17. Union bounding over the $K-1$ possible wrong leaders gives

$$\mathbb{P}_h(\tau_c < \infty, \widehat{H}_{\tau_c} \neq h) \leq (K-1)e^{-c}.$$

Substituting $c = c_\delta$ proves the second statement. □

Corollary 19 (Validity versus eventual assertion)**.** *The bound in Theorem 18 controls the probability of a wrong assertion but, by itself, does not guarantee that the stopping time is finite. If the realized audit design satisfies*

$$\min_{a \neq b} C_{a,b}(t) \longrightarrow \infty \qquad \textit{almost surely},$$

then $\tau_c < \infty$ almost surely for every fixed c. In particular, under the stationary repeated-draw model, any design whose long-run allocation gives strictly positive Bhattacharyya rate to every candidate pair has eventual assertion. Combining this with Theorem 18 yields both eventual stopping and the stated error guarantee.

Proof. If the minimum accumulated information over all candidate pairs diverges, then there is an almost surely finite time after which every pair has accumulated at least c. At that time the leader-centered minimum in the definition of τ_c is also at least c, irrespective of which candidate is the current leader. □

Corollary 20 (Fixed-budget error decomposition)**.** *At any deterministic audit budget B, define the true-mechanism worst-alternative information*

$$C_{\min,h}(B) = \min_{k \neq h} C_{h,k}(B).$$

Then for every $c > 0$,

$$\boxed{\mathbb{P}_h(\widehat{H}_B \neq h) \leq (K-1)e^{-c} + \mathbb{P}_h(C_{\min,h}(B) < c)\,.}$$

Proof. Split the error event according to whether $C_{\min,h}(B) \geq c$. On the event $\{\widehat{H}_B = k \neq h,\ C_{\min,h}(B) \geq c\}$, the conditional-likelihood rule gives $L_B^{k/h} \geq 1$ and the definition of $C_{\min,h}$ gives $C_{h,k}(B) \geq c$. A union bound over $k \neq h$ and Theorem 17 control this part by $(K-1)e^{-c}$. The complementary part is at most $\mathbb{P}_h(C_{\min,h}(B) < c)$. □

Remark 21 (Interpretation of Corollary 20)**.** The bound separates two failure modes:

$$\underbrace{(K-1)e^{-c}}_{\text{wrong despite sufficient pairwise evidence}} + \underbrace{\mathbb{P}_h(C_{\min,h}(B) < c)}_{\text{acquisition failed to accumulate enough evidence}}.$$

The first term is acquisition-policy agnostic. EIG, SR-MKL, random auditing, or any other predictable acquisition rule affects the second term through the realized information trajectory.

Corollary 22 (Recovery of the fixed-design Bhattacharyya bound)**.** *If the audit sequence is fixed before labels are revealed, then every $C_{h,k}(B)$ is deterministic and*

$$\mathbb{P}_h(\widehat{H}_B \neq h) \leq \sum_{k \neq h} e^{-C_{h,k}(B)}.$$

Thus the previous fixed-design result is a special case of the adaptive martingale argument.

Proof. For each $k \neq h$, apply Theorem 17 with $c = C_{h,k}(B)$ at the fixed time B. Since $C_{h,k}(B)$ is deterministic, the event $\{L_B^{k/h} \geq 1\}$ has probability at most $e^{-C_{h,k}(B)}$. If the conditional-likelihood decision is wrong, at least one such event occurs. Union bound over $k \neq h$. □

B.1 Unequal candidate weights

Corollary 23 (MAP certificate with fixed positive weights)**.** *Let $\pi_h > 0$ be candidate weights fixed before the adaptive audit begins, and let*

$$\widehat{H}_t^{\mathrm{MAP}} \in \arg\max_j \pi_j \prod_{s \leq t} p_{j,s}(Y_s \mid \mathcal{G}_s).$$

Define

$$\widetilde{C}_{k,j}(t) = C_{k,j}(t) + \frac{1}{2}\log\frac{\pi_j}{\pi_k}.$$

If

$$\widetilde{\tau}_c = \inf\left\{ t : \min_{j \neq \widehat{H}_t^{\mathrm{MAP}}} \widetilde{C}_{\widehat{H}_t^{\mathrm{MAP}},j}(t) \geq c \right\},$$

then for every true h,

$$\mathbb{P}_h\left(\widetilde{\tau}_c < \infty, \widehat{H}_{\widetilde{\tau}_c}^{\mathrm{MAP}} \neq h\right) \leq (K-1)e^{-c}.$$

The same conclusion holds conditionally and hence unconditionally if the weights are $\mathcal{F}_0$-measurable, for example after a fixed passive-evidence/sentinel initialization stage.

Proof. If the true mechanism is h but MAP selects $k \neq h$, then

$$L_t^{k/h} \geq \frac{\pi_h}{\pi_k}.$$

The stopping condition for the pair (k, h) gives

$$C_{h,k}(t) + \frac{1}{2}\log\frac{\pi_h}{\pi_k} \geq c.$$

Therefore

$$M_t^{k/h} = \sqrt{L_t^{k/h}} e^{C_{h,k}(t)} \geq e^c.$$

Apply Ville's inequality and union bound as in Theorem 18. □

B.2 Necessary trusted-label complexity

The previous results are sufficient conditions for safe assertion. We now give a complementary lower bound. This is a specialization of classical adaptive change-of-measure arguments to monitor-equivalent supervision mechanisms.

Assume stationary channels $p_h(y \mid o)$ on finite support $\mathcal{S}$, and define directed audit information

$$d_{h,k}(o) = D_{\mathrm{KL}}(p_h(\cdot \mid o) \,\|\, p_k(\cdot \mid o)).$$

For true candidate h, define the oracle support-constrained rate

$$C_h^{\mathrm{KL},\star} = \max_{q \in \Delta(\mathcal{S})} \min_{k \neq h} \sum_{o \in \mathcal{S}} q(o) d_{h,k}(o).$$

Assumption 24 (No passive likelihood contribution between the compared mechanisms)**.** *For the lower-bound pair (h, k), every passive/context increment has the same conditional law under h and k given the complete observed history and chosen actions. Equivalently, the likelihood ratio of the full adaptive transcript has no passive-data factor, so its mechanism-dependent increments are exactly the queried trusted-label likelihood ratios. This condition holds in the i.i.d. exact-hidden benchmark model where the monitored state law is identical and items are generated independently across rounds.*

Theorem 25 (Necessary trusted-label complexity)**.** *Consider any adaptive audit procedure with almost surely finite stopping time τ and terminal named-candidate decision $\widehat{H} \in \mathcal{H}$. Suppose it is uniformly δ-correct for $0 < \delta < 1/2$:*

$$\mathbb{P}_h(\widehat{H} \neq h) \leq \delta \qquad \forall h \in \mathcal{H}.$$

Assume additionally that the relevant pairwise KL divergences are finite. Under Assumption 24, for every true mechanism h with $\mathbb{E}_h[\tau] < \infty$,

$$\boxed{\mathbb{E}_h[\tau] \geq \frac{\mathrm{kl}(1-\delta, \delta)}{C_h^{\mathrm{KL},\star}}.}$$

If $C_h^{\mathrm{KL},\star} = 0$, no finite-expected-budget uniformly δ-correct procedure that must return a named candidate exists. A procedure allowed to abstain with nontrivial probability requires an additional completeness/success condition before a label-complexity lower bound of this form is meaningful.

Proof. Fix $k \neq h$ and let $E_h = \{\widehat{H} = h\}$. Uniform δ-correctness gives

$$\mathbb{P}_h(E_h) \geq 1 - \delta, \qquad \mathbb{P}_k(E_h) \leq \delta.$$

By data processing for KL divergence applied to the indicator of E_h,

$$D_{\mathrm{KL}}(P_h^{\mathrm{tr}} \| P_k^{\mathrm{tr}}) \geq \mathrm{kl}(\mathbb{P}_h(E_h), \mathbb{P}_k(E_h)) \geq \mathrm{kl}(1-\delta, \delta),$$

where P_h^{tr} and P_k^{tr} are the laws of the complete stopped audit transcript. Under Assumption 24, passive/context factors and the policy's randomized action kernel cancel in the transcript likelihood ratio. For each deterministic n, the KL chain rule applied to the stopped experiment at $\tau \wedge n$ gives

$$D_{\mathrm{KL}}(P_{h,n}^{\mathrm{tr}} \| P_{k,n}^{\mathrm{tr}}) = \mathbb{E}_h \left[\sum_{t=1}^{\tau \wedge n} d_{h,k}(A_t) \right].$$

The right-hand side increases to the expected stopped information by monotone convergence. The left-hand side increases to the KL divergence on the sigma-field generated by the complete stopped transcript (equivalently, invoke the standard stopped-experiment change-of-measure lemma). Hence

$$D_{\mathrm{KL}}(P_h^{\mathrm{tr}} \| P_k^{\mathrm{tr}}) = \mathbb{E}_h \left[\sum_{t=1}^{\tau} d_{h,k}(A_t) \right].$$

This passage uses the common-support/finite-KL conditions above; without them the identity should be stated in the corresponding extended-valued form. Let $N_o(\tau)$ count audits of state o. Then for every $k \neq h$,

$$\sum_o \mathbb{E}_h[N_o(\tau)] d_{h,k}(o) \geq \mathrm{kl}(1-\delta, \delta).$$

If $n_h = \mathbb{E}_h[\tau] < \infty$, define

$$q_h(o) = \frac{\mathbb{E}_h[N_o(\tau)]}{n_h}.$$

This is a probability distribution over $\mathcal{S}$. Therefore

$$\min_{k \neq h} \sum_o q_h(o) d_{h,k}(o) \geq \frac{\mathrm{kl}(1-\delta, \delta)}{n_h}.$$

By the definition of $C_h^{\mathrm{KL},\star}$,

$$C_h^{\mathrm{KL},\star} \geq \min_{k \neq h} \sum_o q_h(o) d_{h,k}(o),$$

which rearranges to the claimed lower bound. □

Remark 26 (Near-hidden mechanisms)**.** If passive monitor data also distinguish h from k, the change-of-measure lower bound contains both passive and active KL contributions. In that regime, passive evidence can reduce the trusted-label requirement. The clean bound in Theorem 25 is intentionally stated for exact monitor-equivalent mechanisms, where passive KL is zero and trusted labels are information-theoretically necessary.

C Comparison of Bhattacharyya and KL information rates

For binary trusted labels, let $P = \mathrm{Bern}(p)$ and $Q = \mathrm{Bern}(q)$. Suppose all candidate probabilities are clipped to $[\eta, 1-\eta]$, $0 < \eta < 1/2$. Let $B(P, Q)$ denote Bhattacharyya information. Monotonicity of Rényi divergence gives

$$2B(P, Q) = D_{1/2}(P \| Q) \leq D_{\mathrm{KL}}(P \| Q),$$

For the upper bound, the log inequality gives $D_{\mathrm{KL}}(P \| Q) \leq \chi^2(P \| Q) = (p-q)^2/[q(1-q)]$. Writing $A = \sqrt{pq} + \sqrt{(1-p)(1-q)} = e^{-B(P,Q)}$, the total-variation–affinity bound gives $(p-q)^2 \leq 1 - A^2 \leq 2(1-A) \leq 2B(P, Q)$. Since $q(1-q) \geq \eta(1-\eta)$, we obtain

$$\boxed{2B(P, Q) \leq D_{\mathrm{KL}}(P \| Q) \leq \frac{2}{\eta(1-\eta)} B(P, Q).} \tag{9}$$

For a fixed true candidate h, define the candidate-specific Bhattacharyya and directed-KL oracle rates

$$C_h^{B,\star} = \max_{q\in\Delta(\mathcal{S})} \min_{k\neq h} \sum_o q(o) b_{h,k}(o), \qquad C_h^{\mathrm{KL},\star} = \max_{q\in\Delta(\mathcal{S})} \min_{k\neq h} \sum_o q(o) d_{h,k}(o).$$

Applying (9) pointwise and then taking the same min–max operations gives

$$\boxed{2C_h^{B,\star} \le C_h^{\mathrm{KL},\star} \le \frac{2}{\eta(1-\eta)} C_h^{B,\star}.}$$

An analogous constant comparison holds for global robust rates that use a single allocation against all ordered candidate pairs.

This is an *information-geometry comparison*, not by itself a proof that a particular adaptive acquisition algorithm attains the corresponding oracle rate. The lower bound in Theorem 25 uses the candidate-specific directed-KL rate. The anytime theorem instead certifies any realized adaptive path through its accumulated Bhattacharyya information. A separate acquisition analysis is required to upper-bound the stopping time of EIG, SR-MKL, or another specific policy.

D CERTIFIED ESTIMATED-CHANNEL EXTENSION

The exact adaptive martingale assumes the candidate predictive channel used in the likelihood is correct under its corresponding candidate. A conservative extension is possible when estimated channels have certified multiplicative error envelopes.

Let $\widehat{p}_{j,t}(\cdot \mid \mathcal{G}_t)$ be predictable estimated candidate densities, and define

$$\widehat{L}_t^{k/h} = \prod_{s=1}^{t} \frac{\widehat{p}_{k,s}(Y_s \mid \mathcal{G}_s)}{\widehat{p}_{h,s}(Y_s \mid \mathcal{G}_s)},$$

$$\widehat{b}_{h,k,t} = -\log \int \sqrt{\widehat{p}_{h,t}(y \mid \mathcal{G}_t)\widehat{p}_{k,t}(y \mid \mathcal{G}_t)}\, d\mu(y).$$

Assumption 27 (Certified denominator calibration)**.** *For each candidate h there is a predictable nonnegative bound $r_{h,t}$ such that, whenever h is the true mechanism,*

$$\frac{p_{h,t}(y \mid \mathcal{G}_t)}{\widehat{p}_{h,t}(y \mid \mathcal{G}_t)} \le e^{r_{h,t}}$$

for all y in the common support, almost surely.

Proposition 28 (Estimated-channel supermartingale)**.** *Under true mechanism h and Assumption 27,*

$$\widehat{M}_t^{k/h} = \sqrt{\widehat{L}_t^{k/h}} \exp\left\{ \sum_{s=1}^{t} \left(\widehat{b}_{h,k,s} - r_{h,s} \right) \right\}$$

is a nonnegative supermartingale with initial value one.

Proof. Condition on $\mathcal{G}_t$. Then

$$\begin{aligned}
\mathbb{E}_h\left[\sqrt{\frac{\widehat{p}_{k,t}(Y_t \mid \mathcal{G}_t)}{\widehat{p}_{h,t}(Y_t \mid \mathcal{G}_t)}} \,\middle|\, \mathcal{G}_t \right] &= \int p_{h,t}(y \mid \mathcal{G}_t) \sqrt{\frac{\widehat{p}_{k,t}(y \mid \mathcal{G}_t)}{\widehat{p}_{h,t}(y \mid \mathcal{G}_t)}}\, d\mu(y) \\
&= \int \sqrt{\widehat{p}_{h,t}(y \mid \mathcal{G}_t)\widehat{p}_{k,t}(y \mid \mathcal{G}_t)} \frac{p_{h,t}(y \mid \mathcal{G}_t)}{\widehat{p}_{h,t}(y \mid \mathcal{G}_t)}\, d\mu(y) \\
&\le \exp\{r_{h,t} - \widehat{b}_{h,k,t}\}.
\end{aligned}$$

Multiplying by $\exp\{\widehat{b}_{h,k,t} - r_{h,t}\}$ gives conditional expectation at most one for the one-step multiplicative increment. Hence $\widehat{M}_t^{k/h}$ is a nonnegative supermartingale. □

Corollary 29 (Robust effective-evidence certificate)**.** *Let*

$$\widehat{C}^{\mathrm{eff}}_{h,k}(t) = \sum_{s=1}^{t} \left(\widehat{b}_{h,k,s} - r_{h,s} \right).$$

Then

$$\mathbb{P}_h\Big(\exists t : \ \widehat{L}_t^{k/h} \geq 1, \ \widehat{C}^{\mathrm{eff}}_{h,k}(t) \geq c \Big) \leq e^{-c}.$$

Thus certified candidate-channel error consumes accumulated evidence rather than invalidating the sequential argument outright. This pairwise statement is valid even when the effective-evidence penalty is candidate-specific; an observable multihypothesis stopping rule must use envelopes available for every candidate (for example the alternative-index penalty required under each possible true candidate), rather than an unknown truth-specific penalty.

Remark 30 (Operational use)**.** The extension is useful only when $r_{h,t}$ is itself externally certified or bounded with an additional valid uncertainty procedure. Plugging in an unverified training error or an in-sample fit measure does not provide the guarantee. Arbitrary model misspecification remains outside the theorem and should trigger model criticism rather than a forced diagnosis.

D.1 Capability taxonomy and the meaning of "unknown"

The theory distinguishes three different reasons not to assert a named mechanism:

(1) **Support non-identifiability:** some candidate pair has identical trusted-label channels on every currently auditable state, so $\Gamma_B(\mathcal{S}) = 0$. No acquisition rule can resolve the pair on current support.

(2) **Finite-budget insufficiency:** the family is auditable in principle, but the realized audit path has not yet accumulated enough evidence; for example, $\min_{j \neq \widehat{H}_t} C_{\widehat{H}_t,j}(t) < c_\delta$.

(3) **Candidate-family misspecification:** the audited labels are poorly explained by all candidates. This case is not repaired by more confidence in the within-family posterior and requires an absolute-fit/model-criticism check.

These cases should be reported separately whenever possible. Only the second is fundamentally a finite-budget acquisition problem.

Passive monitor	Trusted audit channel on support	Conclusion
Distinguishable	Distinguishable	Passive and active evidence may both help
Indistinguishable	Distinguishable	Hidden-but-auditable; core DynaAudit regime
Indistinguishable	Indistinguishable	Impossible with declared monitor/support
Any	Candidate family fails fit	Out-of-family; return unknown/inconclusive

D.2 Claims supported by this theory

The results justify the following statements.

T1. Exact monitor equivalence can make unlabeled posterior estimation and unlabeled drift detection information-theoretically impossible.

T2. Trusted labels can restore identifiability only if candidate mechanisms differ on states that are actually auditable.

T3. Fully adaptive trusted-label acquisition admits a finite-sample, anytime-valid mechanism certificate based on accumulated pairwise Bhattacharyya information.

T4. The guarantee is acquisition-policy agnostic: EIG, SR-MKL, random, or future predictable policies differ in how quickly they accumulate the required evidence, not in the validity of the certificate.

T5. Any uniformly correct auditor for exact-hidden candidates must spend trusted labels at a rate controlled by support-constrained directed-KL separation.

T6. For clipped Bernoulli channels, candidate-specific Bhattacharyya and directed-KL oracle information rates are constant-equivalent; this does not by itself prove that any particular adaptive acquisition policy attains either oracle rate.

The theory does *not* justify the following stronger claims.

N1. It does not prove that SR-MKL, EIG, or any other proposed acquisition policy is finite-sample optimal.

N2. It does not guarantee recovery when the true supervision mechanism is outside the declared candidate family.

N3. It does not imply that a small one-step unlabeled divergence remains small over arbitrarily long passive monitoring horizons.

N4. It does not make targeted diagnostic labels representative for population estimation; selection-valid adaptation still requires representative sentinels or valid inclusion/propensity correction.

N5. The anytime certificate controls validity of any assertion but does not by itself guarantee that a chosen acquisition policy reaches the stopping threshold quickly, or at all.

D.3 Relationship to established theory

The lower-bound argument in Theorem 5 is a direct two-point/Le Cam argument. The adaptive sample-complexity lower bound in Theorem 25 is a specialization of classical change-of-measure ideas used in controlled sensing and fixed-confidence best-arm identification. The anytime arguments in Lemma 15 and Theorem 17 use a nonnegative likelihood-ratio/Hellinger martingale and Ville's inequality, consistent with the modern e-process/confidence-sequence literature. The contribution claimed for DynaAudit should therefore be the supervision-drift formulation and the resulting chain

$$\boxed{\begin{array}{c}\text{monitor equivalence} \Longrightarrow \text{unlabeled impossibility} \\ \Longrightarrow \text{support auditability} \Longrightarrow \text{adaptive trusted recovery}\end{array}}$$

not the invention of the underlying generic sequential-testing tools.

E Experimental protocols and reproducibility

Controlled suites. The two-source exact-hidden suite contains three accuracy-versus-dependence world pairs, including the construction in Fig. 1; the near-hidden suite perturbs the declared monitor law to one-step total variation 0.015. The multi-source suite uses genuine three- and five-source joint error laws. Accuracy alternatives modify source marginal error rates while retaining conditional independence; dependence alternatives preserve source marginals while modifying the joint error law. The real-label controlled study uses five heterogeneous out-of-fold weak learners trained on disjoint feature groups of the scikit-learn breast-cancer data; audited labels are the original dataset labels, while only the weak-source failure mechanism is controlled.

Shared trial units and inference. All policy comparisons use shared master seeds so that candidate worlds, pools, and revealed labels are paired across acquisition methods. Primary contrasts use paired bootstrap intervals over these shared units. The prespecified SR-MKL and Random outputs are never regenerated during the strong-baseline audit; only the stronger EIG policies are rerun on the immutable worlds and seed namespaces.

Sentinel and complete trusted-label cost. The controlled sentinel stream contains 4,000 arriving observations with change at $t = 1200$ and sentinel rates from 1% to 10%. The primary operating point is 5% with a 5% false-alarm target. Total cost charges all pre-change sentinels, all post-change sentinels through alarm, and the additional diagnostic audits. At the 5% operating point the componentwise medians are 60 pre-change sentinel labels and 43 post-change sentinel labels through alarm, followed by $B = 10$ diagnostic audits. The directly computed run-level median total is 114 labels (mean 119.13; 90th percentile 163), because the median of a sum need not equal the sum of marginal medians.

Selected-label correction. For the prespecified with-replacement diagnostic batch with per-draw probability q_i and B draws, the first-order inclusion probability is

$$\pi_i = 1 - (1 - q_i)^B.$$

The implementation-level propensity audit compares this analytic value with empirical inclusion frequency over 200,000 Monte Carlo repetitions; the maximum absolute discrepancy is 0.00157. Population-estimation experiments deliberately oversample agreement states and compare naive selected-sample estimation, Horvitz–Thompson, Hájek, doubly robust correction, and a randomized-sentinel reference.

Temporal benchmarks. The natural study uses the UCI Gas Sensor Array Drift data (Vergara et al., 2012; Fonollosa et al., 2015): 13,910 labeled observations collected over 36 months. Ten binary gas tasks across five deployment windows give 50 task–window units; 45 exceed the prespecified degradation threshold. Uncertainty is recomputed using task-cluster, window-cluster, and crossed task–window bootstraps. The temporal mechanism benchmark uses the UCI Occupancy Detection time series (Candanedo & Feldheim, 2016). Historical weak-source training ends before 2015-02-05. An objective eligibility rule selects four future weekdays, and rows are subsequently balanced 1:1 by gold class within each difficulty stratum before weak-source error mechanisms are controlled. Exact equivalence is therefore asserted only for the declared state $O = (S, \Lambda)$, not unrestricted raw X or the natural future joint distribution.

Artifacts. The accompanying supplementary archive contains the prespecified protocols, source snapshots, executable audit notebooks, result summaries, cryptographic manifests, verification reports, and SHA-256 identifiers for the full raw-result archives. The theorem appendices above are self-contained and state the assumptions under which the adaptive certificate is valid.

F Deployment guidance

The framework suggests several deployment rules. *Declare the information set before claiming hiddenness:* if raw covariates, embeddings, or telemetry are monitored, they belong in O and can invalidate an apparent equivalence construction. *Keep representative evidence even when the unlabeled monitor is quiet:* exact-hidden drift is invisible to that monitor by definition, while the sentinel rate exposes an explicit cost–power–delay trade-off that should be budgeted end-to-end rather than reporting only post-alarm labels. *Match acquisition to the diagnostic objective, not predictive uncertainty alone:* an example can be maximally label-uncertain yet mechanism-uninformative, and strong EIG or robust maximin designs can each be appropriate depending on candidate geometry.

Deployment software should also preserve reasons for uncertainty. A named mechanism should be withheld when current support cannot separate candidates, when the audit has not accumulated enough evidence, or when the candidate family fails absolute fit. These cases call for different responses: broaden support or sensing, collect more trusted evidence, or revise the mechanism family. Any diagnostically selected labels reused for calibration or model updating should carry their first-order inclusion probabilities (or an otherwise justified sequential estimator); per-draw action probabilities are not automatically population weights. Downstream predictive benefit should be evaluated separately from mechanistic diagnosis because the natural results show that one can improve without the other.

G Acquisition-policy audits

Why the EIG baseline was harmonized. The original controlled implementation used a temperature-softened EIG policy, while a later independently specified temporal implementation used the parameter-free multiplicity-aware rule $q(o) \propto N(o)\,\mathrm{EIG}(o)$. The baseline-strength audit therefore reruns only the stronger EIG rules on the immutable historical worlds. Relative to the legacy temperature-softened EIG, the multiplicity-aware rule improves primary-endpoint accuracy by 5.33 percentage points in the two-source exact benchmark and 5.30 points in the real-label controlled benchmark. These changes motivate the policy-agnostic main-text claim.

Geometry stress test. A separate exact-hidden audit constructs five-source candidate families with identical declared $P(\Lambda)$ and varies both the number of candidate mechanisms, $K \in \{3, 5, 8\}$, and a geometry bottleneck controlling where pairwise mechanism information is concentrated. At $B = 10$, multiplicity-aware EIG has the larger macro-accuracy point estimate in all nine $K \times$geometry cells and the larger mean accumulated worst-alternative Bhattacharyya information $C_{\min}$ in all nine, while worst-candidate accuracy changes ordering across geometries. This supports Eq. (6): policies differ through the evidence-shortfall term, not through the validity of the certificate itself.

Oracle references. Support-constrained directed-KL, Bhattacharyya, and symmetric-KL allocations are used as information-design references in small discrete worlds. They optimize their stated criteria on the empirical support and are not finite-budget upper bounds on diagnosis accuracy. After strong-baseline harmonization, greedy EIG + 10% exploration is statistically indistinguishable from the directed-KL reference at the primary $B = 10$ endpoint in the exact, near-hidden, and real-label oracle-proximity settings. Consequently the main paper does not use oracle proximity as evidence of unique SR-MKL optimality.

H WRENCH WEAK-LABEL SUPPORT STRESS TEST

Design and scope. We use binary YouTube, SMS, and IMDb datasets from WRENCH (Zhang et al., 2021). A historical logistic conditional-label channel is fit using the WRENCH training labels and weak-label votes. The validation and test weak-label states form a fixed deployment pool. On this real weak-label support we define five controlled candidate channels: reference, class-prior, source-accuracy, source-dependence, and combined accuracy/dependence logit shifts. Each candidate changes $P(Y \mid \Lambda)$ while leaving the deployment weak-label state counts unchanged. For every dataset and true channel, 100 trials share a finite pool, an item order, and a latent label bank across policies. These are generated conditional labels under designed shifts, *not observed WRENCH temporal drift or real future labels*. Accuracy is averaged over 500 dataset–mechanism–trial units per policy and budget. The strengthened EIG runs preserve the original candidate channels and shared trial seed namespaces; only EIG acquisition is rerun. Diagnostic budgets are $B \in \{0, 5, 10, 20, 50, 100\}$, without additional critic labels.

Table 3: WRENCH mechanism-identification accuracy (%). "Legacy EIG" uses the original temperature-softened sampler; "mult. EIG" is $q(o) \propto N(o)\,\mathrm{EIG}(o)$; greedy EIG uses 10% availability-proportional exploration. All policies share the five-channel mechanism family and paired latent label banks.

Dataset	B	Random	Legacy EIG	Mult. EIG	Greedy EIG	SR-MKL
IMDb	20	43.8	45.8	46.4	56.8	52.4
IMDb	100	64.2	65.6	76.0	91.4	86.6
SMS	20	33.0	33.2	43.2	59.8	58.6
SMS	100	49.8	52.0	66.8	71.0	70.6
YouTube	20	34.6	35.0	47.2	57.6	55.6
YouTube	100	53.8	56.4	70.2	70.8	70.4

Paired comparisons. At $B = 100$, SR-MKL minus greedy EIG accuracy is -4.8 points on IMDb (paired 95% bootstrap CI $[-7.8, -2.0]$), -0.4 on SMS ($[-2.8, 1.8]$), and -0.4 on YouTube ($[-2.6, 1.8]$). Paired bootstrap units match dataset, true mechanism, trial, and budget; the intervals are descriptive for this post-hoc benchmark audit. Greedy EIG improves markedly on the legacy EIG at both budgets in Table 3; this dataset study strengthens the evidence against a universal acquisition ranking, not for a new SR-MKL performance claim. The accompanying supplement contains all six budget endpoints, raw paired trials, the original experiment, the strong-EIG rerun notebook, and the contrast reconstruction script.

Independent model criticism. A separate critic study first runs the original Random, legacy EIG, or SR-MKL diagnostic policy, then draws *100 additional labels* uniformly from previously unqueried pool items. Absolute fit of the selected candidate is tested on these disjoint critic labels with a conditional Monte Carlo deviance null (2,000 null draws, rejection threshold 0.05). Among 500

in-family trials per dataset and policy, observed rejection ranges from 4.6% to 10.4%; rejection of a wrong selected candidate contributes to this aggregate. An out-of-family alternative perturbs the reference logit in a bounded direction based on weak-label disagreement and abstention. At the largest tested severity, SR-MKL-driven diagnosis followed by the independent critic rejects 99% of IMDb unknowns, but only 1% of SMS and 6% of YouTube unknowns (100 trials each). The alternative's nearest-family divergence on SMS/YouTube support is much smaller than on IMDb. Hence a separate critic sample prevents reuse of the diagnostic evidence, but *does not guarantee useful unknown-mechanism power* at the tested budget. This critic sweep used the original three diagnostic policies; it was not rerun for either stronger EIG policy.