WHERE DOES TASK IDENTITY LIVE? FLAT: A PROVENANCE-COMPLETE FIXED-LOG AUDIT FOR OFFLINE TASK IDENTIFICATION

Anonymous authors
Paper under double-blind review

ABSTRACT

A fixed offline log may lack task-identifying evidence, hold evidence an encoder discards, or carry task identity through task-dependent collection; encoder accuracy alone cannot tell these cases apart. We introduce FLAT (Fixed-Log Audit for Task Identification), which upper-bounds passive task identification on a fixed log relative to a declared generative factorization. Each observed factor is declared COMMON, CALIBRATED, or UNKNOWN before it is sampled, and the declaration must be complete, which the log cannot verify. FLAT stops before the first UNKNOWN factor and reveals the task on stopped episodes, which keeps the bound conservative; full likelihoods, likelihood ratios, or directional KL then give an informative ceiling, a vacuous ceiling, or no number. On reward-free PointRobot logs from 100 task-specific actors, the certified identification bracket rises from $[0.077, 0.088]$ at one step to $[0.993, 1]$ at 20 steps, whereas shared-policy logs whose reward-free factors are all COMMON have the exact prior-level ceiling 0.01. Learned identifiers recover task identity from the task-specific logs even without rewards, showing that collection itself can carry usable task signal. Across 48 search configurations (640 seeded runs), the audit maps informative and vacuous certificates; reached-stop RadSearch and a controlled stochastic Walker variant stay informative under partial modeling, and, under a uniform prior, a controlled family locates most slack in directional-KL summaries. The contribution is the fixed-log provenance formulation, its conservative treatment of unmodeled mechanisms, and certificates that use exactly the information the declaration supplies.

1 INTRODUCTION

Context-based offline meta-reinforcement learning (OMRL) infers a task representation from fixed multi-task data and conditions control on that inferred state (Rakelly et al., 2019; Mitchell et al., 2021; Li et al., 2021; Yuan and Lu, 2022; Zhou et al., 2024; Wang et al., 2024). A common way to evaluate a task encoder is to measure how accurately its representation identifies the underlying task. This accuracy mixes two quantities: how much task information the offline log contains, and how much of it the encoder recovers. Low accuracy may reflect a lack of task information in the log, and high accuracy may arise because task-dependent data collection makes the logged trajectories identifying.

In PointRobot, task-specific collection makes reward-free logs almost fully identifying: the certified 100-way bracket reaches $[0.993, 1]$ at 20 steps (Section 4.2). Shared-policy collection gives the exact prior-level ceiling 0.01 (Section 4.3). Learned reward-free identifiers reach well above chance on the task-specific logs and stay near chance on shared-policy logs. This source contrast motivates the audit question: how much task identity does the fixed log make available, and which logged factors carry it?

Answering this question requires knowing how the log was generated. A valid upper bound has to account for every observed mechanism that can carry task information, because omitting even one can invalidate it (Proposition 1). FLAT (Fixed-Log Audit for Task Identification) therefore asks the auditor to declare the generating mechanisms one by one: reset, behavior policy, transition, observation, reward when present, termination, and any domain-specific read. FLAT requires this

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.

declared observable generative factorization to be complete; completeness is an assumption about the declared model that the fixed log cannot establish statistically.

How the audit works. Before each factor is sampled, the auditor declares it COMMON (its conditional law is the same under every candidate), CALIBRATED (candidate likelihoods, likelihood ratios, or a conservative divergence bound are available), or UNKNOWN. COMMON factors cancel from every comparison between candidates and need no model. The audit stops before the first UNKNOWN factor and, on the episodes that reach it, gives the identifier the true task label in place of everything that follows; this can only help the identifier, so the bound remains valid. The information available before the stop then selects the certificate route (Figure 1). The audit ends with an *informative* ceiling below one, a *vacuous* ceiling equal to one, or a *non-admissible* case in which it reports no number because the declaration fails a required condition or a predeclared diagnostic check fires.

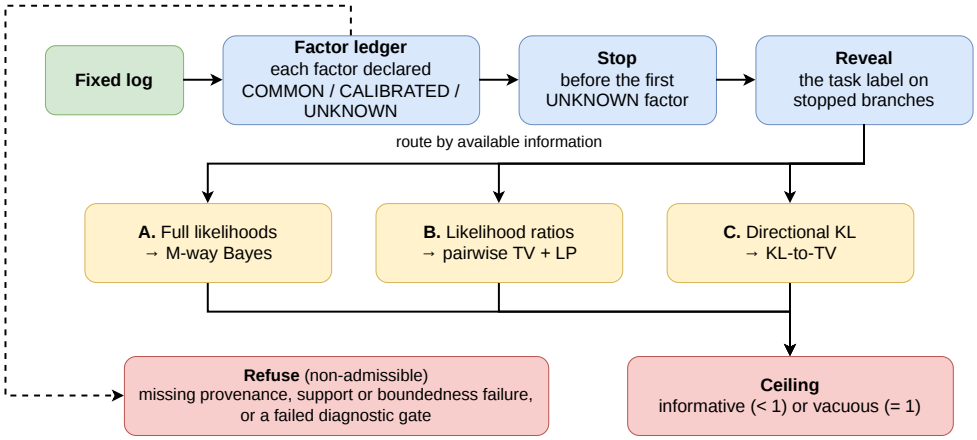


Figure 1: How FLAT audits a log. Each observed factor is declared before sampling. The audit stops before the first UNKNOWN factor and reveals the task label on stopped branches, so UNKNOWN factors are handled before any route is chosen. The information available on the retained prefix selects Route A, B (pairwise TV followed by the decision-image linear program, LP; Section 3.3), or C. The output is an informative or vacuous ceiling, or a refusal.

Contributions.

- **A provenance contract for fixed logs.** The audited object is the candidate-conditioned law of a fixed logged experiment, relative to a declared observable generative factorization. Completeness of this declaration is necessary for any nontrivial ceiling (Proposition 1), and the audit fails closed: it returns an informative ceiling, a vacuous ceiling, or no number.
- **Predictable stopping and information-dependent certificates.** First-UNKNOWN stopping with revealing completion yields an upper experiment that Blackwell-dominates the log (Theorem 1). The certificate then uses exactly the information the declaration supplies: direct M-way Bayes accuracy, stopped pairwise TV with decision-image aggregation, or a classical KL-to-TV transfer, with finite-sample endpoints that hold simultaneously over predeclared choices.
- **Evidence from OMRL and search logs.** PointRobot shows that task-dependent collection can make reward-free logs identifying and that learned identifiers can use this signal, while shared-policy logs stay at the prior level; reached-stop RadSearch and stochastic Walker cases remain informative under partial modeling; and a controlled family locates where certificate slack enters.

2 Related Work

Offline meta-RL and task identifiability. Task inference is a central component of meta-RL, as illustrated by PEARL and VariBAD, which infer task information to support adaptation and control (Rakelly et al., 2019; Zintgraf et al., 2021). In the offline setting, prior work addresses adaptation,

MDP ambiguity, context shift, and transition–reward shift (Mitchell et al., 2021; Dorfman et al., 2021; Gao et al., 2023; Wang et al., 2023). A large body of work improves the task representation itself, with objectives ranging from metric and contrastive learning to sequence modeling, flow-based inference, and causal disentanglement; examples include FOCAL, CORRO, GENTLE, Meta-DT, CERTAIN, ER-TRL, CTMRL, FLORA, and CATAL (Li et al., 2021; Yuan and Lu, 2022; Zhou et al., 2024; Wang et al., 2024; Zhou et al., 2025; Nakhaeinezhadfard et al., 2025; He et al., 2026; Wang et al., 2026; Cong et al., 2026). Closest to our question, BOReL studies identifiability and data collection at the MDP level (Dorfman et al., 2021); CORRO, CSRO, and related methods reduce the dependence of learned representations on the behavior policy or context distribution (Yuan and Lu, 2022; Gao et al., 2023); Jia et al. (2024) disentangle the behavior policy from the task representation by adversarial data augmentation; Qian et al. (2026) learn behavior-invariant task representations with a Transformer world model; and information-theoretic OMRL work studies the information a representation carries about the task (Li et al., 2024; Zhang et al., 2025).

These works show that collection can carry task identity into offline contexts, and they design data strategies or training objectives around it. FLAT turns this observation into a certification objective: bounding how much task identity an already collected log makes available to any passive identifier, under a declared provenance contract. The audited object is the candidate-conditioned law of the logged experiment, which every encoder of that transcript processes; the assumption is a complete declaration of the observed factors with route-specific calibration; and the guarantee is a one-sided finite-sample ceiling that holds for every encoder or identifier using the same transcript.

Bayes-error estimation. Supervised learning also compares classifiers with the best accuracy the data permit, by computing the Bayes error of learned generative models (Theisen et al., 2021) or estimating it from soft labels (Ishida et al., 2023). FLAT brings this comparison to logs whose generating process is only partly modeled and returns a one-sided certified ceiling.

Information flow, testing, and controlled sensing. Quantitative information flow supplies the posterior-vulnerability and channel-refinement viewpoint used by the full-likelihood route (Smith, 2009; Alvim et al., 2012). Classical hypothesis testing and change-of-measure arguments relate likelihood ratios, KL divergence, and total variation (Tsybakov, 2009; Kaufmann et al., 2016; Fedotov et al., 2003). Controlled sensing chooses actions in order to acquire information (Nitinawarat et al., 2013), as do POMDP planning and source-search methods (Kaelbling et al., 1998; Silver and Veness, 2010; Vergassola et al., 2007; Proctor et al., 2021); FLAT is passive and uses these tools to audit a log that has already been collected. Our contribution is the provenance contract that decides when they apply to a fixed log, the conservative treatment of UNKNOWN factors, and fail-closed refusal.

3 Auditing the Provenance of a Fixed Log

3.1 Declaring the factors of a log

There are M candidate tasks. Let $Y \in [M]$ be the task identity, which the identifier does not observe, with declared prior $\pi = (\pi_1, \dots, \pi_M)$. Let Z be the observed context available to a passive identifier, which reads the logged transcript but cannot collect new data, and write $P_i = \mathcal{L}(Z \mid Y = i)$ for its law under task i. For any observed experiment W, $\mathrm{Acc}^\star(W)$ denotes its Bayes-optimal 0–1 task-identification accuracy, that is, the highest accuracy any identifier can reach from W.

The audit describes the log as a sequence of observed factors. Conditional on $Y = i$, the logged experiment is a finite, or almost surely finite, sequence

$$Z = (U_1, \dots, U_R), \qquad U_r \mid H_{r-1}, Y = i \sim K_{i,r}(\cdot \mid H_{r-1}), \tag{1}$$

where $H_{r-1} = (U_1, \dots, U_{r-1})$ is the history before factor r. A factor can be a reset variable, a policy action, a transition outcome, an observation, a reward, a termination indicator, or a domain-specific sensor read, and the kernel $K_{i,r}$ describes how it is generated under task i. This factorization is part of the audit contract: it must account for every observed factor, including any mechanism that depends on the task.

Before U_r is sampled, the audit assigns its factor one of three statuses, using only the history available at that time:

COMMON $K_{i,r}(\cdot \mid H_{r-1})$ is the same for every audited candidate, up to null histories (Appendix A.2), so the factor adds no task information beyond the history;

CALIBRATED candidate likelihoods, likelihood ratios, or a conservative directional-divergence bound are available;

UNKNOWN no claim is made about how the factor depends on the task.

The status rule must be *predictable*: it may depend on the past, but not on the value about to be drawn. Classifying a factor after its value is seen would let the audit keep favorable values and set aside unfavorable ones, invalidating the bound.

For a candidate family, or for a pair of candidates, let σ be the index of the first factor declared UNKNOWN, with $\sigma = R + 1$ if no such factor is reached. The audit stops just before U_σ is sampled and keeps the prefix $S = (H_{\sigma-1}, \sigma)$, that is, the history up to the stop together with the stopping index. Let $E = \{\sigma \le R\}$ be the event that the audit reaches an UNKNOWN factor; on E^c the whole transcript is kept.

Proposition 1 (Completeness of the declaration is necessary)**.** *Fix any declaration of the factors $U_1, \ldots, U_R$, including their statuses and kernels. Appending one further observed factor $V = Y$ yields a log that agrees with the declaration on every declared factor and has $\mathrm{Acc}^\star(Z, V) = 1$. Hence no procedure that uses the declared factors alone can return a ceiling below one that is valid for every log that agrees with the declaration on the declared factors.*

An omitted observed factor is therefore grounds for refusal (Appendix A).

3.2 STOP, REVEAL, AND THE DIRECT ROUTE

The stop leaves one question open: what should the audit assume about the part of the log it did not examine? FLAT assumes the case that is worst for the ceiling, namely that this part reveals the task completely.

Define the *revealing completion* $\bar{Z}$ as follows. On E^c, $\bar{Z}$ is the complete logged transcript. On E, $\bar{Z}$ consists of the stopped prefix S together with the true label Y. The original log can be reconstructed in distribution from $\bar{Z}$: on E^c nothing needs to be done, and on E the omitted continuation can be drawn from its conditional law given the stopped history and the revealed label. The original experiment is therefore a garbling of $\bar{Z}$, so $\bar{Z}$ dominates it in the Blackwell order (Blackwell, 1953), and no identifier can do better on Z than on $\bar{Z}$. We assume standard Borel factor spaces (or sufficient measurable structure) and measurable kernels $K_{i,r}$, ensuring the regular conditional laws needed to reconstruct the omitted continuation.

Theorem 1 (Fixed-log upper experiment under a complete audited factorization)**.** *Under the factorization in Equation 1, the standard-Borel/measurability conditions above, and predictable first-unknown stopping,*

$$\mathrm{Acc}^\star(Z) \;\le\; \mathrm{Acc}^\star(\bar{Z}) \;=\; \mathbb{E}\left[\max_{i \in [M]} \mathbb{P}(Y = i \mid \bar{Z})\right]. \tag{2}$$

If the candidate likelihoods on every non-revealed branch are available, then

$$\mathrm{Acc}^\star(\bar{Z}) = \mathbb{P}(E) + \mathbb{E}\left[\mathbf{1}\{E^c\} \max_i \frac{\pi_i L_i(Z)}{\sum_j \pi_j L_j(Z)}\right], \tag{3}$$

where L_i is the density of the fully calibrated transcript under candidate i with respect to a common dominating measure (Appendix A.2), defined up to any task-common factor. In particular, likelihood contributions from COMMON *factors cancel from the posterior and need not be evaluated candidate-by-candidate. If all retained factors are* COMMON *and E never occurs, then $\mathrm{Acc}^\star(Z) = \max_i \pi_i$.*

Equation 2 states that the ceiling is the expected largest posterior probability under $\bar{Z}$, and Equation 3 splits it in two: episodes that reach an UNKNOWN factor contribute $\mathbb{P}(E)$, and the rest contribute the largest posterior computed from candidate likelihoods. This is Route A, the direct M-way route. It needs likelihoods only for the task-dependent factors retained before the stop, because COMMON factors cancel; in the task-specific PointRobot audit, for example, only the candidate policy likelihoods are evaluated (Section 4.2). A large $\mathbb{P}(E)$ pushes the ceiling toward one, which is how a ceiling becomes vacuous; if every factor is COMMON and no UNKNOWN factor is reached, identification stays at $\max_i \pi_i$.

Example. Consider a reward-free log whose steps record a reset, an action, a transition, and, only near a candidate goal, a sensor reading. If reset and transition are COMMON, the behavior policy's action likelihoods are CALIBRATED for every candidate, and the sensor mechanism is UNKNOWN, the audit keeps reset, action, and transition factors, stops before the first sensor reading, and replaces that reading and everything after it by the task label. By Equation 3, the ceiling is $\mathbb{P}(E)$ plus the expected largest posterior, computed from action likelihoods alone, on episodes that never reach a goal neighborhood. The ceiling falls below one only when these unrevealed transcripts retain task ambiguity under the action likelihoods, and it is low only when, in addition, few episodes reach a goal neighborhood; if every episode reaches the sensor, the ceiling is one.

3.3 PAIRWISE TV AND KL ROUTES

Full candidate likelihoods are not always available. The two remaining routes bound each pair of candidates separately and then combine the pairs into an M-way ceiling. Throughout, $\mathrm{TV}(P, Q) = \sup_A |P(A) - Q(A)|$.

From a stopped pair to a full pair. For a pair (i, j), let S_{ij} be its stopped transcript and E_{ij} the event that it reaches its first UNKNOWN factor, and write $p_{i,ij} = P_i(E_{ij})$ and $p_{j,ij} = P_j(E_{ij})$ for the probability of that event under each candidate. A continuation coupling gives

$$\mathrm{TV}(P_i, P_j) \le \min\Big\{1,\, p_{j,ij} + \mathrm{TV}(P_i^{S_{ij}}, P_j^{S_{ij}}),\, p_{i,ij} + \mathrm{TV}(P_i^{S_{ij}}, P_j^{S_{ij}})\Big\}. \tag{4}$$

In words, the distance between the full logs is at most the distance between the stopped prefixes plus the probability of reaching the stop. The proof replaces one candidate's continuation by the other's (Appendix A).

Route B: likelihood ratios. If the stopped likelihood ratio is available, the stopped TV can be computed directly. Under domination,

$$\mathrm{TV}(P_i^{S_{ij}}, P_j^{S_{ij}}) = \mathbb{E}_i\left[\left(1 - \frac{dP_j^{S_{ij}}}{dP_i^{S_{ij}}}\right)_+\right]. \tag{5}$$

We evaluate the ratio in log space, as $1 - \exp(-\ell)$ with $\ell = \log\big(dP_i^{S_{ij}}/dP_j^{S_{ij}}\big)$, to improve numerical stability over long horizons. A non-finite log-ratio makes the requested endpoint non-admissible.

Route C: directional KL. If only a directional information bound is available, define

$$I_{i\to j} = \mathbb{E}_i\left[\sum_{r<\sigma_{ij}} D_{\mathrm{KL}}(K_{i,r}(\cdot \mid H_{r-1}) \,\|\, K_{j,r}(\cdot \mid H_{r-1}))\right]. \tag{6}$$

By the chain rule, $I_{i\to j}$ equals the KL divergence between the two stopped laws (Appendix A). Let $\psi_{\mathrm{KL}}(D)$ be the sharp classical upper bound on TV implied by $D_{\mathrm{KL}}(P\|Q) \le D$, obtained by inverting Vajda's tight KL–TV boundary (Fedotov et al., 2003). Equation 4 remains valid after replacing the stopped-TV term in each direction by $\psi_{\mathrm{KL}}(I_{i\to j})$ or $\psi_{\mathrm{KL}}(I_{j\to i})$; Pinsker's inequality, $\mathrm{TV} \le \sqrt{D/2}$, is a simpler but weaker alternative. Route C needs no likelihood of the whole transcript: an upper bound on the directional KL of each retained factor, e.g. from a declared sensor channel, suffices for the population bound, and its finite-sample endpoint additionally needs a cap on per-episode information (Section 3.4).

Combining pairs into an M-way ceiling. For $M > 2$, the pair bounds can be combined without certifying all $M(M-1)/2$ pairs. Let $q_i = (q_{i1}, \ldots, q_{iM})$ be the output distribution of an arbitrary identifier when the true task is i. Data processing through the identifier gives $\mathrm{TV}(q_i, q_j) \le \mathrm{TV}(P_i, P_j)$, so every certified pair bound δ_{ij} implies $\sum_k |q_{ik} - q_{jk}| \le 2\delta_{ij}$. The best accuracy allowed by these constraints is the value of a linear program over identifier output laws, which we call the decision-image program,

$$\max_{q \ge 0} \quad \sum_i \pi_i q_{ii} \tag{7}$$

$$\text{s.t.} \quad \sum_k q_{ik} = 1, \qquad \sum_k |q_{ik} - q_{jk}| \le 2\delta_{ij} \quad \text{for certified edges } (i, j), \tag{8}$$

which therefore upper-bounds optimal M-way identification. When only a sparse, predeclared, connected graph of pairs is certified, the triangle inequality along paths bounds the remaining pairs. When every pair has the same bound $\mathrm{TV}(P_i, P_j) \le \delta$, the value of the program is $\max_i \pi_i + (1 - \max_i \pi_i)\delta$ for $\delta \le 1$ (Appendix A.5); under a uniform prior, which every search-domain cell in Section 4.1 uses, this is

$$\mathrm{Acc}^\star(Z) \le \frac{1}{M} + \left(1 - \frac{1}{M}\right)\delta. \tag{9}$$

The program is a valid upper bound but need not equal Bayes accuracy, since pairwise TV does not determine M-way accuracy in general (Appendix A.5 gives a three-task example).

3.4 Estimating the bounds from samples

The quantities above are population values. From a finite log, FLAT reports one-sided confidence endpoints instead. For Route A, each episode contributes a score in $[0, 1]$, namely its largest posterior probability under revealing completion. With independent episodes from each task, a weighted Hoeffding bound over these scores gives a one-sided upper endpoint with confidence at least $1 - \alpha$, clipped at one (Appendix B). Route B can use any valid one-sided bounded-mean endpoint, because its per-sample quantity also lies in $[0, 1]$. Route C needs a guaranteed upper limit on how much information a single episode can carry, or another declared valid endpoint; without one, the endpoint is non-admissible. For localized reads, the probability of reaching the stop is bounded by simultaneous DKW bands (Massart, 1990).

The audit usually chooses among a finite family of options declared in advance, such as margins, certified pairs, or horizons. Endpoints are formed simultaneously over that family using a union bound, so choosing the best option afterwards keeps the stated confidence level. Table 2 in Appendix B.4 lists what each route needs and when it is vacuous or refused.

Predictable refinement robustness. Splitting a fully calibrated factor into a finer predictable sequence, or coarsening such a sequence, preserves the population directional information by the KL chain rule; only finite-sample efficiency can change. When UNKNOWN factors are present, refinement remains safe if the predeclared stopping family retains the original coarse stop: the best eligible population certificate cannot worsen solely because a finer description is available. Factor classification must remain predictable and cannot depend on the value being sampled (Appendix A.8).

3.5 Two special cases: localized reads and shared-policy logs

Localized reads. In a search task, the agent moves and takes sensor reads, and a read is informative about the task mainly when it is taken close to the task's source. Suppose non-read factors are COMMON, reads taken far from the candidates are CALIBRATED, and the first read that enters a neighborhood of radius d around a candidate pair is UNKNOWN. Let $p_i(d)$ be the probability under task i of reaching such a read, which we call the exposure. Assume at most N reads almost surely, and let $D_{i\to j}(d)$ bound the directional KL of a single read taken outside the neighborhood. Then $I_{i\to j}(d) \le N D_{i\to j}(d)$, and Equation 4 yields

$$\mathrm{TV}(P_i, P_j) \le \min\Big\{1,\ p_j(d) + \psi_{\mathrm{KL}}\big(N D_{i\to j}(d)\big),\ p_i(d) + \psi_{\mathrm{KL}}\big(N D_{j\to i}(d)\big)\Big\}. \tag{10}$$

The margin d sets a trade-off: a larger d makes the remaining far reads less informative but makes the neighborhood easier to reach. In the search domains of Section 4.1, N is the protocol's almost-sure read limit and $D_{i\to j}(d)$ comes from the declared read channel, with simultaneous endpoints for exposure and for the channel.

Shared-policy reward-free logs. The last statement of Theorem 1 gives the following structural consequence (Appendix A).

Corollary 1 (Shared-policy reward-free logs)**.** *If reset, behavior policy, physical transition, non-reward observation, and termination or truncation are all* COMMON *across tasks, the reward-free log has the same law under every candidate, every pair TV is zero, and* $\mathrm{Acc}^\star(Z) = \max_i \pi_i$*, which equals* 0.01 *under a uniform prior over 100 tasks.*

Table 3 in Appendix C lists the declarations used for the three kinds of logs studied in Section 4.

4 EXPERIMENTS

Experiments answer four questions: where task identity enters (Sections 4.2–4.3), whether partial modeling can remain informative (Sections 4.1 and 4.3), how broadly the audit behaves across search settings (Section 4.1), and where certificate slack enters (Section 4.4). Main experiments use a uniform prior, and candidate choices and confidence families are declared in advance; Appendices C–E give the protocols.

4.1 SEARCH DOMAINS: BREADTH AND REACHED STOPS

We apply the localized-read bound of Equation 10 to OTTO olfactory source tracking (Loisy and Heinonen, 2023), RadSearch radiation-source search (Proctor et al., 2021), and RockSample (Smith and Simmons, 2004). A task is the location of a hidden source among M candidates (in RockSample, which of M rocks is good). In these domains the first unrestricted task-dependent read defines the UNKNOWN boundary, so the audit combines predictable stopping with Route C. Each domain is audited on a factorial of four channel-strength levels, four candidate counts, and 10 seeds (20 for RockSample), giving 160, 160, and 320 cells (Appendix C.1). Contexts are collected by a random-walk controller that never reads the task or any sensor value, so exposure has the same law under every task and all contexts of a cell share one DKW band (Appendix C).

Table 1: Search-domain audit over 48 configurations and 640 seeded runs (uniform prior, 15 contexts per task, $\alpha = 0.05$). *Informative*: a valid ceiling below one. *Vacuous*: a valid ceiling clipped at one. *Non-admissible*: the audit refuses to report a ceiling. The last column gives the range of informative ceilings; chance $1/M$ lies between 0.06 and 0.25.

Domain	Total	Informative	Vacuous	Non-admissible	Informative ceilings
OTTO	160	111	37	12	0.43–0.96
RadSearch	160	74	74	12	0.36–0.80
RockSample	320	156	156	8	0.27–0.83
Total	640	341	267	32	0.27–0.96

Among the 608 admissible runs, 341 are informative and 267 vacuous. In every domain, weaker sensor levels give informative ceilings and stronger levels saturate them, mapping how channel strength changes audit usefulness across candidate counts. The 32 gate refusals come from 8 task-blind settings and are analyzed in Appendix C.3.

Reached-stop RadSearch. A separate RadSearch design widens the random walk. In 7/10 confirmation seeds, 16.0–18.8% of contexts reach the selected margin; in those seeds the first UNKNOWN read is reached after 6–9 calibrated reads and the informative ceilings are 0.876–0.897. In the remaining 3/10 seeds no context reaches the selected margin and the ceiling is 0.913. A Poisson Bayes decoder reaches 0.265–0.323 against chance 0.25; these runs establish non-vacuity under reached stops, while tightness is studied separately in Section 4.4 (Appendix C.5).

4.2 TASK-SPECIFIC COLLECTION CARRIES TASK IDENTITY

We ask how much task identity reward-free logs carry when each task is collected by its own actor, and how this grows with context length. We use 100 PointRobot tasks with goals equally spaced on a half-circle, one frozen actor per task, 1000 independent episodes per task, horizon 20, and prefixes $H \in \{1, 2, 4, 8, 12, 20\}$. Reset, transition (the next state is the current state plus the action), and termination are COMMON; the behavior policy is CALIBRATED because every candidate's policy likelihood is known. The audit therefore uses Route A with the 100 candidate policy likelihoods alone.

The certificate brackets the Bayes accuracy on the logged state/action transcript. The upper endpoint evaluates every episode under all 100 candidate policy likelihoods on a richer transcript that includes the generation latents; the logged transcript is a deterministic function of that one, so its Bayes accuracy can only be lower (Corollary 2). The lower endpoint is the accuracy of a fixed identifier that

uses candidate likelihoods but sees only the logged pre-action states and float32 actions. The 0.05 error budget is split equally over the twelve jointly covered endpoints (Appendix D); the full Route A certificate for 100 tasks and 10^5 episodes takes about 75 s on one machine (Appendix G).

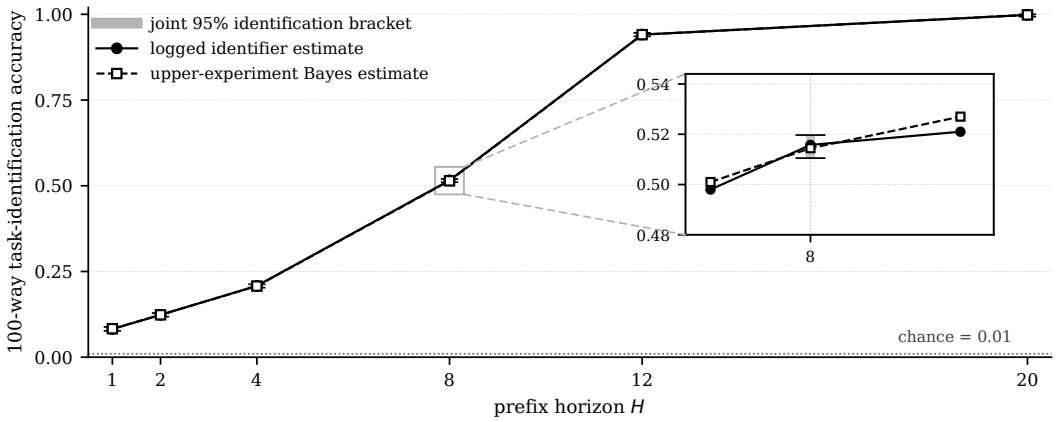


Figure 2: Direct M-way identification on reward-free PointRobot logs from task-specific actors (Family A). Vertical intervals are the jointly valid 95% identification brackets at the six prespecified horizons. The two curves are finite-sample point estimates of the logged identifier and of the upper experiment; their ordering is not part of the certificate. Chance is 0.01; exact values are in Table 7.

Identification rises from about 0.08 at $H = 1$, with bracket $[0.077, 0.088]$, to essentially one at $H = 20$, with bracket $[0.993, 1]$ (Figure 2). Most of the rise happens between $H = 4$ and $H = 12$. Because the lower endpoint is the accuracy of an actual identifier, every bracket pins the Bayes accuracy of the logged transcript to within about 0.01: with full likelihoods, the ceiling is essentially tight. Even without rewards, the logged states and actions identify the task among 100 candidates within 20 steps. Across three training seeds, mean reward-free identification at $H = 20$ reaches 0.252 (range 0.107–0.499) for a FOCAL-style encoder and 0.661 (range 0.632–0.685) for DeepSets on newly collected logs from the same task-specific actors, while evaluations on shared-policy logs stay near chance (0.01; Appendix D.3).

4.3 Shared-policy logs stay at the prior level

We next audit logs in which one policy collects data for every task. In the PointRobot logs, each of three shared actors is run on all 100 tasks, and reset, execution policy, physical transition, non-reward observation, and termination are all COMMON; the HalfCheetah-Vel shared-policy logs satisfy the same declaration. By Corollary 1, the pair TV is exactly zero and the ceiling is exactly 0.01 in both domains. Frozen classifier probes on these logs stay near chance and provide complementary consistency checks (Appendix D). Together with Section 4.2, this gives a source contrast relative to the declared factorization: logs from task-specific actors become nearly perfectly identifying, while shared-policy logs with common reward-free factors admit only the prior level.

Two Walker-based constructions show how available transition calibration changes the audit outcome. With no calibrated transition law, the first transition is UNKNOWN and the ceiling is one; in a controlled stochastic variant with calibrated upright transitions, 16.7–19.5% of episodes later reach UNKNOWN and all 10 confirmation seeds give ceilings of 0.832–0.847, versus a full-transcript oracle of 0.782–0.816 (Appendix D.4)..

4.4 Where certificate slack comes from

Controlled family. A four-task family with binary states and actions varies whether the behavior policy and transition dynamics are shared or task-specific and calibrated, while keeping Bayes accuracy exactly computable (Appendix E.1). When both factors are task-specific, exact Bayes accuracy is 0.349507, and feeding the exact pair TVs to the decision-image program reproduces it to

numerical precision (residual below 4×10^{-15}). Feeding exact directional KL values directly to the output-law program gives 0.372007, and the operational Route C ceiling is 0.372070: of the total gap of 0.022563, only 6.4×10^{-5}, under 0.3%, comes from the final scalar KL-to-TV transfer. A prespecified criterion that asked the direct-KL diagnostic to recover at least half of this slack therefore fails (Appendix E.1). The two other regimes with nonzero slack show the same decomposition, and with policy and dynamics both shared every value equals the prior maximum 0.25 (Table 10). Under the uniform prior used in this decomposition, reducing pair laws to directional-KL summaries accounts for nearly all slack; the exact agreement of the pair-TV program is a property of this family (Appendix A.5).

Information ladder on a fixed PointRobot log. On the task-specific PointRobot log, likelihood ratios on a chain of 99 adjacent-goal pairs (Route B) stay within 0.0004 of the full-likelihood route before confidence adjustment, directional-KL summaries raise the plug-in ceiling by up to 0.124, and certified Route B ceilings are up to 0.060 higher because of confidence allocation; three of five prespecified conditions pass, so the comparison is diagnostic (Appendix E.2).

5 Discussion and Limitations

Encoder accuracy mixes together three different questions: how much task information is available in the logged data, which parts of the data-collection process put that information there, and how much of it a learned representation keeps. PointRobot makes this distinction concrete. Task-specific collection produces reward-free logs that become almost fully identifying, and learned identifiers can use that signal; under shared-policy collection, the corresponding reward-free logs stay at the prior level. The representation study in Appendix F shows the other side: task information can be present in the raw context but be substantially reduced by a learned embedding.

The partial-modeling experiments clarify what an UNKNOWN factor means in practice. Reaching UNKNOWN does not automatically make the certificate vacuous. RadSearch and the controlled stochastic Walker variant remain informative when enough calibrated evidence is available before the stop. The ceiling therefore depends not only on whether some mechanism is unmodeled, but also on where that mechanism appears and how much task information comes before it. The stop-and-reveal construction keeps this earlier evidence while treating everything after the stop conservatively.

The tightness studies show where looseness enters the certificate. In the controlled family, exact pairwise TVs reproduce Bayes accuracy to numerical precision, whereas directional-KL summaries account for nearly all of the observed gap under the uniform prior; on the fixed PointRobot log, finite-sample confidence adjustment adds further slack. The guarantee also remains relative to the declared factorization and route conditions: a fixed log cannot establish that the declaration is complete, and coarse calibration can leave a nonzero population floor even as sample size grows. For large candidate families, dense pairwise aggregation can also become expensive, motivating sparse pair graphs when the important confusions are local.

Our future work will extend the same audit objective to continuous task spaces, improve scalable aggregation for large candidate families, and develop calibration methods that preserve more of the information available in the logged data.

6 Conclusion

FLAT studies task identifiability at the level of the fixed log rather than through one particular encoder. The declared factorization determines which observed factors are shared across tasks, which can be calibrated, and where the audit must stop. The resulting certificate answers a question that encoder accuracy alone cannot: how much task identity the logged data itself makes available under the declared data-generating process.

The experiments show why this distinction matters. Task-dependent collection can make reward-free trajectories strongly identifying, shared-policy collection can remove that source of identity, and partial modeling can still give informative certificates when useful calibrated evidence appears before an UNKNOWN factor. Together, these results show that understanding where task information enters the log is important for interpreting what an offline task representation has actually learned.

REPRODUCIBILITY STATEMENT

The anonymous supplement lets reviewers check file hashes, rebuild every table from saved records, and recompute the main results. The PointRobot collection code regenerates the 100-task log bitwise, and the 640 search-domain certificates can be recomputed from stored sufficient statistics without rerunning the simulators. For the released pinned HalfCheetah actors, re-collection of three tasks per actor reproduces all 2,400 checked panel files bit-for-bit on the recorded CPU stack; actor retraining does not reproduce the historical checkpoint bytes. Appendix H lists which claims each check covers.

ETHICS STATEMENT

All experiments are conducted in simulation and use no human subjects or personal data. The main practical risk is scientific misinterpretation: task identity carried by a behavior policy or other collection factor can be mistaken for recovery of task semantics. The proposed audit is intended to make such provenance and side channels explicit.

AI USE STATEMENT

We used generative AI tools to support manuscript preparation, literature discovery, research development, mathematical reasoning, and code implementation. Their specific roles are described below.

Writing and Presentation. AI assisted with drafting and revising manuscript text, organizing theoretical and empirical material, improving wording and readability, and checking LaTeX formatting and presentation consistency.

Retrieval and Discovery. AI supported literature discovery, related-work comparison, and preliminary novelty assessment. References and bibliographic information used in the manuscript were checked against the corresponding source publications.

Research Ideation and Execution. AI assisted with exploring research directions, refining methodological ideas, implementing and debugging research code, designing experiments, and interpreting empirical results. Experimental outputs were checked by rerunning the relevant scripts and comparing the reported quantities with the saved experiment outputs.

Mathematical Claims and Proofs. AI assisted with formulating and refining mathematical claims, developing proof strategies, checking assumptions, and identifying potential gaps. The resulting claims and proofs were subsequently checked against their stated assumptions and re-examined by the authors.

Synthetic Data. Generative AI was not used to generate synthetic datasets. All experimental data were obtained through the research procedures described in the paper.

AI-assisted mathematical arguments were checked against their assumptions; experimental code and figure-generation scripts were rerun against the saved outputs; and cited literature was checked against the corresponding source publications. The authors take responsibility for the final manuscript, results, proofs, and code.

REFERENCES

Mário S. Alvim, Konstantinos Chatzikokolakis, Catuscia Palamidessi, and Geoffrey Smith. Measuring information leakage using generalized gain functions. In *IEEE Computer Security Foundations Symposium*, pages 265–279, 2012.

David Blackwell. Equivalent comparisons of experiments. *The Annals of Mathematical Statistics*, 24(2):265–272, 1953.

Shan Cong, Chao Yu, and Xiangyuan Lan. CATAL: Causally disentangled task representation learning for offline meta-reinforcement learning. In *AAAI*, 40(25):20634–20641, 2026.

Ron Dorfman, Idan Shenfeld, and Aviv Tamar. Offline meta reinforcement learning—identifiability challenges and effective data collection strategies. In *NeurIPS*, 34:4607–4618, 2021.

Alexei A. Fedotov, Peter Harremoës, and Flemming Topsøe. Refinements of Pinsker's inequality. *IEEE Transactions on Information Theory*, 49(6):1491–1498, 2003.

Yunkai Gao, Rui Zhang, Jiaming Guo, Fan Wu, Qi Yi, Shaohui Peng, Siming Lan, Ruizhi Chen, Zidong Du, Xing Hu, Qi Guo, Ling Li, and Yunji Chen. Context shift reduction for offline meta-reinforcement learning. In *NeurIPS*, 36, 2023.

Hongcai He, Zetao Zheng, Anjie Zhu, Deqiang Ouyang, and Jie Shao. Improving generalization in offline meta-reinforcement learning via cross-task contexts. In *AAAI*, 40(26):21637–21644, 2026.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

Takashi Ishida, Ikko Yamane, Nontawat Charoenphakdee, Gang Niu, and Masashi Sugiyama. Is the performance of my deep network too good to be true? A direct approach to estimating the Bayes error in binary classification. In *ICLR*, 2023.

Chengxing Jia, Fuxiang Zhang, Yi-Chen Li, Chen-Xiao Gao, Xu-Hui Liu, Lei Yuan, Zongzhang Zhang, and Yang Yu. Disentangling policy from offline task representation learning via adversarial data augmentation. In *AAMAS*, pages 944–953, 2024.

Leslie Pack Kaelbling, Michael L. Littman, and Anthony R. Cassandra. Planning and acting in partially observable stochastic domains. *Artificial Intelligence*, 101(1–2):99–134, 1998.

Emilie Kaufmann, Olivier Cappé, and Aurélien Garivier. On the complexity of best-arm identification in multi-armed bandit models. *Journal of Machine Learning Research*, 17(1):1–42, 2016.

Lanqing Li, Rui Yang, and Dijun Luo. FOCAL: Efficient fully-offline meta-reinforcement learning via distance metric learning and behavior regularization. In *ICLR*, 2021.

Lanqing Li, Hai Zhang, Xinyu Zhang, Shatong Zhu, Yang Yu, Junqiao Zhao, and Pheng-Ann Heng. Towards an information theoretic framework of context-based offline meta-reinforcement learning. In *NeurIPS*, 37, 2024.

Aurore Loisy and Robin A. Heinonen. Deep reinforcement learning for the olfactory search POMDP: a quantitative benchmark. *The European Physical Journal E*, 46(3):17, 2023.

Pascal Massart. The tight constant in the Dvoretzky–Kiefer–Wolfowitz inequality. *The Annals of Probability*, 18(3):1269–1283, 1990.

Eric Mitchell, Rafael Rafailov, Xue Bin Peng, Sergey Levine, and Chelsea Finn. Offline meta-reinforcement learning with advantage weighting. In *ICML*, pages 7780–7791, 2021.

Mohammadreza Nakhaeinezhadfard, Aidan Scannell, and Joni Pajarinen. Entropy regularized task representation learning for offline meta-reinforcement learning. In *AAAI*, 39(18):19616–19623, 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.

Philippe Proctor, Christof Teuscher, Adam Hecht, and Marek Osiński. Proximal policy optimization for radiation source search. *Journal of Nuclear Engineering*, 2(4):368–397, 2021.

Fuyuan Qian, Menglong Zhang, Song Wang, and Quanying Liu. Behavior-invariant task representation learning with Transformer-based world models for offline meta-reinforcement learning. In *ICML*, 2026.

Kate Rakelly, Aurick Zhou, Chelsea Finn, Sergey Levine, and Deirdre Quillen. Efficient off-policy meta-reinforcement learning via probabilistic context variables. In *ICML*, pages 5331–5340, 2019.

David Silver and Joel Veness. Monte-Carlo planning in large POMDPs. In *NeurIPS*, pages 2164–2172, 2010.

Geoffrey Smith. On the foundations of quantitative information flow. In *FoSSaCS*, pages 288–302, 2009.

Trey Smith and Reid Simmons. Heuristic search value iteration for POMDPs. In *UAI*, pages 520–527, 2004.

Ryan Theisen, Huan Wang, Lav R. Varshney, Caiming Xiong, and Richard Socher. Evaluating state-of-the-art classification models against Bayes optimality. In *NeurIPS*, 34:9367–9377, 2021.

Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer Series in Statistics. Springer, 2009.

Massimo Vergassola, Emmanuel Villermaux, and Boris I. Shraiman. Infotaxis as a strategy for searching without gradients. *Nature*, 445:406–409, 2007.

Jianhao Wang, Jin Zhang, Haozhe Jiang, Junyu Zhang, Liwei Wang, and Chongjie Zhang. Offline meta reinforcement learning with in-distribution online adaptation. In *ICML*, pages 36626–36669, 2023.

Zhi Wang, Li Zhang, Wenhao Wu, Yuanheng Zhu, Dongbin Zhao, and Chunlin Chen. Meta-DT: Offline meta-RL as conditional sequence modeling with world model disentanglement. In *NeurIPS*, 37, 2024.

Min Wang, Xin Li, Mingzhong Wang, and Hasnaa Bennis. Offline meta-reinforcement learning with flow-based task inference and adaptive correction of feature overgeneralization. In *AAAI*, 40(31):26390–26397, 2026.

Haoqi Yuan and Zongqing Lu. Robust task representations for offline meta-reinforcement learning via contrastive learning. In *ICML*, pages 25747–25759, 2022.

Hai Zhang, Boyuan Zheng, Tianying Ji, Jinhang Liu, Anqi Guo, Junqiao Zhao, and Lanqing Li. Scrutinize what we ignore: Reining in task representation shift of context-based offline meta reinforcement learning. In *ICLR*, 2025.

Renzhe Zhou, Chen-Xiao Gao, Zongzhang Zhang, and Yang Yu. Generalizable task representation learning for offline meta-reinforcement learning with data limitations. In *AAAI*, pages 17132–17140, 2024.

Hongtu Zhou, Ruiling Yang, Yakun Zhu, Haoqi Zhao, Hai Zhang, Di Zhang, Junqiao Zhao, Chen Ye, and Changjun Jiang. CERTAIN: Context uncertainty-aware one-shot adaptation for context-based offline meta reinforcement learning. In *ICML*, pages 79529–79547, 2025.

Luisa Zintgraf, Sebastian Schulze, Cong Lu, Leo Feng, Maximilian Igl, Kyriacos Shiarlis, Yarin Gal, Katja Hofmann, and Shimon Whiteson. VariBAD: Variational Bayes-adaptive deep RL via meta-learning. *Journal of Machine Learning Research*, 22(289):1–39, 2021.

A Proofs and Structural Details

A.1 Proof of Theorem 1

Proof. On E^c, the revealing completion equals the complete observed transcript. On E, it contains the stopped prefix and the true task label. The standard-Borel and measurability assumptions of Section 3.2 guarantee a regular conditional kernel for the continuation. From $\bar{Z}$, construct a draw from the original experiment by leaving the transcript unchanged on E^c and, on E, sampling the remaining factors from this conditional continuation given the stopped history and the revealed Y. The original log is therefore a stochastic post-processing, that is, a Blackwell garbling, of $\bar{Z}$, so $\bar{Z}$ Blackwell-dominates the original experiment. Bayes-optimal 0–1 identification cannot improve under post-processing, which gives the inequality in Equation 2.

For a finite candidate family, the Bayes-optimal identification accuracy of any experiment W is the posterior vulnerability $\mathbb{E}[\max_i \mathbb{P}(Y = i \mid W)]$. On E, the revealing completion contains Y, so its largest posterior is one. On E^c, every observed factor is COMMON or CALIBRATED, and the posterior follows from the candidate likelihoods by Bayes' rule, which gives Equation 3; task-common factors multiply every L_i by the same quantity and cancel. If all factors are COMMON and E never occurs, W is independent of Y, the posterior remains π, and the accuracy is $\max_i \pi_i$. □

A.2 Conventions and an auxiliary upper experiment

Let

$$\mu = \frac{1}{M} \sum_{j=1}^{M} P_j,$$

which dominates every candidate law of the retained experiment, and let $L_i = dP_i/d\mu$. If $L_i = 0$ on a transcript, candidate i has posterior zero there, so differences in support are handled by Bayes' rule without a special case. For factor r, let

$$\mu_{r-1} = \frac{1}{M} \sum_{j=1}^{M} \mathcal{L}(H_{r-1} \mid Y = j)$$

be the corresponding mixture law of the history. A factor is COMMON when its kernels agree μ_{r-1}-almost everywhere; histories null under every candidate do not affect any candidate law.

Corollary 2 (Auxiliary upper experiment)**.** *If the logged transcript is a measurable function of a richer record, $Z = f(W)$, then $\mathrm{Acc}^\star(Z) \le \mathrm{Acc}^\star(W)$.*

Proof. Z is a deterministic garbling of W, and Bayes-optimal identification cannot improve under garbling. □

In PointRobot, W adds the pre-squash Gaussian latents of the actions; the logged action is the tanh of the latent stored in float32, so Z is a function of W. Only the auditor uses W; no identifier observes it.

A.3 Proof of Proposition 1

Proof. Fix a declaration of $U_1, \ldots, U_R$ and a log whose factors follow the declared kernels. Append one further observed factor V with $V = Y$ deterministically. Every declared factor keeps its declared conditional law given its history, so any quantity computed from the declaration alone is unchanged. The extended transcript (Z, V) determines Y, so its Bayes accuracy is one. A procedure that returns a ceiling below one from the declaration alone is therefore invalid for this log. The construction applies to any finite M and any prior. A nontrivial ceiling thus requires every observed factor to be declared COMMON or CALIBRATED, or to be placed behind the stopping boundary. □

A.4 PAIRWISE CONTINUATION BOUND

Proof of Equations 4–6. Fix a pair (i, j) and stop immediately before its first UNKNOWN factor. Write the full laws as $P_i = P_i^{S_{ij}} K_i$ and $P_j = P_j^{S_{ij}} K_j$, where K_i and K_j are the continuation kernels after the stop. Consider the hybrid $Q_{j\to i} = P_j^{S_{ij}} K_i$. Since P_i and $Q_{j\to i}$ apply the same kernel K_i to different prefixes, data processing gives $\mathrm{TV}(P_i, Q_{j\to i}) \le \mathrm{TV}(P_i^{S_{ij}}, P_j^{S_{ij}})$. The laws $Q_{j\to i}$ and P_j share the prefix law $P_j^{S_{ij}}$ and differ only through the continuation. On E_{ij}^c the stopped transcript is the whole transcript and the continuation is empty, so K_i and K_j agree there, and $\mathrm{TV}(Q_{j\to i}, P_j) \le P_j(E_{ij})$. By the triangle inequality,

$$\mathrm{TV}(P_i, P_j) \le P_j(E_{ij}) + \mathrm{TV}(P_i^{S_{ij}}, P_j^{S_{ij}}).$$

The symmetric hybrid $Q_{i\to j} = P_i^{S_{ij}} K_j$ gives the same bound with $P_i(E_{ij})$. Taking the smaller bound and clipping at one yields Equation 4. Under domination, the standard likelihood-ratio identity gives Equation 5. The chain rule for stopped predictable factors gives

$$D_{\mathrm{KL}}\big(P_i^{S_{ij}} \,\|\, P_j^{S_{ij}}\big) = I_{i\to j},$$

so any valid KL-to-TV transfer may be substituted for the stopped-TV term. The audit uses the sharp classical envelope ψ_{KL}; Pinsker's inequality suffices for the simpler relaxation. □

A.5 DECISION-IMAGE AGGREGATION

Proof that the decision-image program upper-bounds Bayes accuracy. A deterministic or randomized classifier is a Markov kernel from the transcript to an output label in $[M]$. Let $q_i = (q_{i1}, \dots, q_{iM})$ be its output law under task i, so $q_{ik} = P_i(\widehat{Y} = k)$. Total variation contracts under this post-processing:

$$\mathrm{TV}(q_i, q_j) = \frac{1}{2}\sum_k |q_{ik} - q_{jk}| \le \mathrm{TV}(P_i, P_j) \le \delta_{ij}.$$

Every classifier therefore satisfies the constraint $\sum_k |q_{ik} - q_{jk}| \le 2\delta_{ij}$ of Equation 8, and its accuracy is exactly $\sum_i \pi_i q_{ii}$. Maximizing over the feasible set of output laws thus upper-bounds Bayes accuracy. □

Because the row difference $q_i - q_j$ sums to zero, each coordinate deviation is at most half of its ℓ_1 norm, so $|q_{ik} - q_{jk}| \le \delta_{ij}$ follows as a corollary; the converse need not hold. If only a graph of edge bounds is available, the triangle inequality supplies path bounds before the same program is solved. Appendix G reports the size of the program for $M = 100$.

Equal pair bounds and general priors. Suppose $\delta_{ij} = \delta \le 1$ for every pair and let $m \in \arg\max_i \pi_i$. For any feasible q and $i \ne m$, the coordinate corollary gives $q_{ii} \le q_{mi} + \delta$, so

$$\sum_i \pi_i q_{ii} \le \pi_m q_{mm} + \sum_{i\ne m} \pi_i q_{mi} + \delta(1-\pi_m) \le \pi_m \sum_i q_{mi} + \delta(1-\pi_m) = \pi_m + (1-\pi_m)\delta.$$

Conversely, the rows $q_m = e_m$ and $q_i = \delta e_i + (1-\delta)e_m$ for $i \ne m$ are feasible, since every pair of rows is at TV distance exactly δ, and they attain this value. The value of the program is therefore $\min\{1, \max_i \pi_i + (1 - \max_i \pi_i)\delta\}$, and a uniform prior gives Equation 9.

Pairwise TV need not determine Bayes accuracy. With three tasks and a uniform prior, let P_1, P_2, and P_3 be uniform on $\{a, b\}$, $\{b, c\}$, and $\{c, a\}$. Every pair has TV $1/2$, so the program gives $1/3 + (2/3)(1/2) = 2/3$, whereas each observation is consistent with exactly two tasks and the Bayes accuracy is $1/2$. The exact agreement between the pair-TV program and Bayes accuracy in the controlled family of Appendix E.1 is therefore a property of that family.

A.6 LOCALIZED-READ ADAPTER

Proof of Equation 10. For a pair (i, j), define the boundary as the first read whose location lies inside the declared pair neighborhood of radius d. Before that boundary, non-read kernels are COMMON and every read has directional KL at most $D_{i\to j}(d)$. If the protocol permits at most N reads almost surely, predictable KL accumulation gives $I_{i\to j}(d) \le N D_{i\to j}(d)$. Substituting into the pairwise continuation bound yields Equation 10. □

A global symmetric exposure/channel envelope, in which one bound δ holds for every pair, is a convenient relaxation of these pair-local quantities; combined with Equation 9, it is the form used in the search-domain factorial (Appendix C.1).

A.7 SHARED-POLICY REWARD-FREE COROLLARY

Proof of Corollary 1. Factor a reward-free trajectory into reset, action, transition, non-reward observation, and termination or truncation. Under a shared execution policy and task-invariant reward-free environment semantics, each conditional factor is COMMON across candidate tasks. Their product therefore defines the same path law under every task. Pairwise TV is zero, and the posterior task probabilities equal the prior for every observed path, which is the last statement of Theorem 1. □

A.8 PREDICTABILITY AND REFINEMENT

The pre-sampling status requirement is essential. If a factor could be labeled CALIBRATED only after its realized value is observed, the audit could stop or retain samples preferentially according to task-informative outcomes, which would invalidate both the stopped factorization and its finite-sample endpoint.

For a predictable refinement $U = (U^{(1)}, \dots, U^{(m)})$ of a fully calibrated factor, the conditional KL chain rule shows that the sum of the refined expected increments equals the KL of the coarse factor. Splitting or regrouping a fully calibrated factor while preserving the observed experiment may change finite-sample efficiency but not the population directional information. Around UNKNOWN factors, refinement is safe when the predeclared stopping family retains the original coarse stop: a finer declaration then cannot make the best eligible population certificate worse, because the original option is still available. Classifying a factor after it has been sampled is excluded, since it is not predictable.

B FINITE-SAMPLE ENDPOINTS

B.1 DIRECT M-WAY UPPER AND IDENTIFIER LOWER ENDPOINTS

For task i, let $G_{ir} \in [0, 1]$ be the largest posterior probability of the upper experiment for episode r, and let n_i be the number of independent episodes. The prior-weighted target is $V = \sum_i \pi_i \mathbb{E}_i[G]$, estimated by $\widehat{V} = \sum_i \pi_i \bar{G}_i$, where $\bar{G}_i$ is the mean score for task i. The estimator is a sum of independent terms $\pi_i G_{ir}/n_i$, each with range π_i/n_i, so the sum of squared ranges is $\sum_i \pi_i^2/n_i$. Hoeffding's inequality then gives, with probability at least $1 - \alpha$, the upper endpoint

$$\widehat{V} + \sqrt{\frac{1}{2} \log \frac{1}{\alpha} \sum_{i=1}^{M} \frac{\pi_i^2}{n_i}}, \tag{11}$$

clipped at one. A family of K prespecified horizons receives simultaneous coverage by splitting α across the K endpoints before any outcome is inspected. The logged-only identifier uses the same construction, with Bernoulli correctness indicators in place of G_{ir} and a lower endpoint in place of an upper one.

For the PointRobot result in Table 7, the family-wise level is 0.05. The upper and lower families receive 0.025 each, and each family contains the six fixed horizons $\{1, 2, 4, 8, 12, 20\}$. All twelve raw endpoints are formed first; monotone closure in horizon is a deterministic post-processing of the simultaneous family.

B.2 DIRECT-TV AND KL ENDPOINTS

When the stopped likelihood ratio $R = dP_j^{S_{ij}}/dP_i^{S_{ij}}$ is available, $(1 - R)_+ \in [0, 1]$ has mean $\mathrm{TV}(P_i^{S_{ij}}, P_j^{S_{ij}})$ under P_i, and it is computed from the log-ratio as described in Section 3.3. Any valid one-sided bounded-mean endpoint can therefore be used, simultaneous over the predeclared family of ordered edges. When only conditional KL increments are available, the audit requires a declared almost-sure cap on the information in an episode, or another externally justified concentration condition. Support mismatch, non-finite conditional KL, or a violated cap makes the requested finite endpoint non-admissible.

Stop probabilities. Equation 4 also needs $p_{i,ij}$ and $p_{j,ij}$. Each is the mean of $\mathbf{1}\{E_{ij}\}$ under one candidate, so any one-sided Bernoulli upper endpoint, such as Clopper–Pearson, applies to that candidate's episodes. Stop-probability and stopped-TV endpoints over all ordered edges and horizons form one family, and α is split across it by a union bound before any outcome is seen. In our Route B experiments no factor is UNKNOWN, so both stop probabilities are exactly zero.

B.3 LOCALIZED EXPOSURE AND SELECTED MARGINS

For a localized-read context r, let $D_r^{\min}$ be the minimum distance between any read location and the candidate set, and define the empirical exposure curve

$$\widehat{p}_i(d) = \frac{1}{n_i} \sum_{r=1}^{n_i} \mathbf{1}\{D_r^{\min} < d\}.$$

All 640 search-domain cells use one pooled two-sided DKW band (Massart, 1990), uniform over the frozen margin grid, with radius $\sqrt{\log(2/\alpha)/(2n)}$, where $n = 15M$ is the number of contexts in the cell. Pooling is valid because the controller does not depend on the task (Appendix C), so $p_i(d)$ is the same for every i and all contexts are i.i.d. draws from one exposure law. When the controller can react to task-dependent reads, pooling is not valid and taskwise bands with a Bonferroni split over tasks are used instead, as in Appendix C.4. The channel term is either known from the simulator or receives its own simultaneous calibration endpoint. The margin with the smallest final ceiling is selected only after the joint exposure/channel event has been formed, so post-selection validity follows from the union bound. A frozen diagnostic with 48 primary cells (200–1500 contexts each, 10,000 replicates per cell) repeats margin selection in each replicate to check post-selection coverage. The minimum observed coverage is 0.9989, with no confident undercoverage.

B.4 SUMMARY OF THE THREE CERTIFICATE ROUTES

Table 2 lists, for each route, the information it needs, the object it bounds, its finite-sample requirement, and when it is vacuous or refused. The route is chosen by the likelihood information available on the retained experiment, not by a ranking of certificate strength, and predictable stopping and revealing completion come before all three routes.

Table 2: The three certificate routes under the declared provenance contract.

Available information	**Route**	**Object bounded**	**Finite-sample requirement**	**Vacuity / refusal**
Full candidate likelihoods	A: Direct M-way	Bayes accuracy of the revealing-completion experiment	Valid likelihoods; bounded-score endpoints simultaneous over predeclared choices	Ceiling at 1: vacuous. Invalid contract: refuse.
Pairwise likelihood ratios	B: Pairwise TV	Pairwise TV followed by M-way decision-image aggregation	Valid likelihood ratios; simultaneous pairwise-TV endpoints	Ceiling at 1: vacuous. Unsupported likelihood ratios: refuse.
Directional KL or calibrated channel envelope	C: KL→TV	Directional information, pairwise TV, and the resulting M-way ceiling	Valid KL endpoints, with declared information caps or other justified concentration conditions when needed	Ceiling at 1: vacuous. Unsupported information endpoint: refuse.

C Details of the Search-Domain Experiments

The search-domain factorial uses three identifying channels: the published olfactory-search count mechanism in OTTO (Loisy and Heinonen, 2023), the radiation-source observation model in Rad-Search (Proctor et al., 2021), and the known sensor model of RockSample (Smith and Simmons, 2004). In every domain, task-free movement and transition components are COMMON when the benchmark contract justifies it, localized identifying reads are CALIBRATED outside the selected neighborhood, and the first unrestricted near-field read defines the UNKNOWN boundary. RockSample read counts use the protocol's almost-sure bound.

Controller and exposure. In all three domains, contexts are collected by a uniform random walk that starts at the centre of the map, stays within a fixed radius of it, and never reads the task or any sensor value. Within each cell, rollouts consume fresh controller draws and action paths are not replayed across task labels. The path of read locations therefore has the same law under every task. Exposure is measured at the read locations by the distance to the nearest of all M candidates, which upper-bounds the exposure of every pair; read counts use the protocol's almost-sure maximum; and the margin is selected only after the exposure and channel endpoints are formed. The asymmetric-exposure stress test of Appendix C.4 differs: its controller reacts to observed reward values, so exposure depends on the task.

Table 3: Main factor declarations for the three kinds of logs audited in Section 4: the localized-read search domains, the task-specific PointRobot logs, and the shared-policy logs. The route follows from the available information.

Setting	COMMON factors	CALIBRATED / UNKNOWN factors	Route
Localized search (Sec. 4.1)	reset, movement, task-free transition	far reads calibrated; first near read unknown	C: KL→TV
Task-specific PointRobot (Sec. 4.2)	reset, physical transition, termination	policy likelihoods calibrated for all 100 candidates	A: direct M-way
Shared-policy OMRL (Sec. 4.3)	reset, policy, transition, non-reward observation, termination	none	none needed: ceiling $= \max_i \pi_i$

C.1 Factorial design

Each domain is audited on a full factorial of channel strength, candidate count, and seed (Table 4). Every cell uses a uniform prior over its candidates, 15 contexts per task, horizon 40, and family-wise

level $\alpha = 0.05$. The largest pairwise bound from Equation 10 is used as a single envelope δ for all pairs, and the reported ceiling is $1/M + (1 - 1/M)\delta$ from Equation 9. Table 5 breaks the outcomes of Table 1 down by channel level. Within each domain, the outcome category of the admissible cells is determined by the channel level alone: the weaker levels give informative ceilings and the stronger levels give vacuous ones, for every candidate count.

Table 4: Axes of the search-domain factorial.

Domain	Channel-strength parameter	Levels	Candidate counts	Seeds
OTTO	dispersion length $\lambda/\Delta x$	$1.0, 1.3, 1.6, 2.0$	$M \in \{4, 8, 12, 16\}$	10
RadSearch	source intensity	10^5, 2.5×10^5, 10^6, 5×10^6	$M \in \{4, 8, 12, 16\}$	10
RockSample	half-efficiency distance	$0.5, 0.8, 2.0, 3.0$	$\{4, 6, 8, 12\}$ rocks	20

Table 5: Search-domain outcomes by channel level, summed over candidate counts and seeds.

Domain	Channel level	Informative	Vacuous	Non-admissible
OTTO	1.0, 1.3, 1.6 (each)	37	0	3
OTTO	2.0	0	37	3
RadSearch	10^5, 2.5×10^5 (each)	37	0	3
RadSearch	10^6, 5×10^6 (each)	0	37	3
RockSample	0.5, 0.8 (each)	78	0	2
RockSample	2.0, 3.0 (each)	0	78	2

Structure of the informative ceilings. In every informative cell, no pooled context reached the selected neighborhood, so the exposure endpoint equals the pooled two-sided DKW radius $\sqrt{\log(2/\alpha)/(2 \cdot 15M)}$, which ranges from 0.088 ($M = 16$) to 0.175 ($M = 4$). The far-read term is fixed by the declared channel at the selected margin, and the selected margin is the same in every seed of a configuration. The certified ceiling is therefore identical across seeds within each (channel level, candidate count) configuration; seeds change the sampled contexts, and hence the diagnostic gates and the decoder-based violation check, but not the ceiling. The 341 informative cells thus correspond to the 28 configurations in Table 6. In every informative OTTO cell, the selected margin is smaller than the smallest possible read-to-candidate distance (candidates lie at least 11.3 from the centre and the controller stays within 3 of it), so no context can reach the neighborhood and the exposure term is only a sampling allowance. The exposure term shrinks as $n^{-1/2}$, and the best fixed decoders stay within 0.06 of chance. These experiments therefore characterize outcome separation and finite-sample validity; certificate tightness is examined separately in Section 4.4.

Table 6: Certified informative ceilings and their parts for each (channel level, candidate count) configuration; each ceiling is shared by every admissible seed of its configuration. M is the candidate count (rocks for RockSample) and d the selected margin. *Exposure*: pooled DKW radius (no context reached the neighborhood in any informative cell). *Channel*: KL-to-TV bound for the far reads at margin d. *Ceiling* $= 1/M + (1 - 1/M)$(Exposure + Channel). *Decoder*: mean point accuracy over seeds of the best decoder not fitted to the audited contexts.

Domain	Level	M	d	Exposure	Channel	Ceiling	Decoder	$1/M$
OTTO	1.0	4	9	0.175	0.174	0.512	0.270	0.250
OTTO	1.0	8	8	0.124	0.308	0.503	0.126	0.125
OTTO	1.0	12	8	0.101	0.308	0.458	0.067	0.083
OTTO	1.0	16	8	0.088	0.308	0.433	0.067	0.062
OTTO	1.3	4	9	0.175	0.383	0.668	0.275	0.250
OTTO	1.3	8	8	0.124	0.590	0.750	0.133	0.125
OTTO	1.3	12	8	0.101	0.590	0.717	0.068	0.083
OTTO	1.3	16	8	0.088	0.590	0.698	0.072	0.062
OTTO	1.6	4	9	0.175	0.604	0.834	0.275	0.250
OTTO	1.6	8	8	0.124	0.830	0.959	0.143	0.125
OTTO	1.6	12	8	0.101	0.830	0.937	0.074	0.083
OTTO	1.6	16	8	0.088	0.830	0.922	0.079	0.062
RadSearch	10^5	4	500	0.175	0.229	0.553	0.272	0.250
RadSearch	10^5	8	500	0.124	0.229	0.434	0.142	0.125
RadSearch	10^5	12	500	0.101	0.229	0.386	0.091	0.083
RadSearch	10^5	16	500	0.088	0.229	0.359	0.067	0.062
RadSearch	2.5×10^5	4	500	0.175	0.552	0.796	0.308	0.250
RadSearch	2.5×10^5	8	500	0.124	0.552	0.717	0.168	0.125
RadSearch	2.5×10^5	12	500	0.101	0.552	0.682	0.104	0.083
RadSearch	2.5×10^5	16	500	0.088	0.552	0.662	0.077	0.062
RockSample	0.5	4	3	0.175	0.099	0.456	0.267	0.250
RockSample	0.5	6	3	0.143	0.099	0.368	0.175	0.167
RockSample	0.5	8	2.5	0.124	0.197	0.406	0.122	0.125
RockSample	0.5	12	3	0.101	0.099	0.267	0.079	0.083
RockSample	0.8	4	3	0.175	0.459	0.725	0.284	0.250
RockSample	0.8	6	3	0.143	0.459	0.668	0.188	0.167
RockSample	0.8	8	2.5	0.124	0.680	0.829	0.134	0.125
RockSample	0.8	12	3	0.101	0.459	0.596	0.087	0.083

C.2 Sample-size sweep

A follow-up varies contexts per task over $n \in \{15, 50, 100, 500\}$ on three frozen search settings. The worst-case Route C ceiling decreases with n but approaches a nonzero population floor: RadSearch 10^5, $M = 4$ moves from 0.553 to 0.445 with floor 0.422; RadSearch 2.5×10^5, $M = 8$ moves from 0.717 to 0.627 with floor 0.608; and RockSample 0.8, $M = 8$ moves from 0.829 to 0.739 with floor 0.720. The exposure term shrinks with sample size, while the worst-case per-read channel charge does not.

C.3 Admissibility and the violation criterion

A cell is reported as admissible under its declared premises only if no predeclared diagnostic gate fires. There are two gates. The geometry probe fits a logistic-regression classifier to the non-read geometry of the contexts and reports its mean binary accuracy over 40 random task pairs; chance is 0.5, and the gate fires above 0.60. The observation probe reports the M-way accuracy of a classifier that sees only non-read observations; the gate fires above chance plus 0.10. The gates can falsify the declared ledger, but passing them is not evidence that the ledger is complete. All 32 non-admissible cells are refusals triggered by a gate; no other refusal condition fired. Refusals occur at every channel level (Table 5), and in RadSearch and RockSample only at $M = 4$. The 32 refusals come from 8 distinct (domain, candidate count, seed) samples: the probes use only non-read geometry and

observations, which do not depend on the channel level, so each firing repeats across the four levels. Seven are geometry-probe firings and one is an observation-probe firing. Because the controller never reads the task (Appendix C), the probed quantities have the same law under every task, so these firings are false alarms (8 of 160 samples, 5%); the audit still refuses, as its fail-closed rule requires. In the outside-channel control design of Appendix C.4, where the declaration is correct, the gates fire in 0 of 5 control seeds.

A follow-up whole-context permutation calibration uses 99 permutations and fires when the minimum permutation p-value is at most 0.025. Replaying the 160 historical task-blind settings gives 6/160 firings (3.75%), all within the legacy firing set; in a post-hoc follow-up on 30 additional RadSearch $M = 4$ seeds it fires in 3/30 settings versus 9/30 for the legacy threshold. In controlled shift tests it detects 11/20 shifts of size 0.05 and 20/20 shifts of at least 0.1. This calibration is reported as a follow-up and does not relabel the frozen 640-run table.

A confident violation is recorded when a fixed decoder that is not fitted to the audited contexts, namely the spatial-centroid, matched-filter, or Bayes decoder (the last only in RadSearch and RockSample), has a Hoeffding lower confidence bound on accuracy, simultaneous over the eligible decoders, that exceeds the certified ceiling. A fitted mean-pool decoder is evaluated but is not eligible for this criterion. No admissible cell records a confident violation. The 95% confidence allocation applies to eligible decoders within each cell, not jointly across all 608 admissible cells. The criterion has low power in this design. Vacuous cells cannot be violated, and in the informative cells the largest lower confidence bound of an eligible decoder is 0.125 (OTTO), 0.199 (RadSearch), and 0.215 (RockSample), and in every informative cell the ceiling exceeds the best such bound by at least 0.25 (smallest per-cell gaps: 0.404 in OTTO, 0.343 in RadSearch, and 0.251 in RockSample). The criterion is therefore a consistency diagnostic; the within-channel misspecification test in Appendix C.4 shows a declaration error that an independent decoder check does detect.

C.4 Stress tests

The stress tests were prespecified as characterizations of failure modes; their parameters were not tuned after the results were observed. For a margin grid $\mathcal{D}$, write E_d for the event of reaching the neighborhood of margin d, and define the exposure gap

$$\Delta_{\mathrm{exp}} = \max_{d \in \mathcal{D}} \max_{i,j} \big| P_i(E_d) - P_j(E_d) \big|,$$

computed exactly from the design.

Asymmetric exposure. A controlled design uses one controller for all tasks. The controller never sees the task, but it reacts to observed reward values, so different tasks lead to different exposure. The exposure gap is $\Delta_{\mathrm{exp}} = 0.576245$; at the margin that attains it, the two tasks reach the neighborhood with probabilities 0.7990 and 0.2227. The taskwise simultaneous DKW bounds remain valid, but the selected certificate is vacuous in all 5,000 replications; pooling exposure across tasks would not be admissible here. A RadSearch design provides the informative counterpart: its exposure gap is 0.102739 (probabilities 0.1450 and 0.0423), all 10 frozen seeds yield informative certificates, the mean ceiling is 0.940596, and the margin $d = 900$ is selected in every seed. Together, the two designs separate validity under asymmetric exposure from usefulness of the resulting bound.

Within-channel misspecification. In a binary problem with true per-read KL 0.06 and true Bayes accuracy 0.687897, we deliberately underestimate the per-read KL by declaring $\gamma \times 0.06$, for $\gamma \in \{1, 0.75, 0.5, 0.25\}$. At $\gamma = 0.75$ no check fires and the ceiling, 0.712, still lies above the true accuracy. At $\gamma \leq 0.5$ the ceiling falls below the true accuracy, and an independent check detects the contradiction in every seed: the lower confidence bound on the accuracy of a fixed decoder, evaluated on held-out data, exceeds the declared ceiling.

Outside-channel leakage. A task-dependent non-read observation is added to the log while the ledger declares it COMMON. The geometry probes reach 0.9829 with logistic regression and 0.9846 with an MLP, the diagnostic gates fire and the audit refuses in all five leak seeds; no gate fires in any of the five control seeds. A control design without the leak certifies normally. Channel calibration and factor completeness are therefore distinct obligations: empirical probes can falsify a claimed common factor, but they cannot verify that no task-dependent factor has been omitted.

C.5 REACHED-STOP RADSEARCH CONFIRMATION

A RadSearch follow-up keeps the same observation channel and Route C construction as the search factorial but widens the task-blind random-walk confinement radius from 400 to 650 cm. The design was frozen after separate pilot seeds. With $M = 4$, source intensity 10^5, and 100 contexts per task, all 10 confirmation seeds return informative certificates with no gate firing or confident violation. In 7/10 seeds the selected margin is 300 cm and 16.0–18.8% of contexts reach it; per-task stop frequencies range from 0.11 to 0.24, and the first stop occurs after 6–9 retained calibrated reads. Their certified ceilings are 0.876–0.897. In the remaining 3/10 seeds the selected margin is 250 cm, no context reaches it, and the ceiling is 0.913. Across all 10 seeds, a Poisson Bayes decoder reaches 0.265–0.323 against chance 0.25. At $M = 8$, the pilot search did not find a setting that combined a reached stop with an informative ceiling. The seven reached-stop seeds therefore give natural cases with $0 < P(E) < 1$ in which revealing completion remains informative.

D DETAILS OF THE OMRL EXPERIMENTS

D.1 TASK-SPECIFIC POINTROBOT PROTOCOL

Protocol. The direct M-way result uses the fixed 100-task half-circle PointRobot family and the same frozen task-specific actor family for all six prefixes. Exactly 1000 independently seeded episodes are collected per task, each to horizon 20. Episode seeds follow `SeedSequence([20260916016, task, episode, tag])`, with tag 613 for the environment stream and 977 for the policy stream. The upper certificate computes the largest posterior under the uniform prior using all 100 candidate policy likelihoods on the generation-latent transcript. The logged lower identifier receives only pre-action states and float32 actions. Before the inverse tanh, logged actions are clipped symmetrically in float64 to $\pm(1 - 2^{-53})$; this clipping is inactive for all logged actions in Families A–C. The tanh Jacobian is the same for every candidate and cancels, and the logged identifier is a fixed rule, so its lower endpoint stays valid. No task subset, edge graph, horizon, confidence allocation, or sample-size extension is selected from the result; Table 7 reports the entire confirmatory family, and Figure 2 in Section 4.2 plots it.

Table 7: Direct M-way estimates and jointly valid 95% identification brackets for PointRobot Family A across all six prefixes. Both estimates are finite-sample quantities computed from the same episodes, so the identifier estimate can exceed the upper-experiment estimate ($H = 4, 8, 12$); the certificate is the bracket.

H	Logged identifier estimate	Upper-experiment Bayes estimate	Joint bracket
1	0.082010	0.083032	[0.076775, 0.088267]
2	0.123460	0.123634	[0.118225, 0.128869]
4	0.208100	0.207468	[0.202865, 0.212703]
8	0.515750	0.514437	[0.510515, 0.519672]
12	0.940870	0.940732	[0.935635, 0.945967]
20	0.998310	0.998386	[0.993075, 1.000000]

Actor-family replication. To check that the result does not depend on one trained actor family, the same protocol is repeated for three families that differ only in their training seed (0, 1, and 2). The task family, goal geometry, network architecture, optimizer settings, 50000 training steps, collection producer, 1000 episodes per task, horizon 20, and six prefixes are identical, and task goals are recomputed and checked for equality. The collection seed rule has no family term, so all families use the same episode-seed integers. Shared seeds reduce variation from evaluation-seed differences but do not make cross-family errors independent; logs are therefore not pooled, and no independence-based cross-family inference is made.

Family A is the frozen reference; a subsequent read-only audit applied analogues of all eight provenance checks and passed 8/8. For B and C, the checks cover all-task manifest completeness, training configuration, actor-hash agreement, distinctness and disjointness from A, recomputed goals, the collection producer, unchanged A records, and the input manifest. The protocol, the checks, and the rule of reporting every family were specified before B and C were trained.

Table 8 reports all three families. They show a similar rise in identification accuracy across the six prefixes: at every prefix, the difference between families is smaller than each family's bracket width, although at $H = 20$ the widths are limited because the upper endpoints saturate at one. The three families share one task family, architecture, training algorithm, and collection protocol and differ only in training seed, so they characterize seed-to-seed variation within this protocol; results for other training algorithms, architectures, or domains would require their own audits.

Table 8: Direct M-way identification brackets for three PointRobot actor families. Each family has its own simultaneous 95% guarantee over its twelve endpoints, not a joint guarantee over all 36. Family A (seed 0) is the frozen reference; B and C (seeds 1 and 2) are independently trained and passed eight provenance checks each. Because evaluation seeds are shared, cross-family comparisons are descriptive.

H	Family A (seed 0)	Family B (seed 1)	Family C (seed 2)
1	$[0.0768, 0.0883]$	$[0.0740, 0.0858]$	$[0.0766, 0.0880]$
2	$[0.1182, 0.1289]$	$[0.1146, 0.1255]$	$[0.1178, 0.1288]$
4	$[0.2029, 0.2127]$	$[0.1978, 0.2078]$	$[0.2025, 0.2126]$
8	$[0.5105, 0.5197]$	$[0.5032, 0.5123]$	$[0.5062, 0.5158]$
12	$[0.9356, 0.9460]$	$[0.9327, 0.9428]$	$[0.9369, 0.9465]$
20	$[0.9931, 1.0000]$	$[0.9930, 1.0000]$	$[0.9932, 1.0000]$

D.2 Shared-policy panels and probe diagnostics

For PointRobot, each frozen shared actor is run on all 100 tasks, while reward-free reset, physical dynamics, non-reward observation, and termination remain common. For HalfCheetah-Vel, the replacement shared-policy panels satisfy the analogous common reward-free ledger over 100 tasks. The population certificate is therefore $\mathrm{TV} = 0$ and uniform-prior accuracy 0.01 for each shared panel.

Frozen reward-free classifier probes are consistent with this statement but are not inputs to it. The median temporal-probe accuracies are 0.0097, 0.0107, and 0.0088 for the three prespecified PointRobot shared actors, and 0.0097 and 0.0093 for the two prespecified HalfCheetah shared actors. These probes are diagnostics only: equality of the laws already follows from the audited common-factor ledger.

D.3 Encoder by collection policy

We pair the PointRobot source contrast with learned task identifiers on the same 100-task family. Training uses 200 episodes per task and evaluation uses 100 episodes per task. Task-specific (TS) collection uses the frozen Family-A actor for each task; shared (SH) collection uses one task-50 actor for every task, with a task-25 shared actor reserved for test-only checks. We run reward-free (s, a, s') and reward-bearing (s, a, r, s') branches with a FOCAL-style encoder and a supervised DeepSets identifier. All 36 declared runs complete across three training seeds; chance is 0.01.

At $H = 20$, reward-free FOCAL-style identification is 0.252 on TS→TS and about chance on TS→SH, SH→TS, and SH→SH. Reward-free DeepSets reaches 0.661 on TS→TS and 0.007–0.010 on the three shared/cross-collection evaluations. With rewards, the full DeepSets model reaches 0.650 on TS→TS and 0.658 on SH→SH. On TS logs, an (s, a)-only DeepSets variant reaches 0.653, essentially matching the full model, while an (s', r)-only variant reaches 0.648. These results show that task-dependent collection can supply task identity that learned identifiers use; the cross-collection results also show that the learned decision rule depends on the collection distribution.

D.4 Task-dependent continuous-control transition

The shared-policy panels of Section 4.3 have task-common reward-free physical transitions. Walker2d with randomized physical parameters provides a different ledger (Table 9). Each of eight tasks draws `body_mass`, `dof_damping`, `body_inertia`, and `geom_friction`, which are written to the simulator model, so the physical transition depends on the task by construction. No validated candidate-conditioned transition likelihood is supplied for the *observed* Walker transcript, so this

factor is declared UNKNOWN. The observation map is task-independent given the simulator state, but it drops the global horizontal coordinate and clips velocities; the observed history may not uniquely determine the simulator state, so post-transition observation laws are not declared task-common.

Reset and the shared task-blind collection policy are COMMON. The status of the transition is fixed before the first action is sampled, and the first UNKNOWN boundary is the first transition, at $t = 0$. The retained prefix (s_0, a_0) carries no task information, and every episode reaches the UNKNOWN factor. Revealing completion therefore reveals the task on every branch, and the ceiling is the admissible but vacuous value 1, the prescribed consequence of the information the auditor has declared.

Table 9: Walker2d random-parameters factor ledger. The ceiling follows from the pre-sampling ledger; the probe distances in the text are finite-sample diagnostics, not population certificates.

Field	Declared setting or outcome
Environment	Walker2d random parameters, reward-free transcript
Candidates	Eight tasks, uniform prior; `body_mass`, `dof_damping`, `body_inertia`, `geom_friction`
Collection	Shared task-blind, fixed-seed uniform policy
Task-dependent factor	Physical transition, declared UNKNOWN before sampling
Stop	First transition, $t = 0$; prefix (s_0, a_0); stop probability 1; no task signal before stop
Ceiling and status	1.0; admissible, vacuous (revealing completion)

A paired diagnostic holds each reset draw and action fixed across candidates for 32 trials. The reset observation has a maximum cross-task spread of 0.0 to machine precision. The first post-reset observations have pairwise Euclidean distances of 0.114 (minimum), 2.399 (median), and 10.180 (maximum) over 896 pooled pairs, against a reset-noise half-width of 0.005. Under the blind policy, episodes last 9.31 steps on average (minimum 3, maximum 20). These observations serve as simulator-instantiation checks; the population certificate is provided by the calibrated experiment below.

A stochastic Walker variant provides the calibrated counterpart. It uses eight random-parameter tasks with full simulator state logged. While the walker is upright, the task-dependent transition is modeled as $s' = F_k(s, a) + \sigma\epsilon$, with Gaussian noise standard deviation 0.01 on positions and 1.0 on velocities, so the candidate transition density is available and declared CALIBRATED. After the walker leaves the upright region, an unmodeled task-dependent noise inflation makes the next transition UNKNOWN; reset and the blind uniform policy remain COMMON.

A first pilot at the benchmark parameter spread (log-scale 3) produced no eligible informative setting among 24 configurations because the calibrated transitions already made the tasks almost fully identifiable. The follow-up design was then frozen at log-scale 0.5, $c = 1$, $h_{\text{cal}} = 1.15$, and horizon 10. Across 10 confirmation seeds with 250 episodes per task, 16.7–19.5% of episodes reach UNKNOWN, usually at transitions 7–9, and all 10 seeds return informative ceilings of 0.832–0.847. Prefix Bayes accuracy is 0.766–0.790, while full-transcript oracle accuracy is 0.782–0.816. In 100 independent re-collections, all 100 certificates cover the population estimate 0.810. Together with the original Walker ledger above, the two cases show that the outcome depends on the information declared before the stop: an immediate unmodeled transition gives ceiling one, while calibrated task-dependent dynamics can accumulate useful prefix information before a later UNKNOWN boundary. Walker remains outside the search-factorial counts in Table 1.

E Full Ablation Results

This appendix gives the full versions of the two analyses summarized in Section 4.4.

E.1 Controlled certificate tightness

A controlled four-task OMRL-like family, with binary states and actions and Bernoulli kernels bounded away from 0 and 1, makes the behavior policy and the transition dynamics task-specific but

calibrated, while keeping the Bayes decision problem small enough to enumerate exactly. Table 10 reports population quantities, so the exact Bayes accuracy and FLAT ceilings do not carry finite-sample estimation error. The last column, Pinsker slack, is the ceiling obtained when Pinsker's inequality replaces the sharp envelope ψ_{KL} in Route C, minus the exact Bayes accuracy.

Table 10: Controlled-family decomposition of population certificate slack. *Exact-TV envelope*: exact pair TV fed to the decision-image program. *Direct-KL diagnostic*: exact directional KL fed directly to the output-law program of Equation 12. *KL→TV envelope*: the operational Route C. *Exact Bayes* is the target value and the next three numeric columns are ceilings; *Pinsker slack* is a difference: the Route C ceiling computed with Pinsker's inequality minus exact Bayes accuracy.

Regime	Exact Bayes	Exact-TV envelope	Direct-KL diagnostic	KL→TV envelope	Pinsker slack
Both task-specific	0.349507	0.349507	0.372007	0.372070	0.022928
Dynamics shared	0.321961	0.321961	0.338595	0.338614	0.016791
Policy shared	0.319705	0.319705	0.333470	0.333478	0.013889
Both shared	0.250000	0.250000	0.250000	0.250000	0.000000

Feeding exact pair TV to the decision-image program reproduces exact Bayes accuracy to numerical precision; the largest aggregation residual is below 4×10^{-15}. Even when exact directional KL values are used directly in the output-law program, almost the entire gap remains. In the regime where both factors are task-specific, exact Bayes accuracy is 0.349507074, the direct-KL diagnostic is 0.372007, and the operational KL→TV envelope is 0.372070489. The gap due to the directional-KL summary is about 0.02250, while the additional gap from the scalar KL-to-TV transfer is about 6.4×10^{-5}, under 0.3% of the total gap of 0.022563. The other regimes with a nonzero gap show the same decomposition. Under the uniform prior used in this decomposition, most of the slack comes from replacing exact pairwise information with directional-KL summaries, decision-image aggregation is tight, and the final KL-to-TV step adds little; the tightness of the aggregation is a property of this family (Appendix A.5). When both policy and dynamics are shared, all envelopes equal the prior maximum 0.25. In a 1000-replication finite-sample extension at 160 episodes per task, every ceiling covers exact Bayes accuracy in all four regimes; in the both-task-specific regime, all 1000 runs also cover all 12 population directional-information endpoints. The mean ceiling is 0.37622 (standard deviation 7×10^{-5}), compared with population Route C value 0.372070 and exact Bayes accuracy 0.349507.

The direct pairwise-KL envelope. Let π be the prior over M candidates, $\mathcal{E}$ a set of directed candidate pairs, and d_{ij} a bound on $D_{\mathrm{KL}}(P_i \| P_j)$ for $(i,j) \in \mathcal{E}$. Define the population functional

$$V_{\mathrm{KL}}(\pi, \mathcal{E}, d) = \max_{Q \geq 0} \sum_{i=1}^{M} \pi_i q_{ii} \quad \text{s.t.} \quad \sum_{k=1}^{M} q_{ik} = 1 \ (\forall i), \quad D_{\mathrm{KL}}(q_i \| q_j) \leq d_{ij} \ (\forall (i,j) \in \mathcal{E}), \quad (12)$$

where q_i, the i-th row of Q, is the identifier's output law under task i.

Proposition 2. *If $D_{\mathrm{KL}}(P_i \| P_j) \leq d_{ij}$ for all $(i,j) \in \mathcal{E}$, then the Bayes task-identification accuracy under prior π is at most $V_{\mathrm{KL}}(\pi, \mathcal{E}, d)$. The bound is sharp over the class of experiments constrained by exactly those directional KL caps.*

Proof. By data processing, every randomized identifier induces rows q_i that satisfy the directional constraints, and its accuracy is $\sum_i \pi_i q_{ii}$. Maximizing over identifiers therefore gives

$$\mathrm{Acc}^\star \leq V_{\mathrm{KL}}(\pi, E, d).$$

For sharpness, let $Q^\star$ attain the maximum in Equation (12), and define an experiment on the decision alphabet with $P_i = q_i^\star$. This experiment satisfies the prescribed directional KL caps, and the identity identifier attains

$$\sum_i \pi_i q_{ii}^\star = V_{\mathrm{KL}}(\pi, E, d).$$

Hence its Bayes accuracy is at least $V_{\mathrm{KL}}(\pi, E, d)$, while the upper bound just proved gives the reverse inequality. Its Bayes accuracy therefore equals $V_{\mathrm{KL}}(\pi, E, d)$, establishing sharpness for this information class. □

The functional is a population diagnostic for the controlled family, not an additional finite-sample route. This analysis was specified in advance (artifact file `theory/direct_kl_envelope_preregistration.md`). Its criterion M1 required the envelope to recover at least half of the slack, i.e. a ceiling of at most 0.360789; it gives 0.372007 and recovers 0.28%, so M1 fails. Criterion M2 required an improvement of at least 0.02 over the KL-to-TV route on a non-saturated PointRobot prefix; the largest measured improvement is 0.0013, at $H = 12$, the only such prefix where the envelope is precise enough to decide, so M2 also fails. We retain the envelope as an appendix diagnostic for how directional-KL summaries affect the ceiling.

Prior sensitivity and an information baseline. The controlled family also permits exact population comparisons under nonuniform priors. Under $\pi = (0.4, 0.3, 0.2, 0.1)$ with both factors task-specific, Bayes accuracy is 0.416172, the exact-TV program gives 0.479257, the direct directional-KL envelope gives 0.439712, and operational Route C gives 0.497212. Thus the exact-TV/Bayes coincidence under the uniform prior is not prior-independent. As a separate baseline, exact mutual information gives a Fano-based upper bound on identification accuracy of 0.4121 under the uniform prior, compared with Bayes 0.3495, exact-TV 0.3495, and direct KL 0.3720. Across the 16 checked regime/prior combinations the Fano ceiling remains valid; the comparison illustrates that different information summaries can order differently once the prior changes.

Numerical certification. We solve Equation 12 using a cutting-plane outer approximation. Because KL is jointly convex, each supporting-hyperplane cut is a necessary condition, so every outer linear-program optimum bounds the program from above even before convergence. A KL-feasible primal witness supplies the lower bound. In the regime where both factors are task-specific, the certified outer ceiling is 0.372007 and the feasible lower witness is 0.372005; the bracket width is 2.0×10^{-6} at a reporting quantum of 10^{-6} (unquantized width 8.87×10^{-7}). The primal maximum KL residual is 0.0, the outer candidate's maximum KL violation is 7.60×10^{-7}, and the maximum simplex residual is 1.11×10^{-16}. The computation uses 325 cuts, 49 linear-program solves, and 12 directed edges, with a cut-validity margin of 10^{-11}.

E.2 Information availability on one fixed PointRobot log

To vary the likelihood information declared by the auditor, we keep the PointRobot actors, episode identities, state and action trajectories, prior, and six prefixes fixed. Likelihood calculations that need the generation coordinate use the stored pre-squash latents from the same episodes. These latents augment the generation record for the audit; they are not observations given to the logged-only identifier. The four regimes are:

- R1: all candidate likelihoods and the direct M-way functional (Route A);
- R2: likelihood ratios on a sparse chain of 99 pairs, with decision-image aggregation (Route B);
- R3: an exact-kernel KL plug-in with no confidence adjustment;
- R4: the behavior policy declared UNKNOWN, so the first-unknown rule reveals the task and the ceiling is structurally one.

The chain links geometrically adjacent goals: the 100 goals are equally spaced on the half-circle, so index adjacency is geometric adjacency, and the chain was fixed before any likelihood was evaluated. For each orientation of each edge, R2's endpoint is the smaller of an empirical-Bernstein and a Bernoulli-KL bound, each computed at half the endpoint's error level so that their minimum keeps the stated level; the Bernoulli-KL bound is valid for $[0, 1]$-valued samples by Hoeffding's convexity argument (Hoeffding, 1963).

Table 11 compares the regimes using plug-in values, without finite-sample confidence adjustment.

Table 11: Like-for-like plug-in changes on the fixed PointRobot log, without confidence adjustment. Each column reports a difference of plug-in ceilings: R1 → R2 is R2 minus R1, and R2 → R3 is R3 minus R2. R2 uses a sparse chain of 99 pairs; R3 is an exact-kernel KL plug-in.

H	R1 → R2	R2 → R3
1	$+0.000016$	$+0.018067$
2	-0.000188	$+0.028350$
4	-0.000135	$+0.048941$
8	$+0.000255$	$+0.124478$
12	$+0.000341$	$+0.057395$
20	-0.000011	$+0.001625$

The plug-in ceilings for R1 and R2 differ by less than 0.0004 at every prefix, with differences in both directions. At the population level the R2 value can only exceed the R1 value, because the pair-TV program upper-bounds Bayes accuracy; the three negative plug-in differences are therefore estimation error, and all differences are an order of magnitude smaller than the 0.0052 Hoeffding radius of a single R1 endpoint. Restricting the auditor to likelihood ratios on the 99-pair chain thus changes the plug-in result very little on this fixed log, which is consistent with confusions concentrating between neighboring goals. Compressing those ratios into per-edge directional-KL summaries instead raises the plug-in ceiling by up to 0.124478. At $H = 12$, the R3 plug-in ceiling is 0.998468, while the certified R2 ceiling is about 0.967. A later full-4950-pair attempt completed the 100-candidate statistics but did not return an R2-full ceiling within the frozen compute budget: the decision-image LP has about 5.05×10^5 variables and 9.95×10^5 inequalities. The result is therefore a compute-limited record for this protocol, while the reported R2 comparison still uses the 99-edge chain.

The certified R2–R1 gaps at $H = 1, 2, 4, 8, 12, 20$ are 0.033, 0.039, 0.050, 0.060, 0.021, and -0.000. The plug-in ceilings differ very little, but the certified R2 ceilings can be substantially higher. This difference is mainly associated with the finite-sample confidence cost: R2 splits its error budget over 1188 endpoints (2 directions $\times$ 99 pairs $\times$ 6 prefixes) using a Bonferroni allocation. At $H = 20$, R2's certified ceiling is 0.999928, slightly below R1's 1.000000, because the two routes use different confidence families and R1 is clipped at one. Each confidence family is separately family-wise 95%, one containing the twelve direct-route endpoints and the other the prespecified R2 endpoints, so their simultaneous coverage is at least 90% by a union bound. R4 is structural and deterministic. R3 carries no finite-sample confidence statement; its values are diagnostics, not certified or population quantities.

Prespecified conditions. The prespecified protocol (artifact file `theory/information_availability_ladder_preregistration.md`) has five conditions, and three of them pass. The certified magnitude and breadth conditions pass: the largest certified R2–R1 gap is 0.0604, and five of the six prefixes have a certified gap of at least 0.02. The structural fail-closed condition passes because R4 equals one at all six prefixes. The remaining two prespecified conditions fall below their thresholds: at $H = 12$, R3 reaches 0.998468 versus 0.999 for diagnostic separation, and the largest plug-in R2–R1 gap is 0.00034 versus 0.03 for attribution. We therefore use this experiment to characterize route differences on the fixed log.

F Representation Diagnostics

FLAT bounds the task information available in a fixed log, while a learned representation may retain only part of that information. This appendix examines this second stage through controlled and within-model representation diagnostics.

A controlled Gaussian family. The family uses a two-dimensional point environment with 120 tasks, formed from 5 templates with 24 tasks each and split into 70 training and 50 evaluation tasks. Each context contains $N = 20$ reads from a Gaussian read channel with constants $w = 0.12$, $\rho = 0.60$, and $\sigma_0 = 0.10$, and candidate counts are $M \in \{2, 4, 7, 10, 14\}$. The candidate set, observation field, noise, channel constants, and estimator are held fixed, and only the context law changes. Under the evaluated random occupancy, produced by a uniform random policy started at

the origin, no context enters a candidate neighborhood, and the passive certificate is informative for every candidate count. Under the information-rich deployment occupancy, produced by a trained policy, every context enters a candidate neighborhood, so the localized certificate is vacuous, while a matched filter applied to the raw context reaches 0.999 task identification. The geometry probe stays near chance (≈ 0.50), and both settings are reported as admissible under their declared contracts. In the second setting, a learned representation whose accuracy is compatible with chance therefore cannot be read as evidence that the raw log lacks task information.

Encoders on an information-rich panel. On the frozen panel, the clearest within-model comparison uses the CTMRL classifier: task identification is 0.875 from the raw context and 0.349 from the frozen embedding, with raw-minus-embedding interval $[+0.444, +0.606]$. A matched filter on the same raw context reaches 1.000. Additional frozen encoder measurements give 0.519 for GENTLE and 0.304 for CTMRL. These results complement the collection-policy study by showing the second case in the paper's motivation: task information can be present in the log and only partly retained by a learned representation.

G Computational Cost

Table 12 reports the cost of the operational certificates on fixed PointRobot inputs. These are execution measurements, not certificate results.

Table 12: Operational certificate cost on fixed PointRobot inputs. Three repetitions of each stage on one machine (24 logical cores, 140 GB RAM, NVIDIA Quadro RTX 6000). Time is minimum/median/maximum wall clock; memory is peak child-process resident set size. Collection and retraining are excluded, and GPU memory was not measured.

Workload	Tasks	Episodes/task	Route	Wall clock (s)	Peak RSS
Direct M-way certificate	100	1000	A	74.4 / 75.5 / 75.6	1391 MiB
Decision-image, cached statistics	100	1000	B + C	4387 / 4521 / 4669	425 MiB
Statistics pass and aggregation	100	1000	shared + B + C	4454 / 4546 / 4626	1455 MiB

Route A. The direct route evaluates every episode under every candidate at every step, so its cost grows as M^2nH candidate-state evaluations for M tasks, n episodes per task, and horizon H. For this configuration that is $100^2 \times 1000 \times 20 = 2 \times 10^8$ candidate-state evaluations and 4×10^8 Gaussian log-density evaluations, counted analytically rather than instrumented. The six prefixes reuse candidate log-likelihoods accumulated in a single pass through the episodes, rather than six independent evaluations. The measured time includes actor and episode loading.

Routes B and C. Both aggregation workloads take roughly 75 minutes, most of it spent in decision-image aggregation and in the direct-KL cutting-plane loop. The workload that also computes the statistics has a median runtime only about 25 seconds longer than the cached-statistics workload. For $M = 100$ and 99 certified edges, the aggregation program has $M^2 + |\mathcal{E}|M = 19900$ variables and 19899 inequality rows. The number of direct-KL solves depends on the workload.

Scaling. Apart from this $M = 100$ configuration, the only measurements over other candidate counts concern the small search-domain families with $M \in \{4, 8, 12, 16\}$ (OTTO and RadSearch) and $\{4, 6, 8, 12\}$ rocks (RockSample). Empirical scaling beyond $M = 100$ remains unmeasured.

H Reproducibility and Claim Provenance

The anonymous artifact supports three levels of reproducibility. *Integrity verification* checks the released files, manifests, and record hashes. *Frozen-record parity* rebuilds manuscript tables or reported headline values from saved records. *Scientific recomputation* reruns the relevant model evaluations or scientific calculations from their required inputs. The release keeps these evidence levels distinct.

For the task-specific PointRobot family, the artifact releases the collection producer together with a deterministic generation contract. Running the producer reproduces the frozen 100-task logged family bitwise. It also reconstructs the seed families recorded in the frozen manifest and checks their consistency with the released generation procedure. These checks establish reproducibility of the released generation procedure and logged bytes.

For the search-domain factorial, the artifact stores the sufficient statistics needed to recompute the certificate layer without replaying the simulators. The recomputation agrees with the frozen certificate surface under the predeclared parity criterion. Because these statistics were stored after the original simulator runs, they support deterministic recomputation of the certificate layer from the saved sufficient statistics.

The artifact provides scientific recomputation paths for checkpoint-backed evaluations, the PointRobot certification analyses, and the controlled-family exact quantities. The historical HalfCheetah rollout manifest is unavailable. The artifact instead ships two pinned SAC actor files and a fixed panel-generation path. Re-collection of tasks 0, 50, and 99 for each actor reproduces all 2,400 checked panel files bit-for-bit on the recorded CPU stack. Actor retraining does not reproduce the historical actor bytes, so reproducibility is claimed for panel generation from the pinned actors rather than for bitwise actor training. The artifact labels claims separately as population, structural, finite-sample certified, diagnostic, parity, or scientific recomputation. If a required input or contract condition is missing, the corresponding recomputation path fails closed. The artifact README documents the exact commands, file locations, hashes, environment requirements, and verifier outputs.