AI SYSTEMS · PREPRINT · PRE-REGISTERED, ADVERSARIAL

# Inference Placement: Where Learned Inference Earns Authority in a Symbolic System

*The question is not whether machine learning is good. It is where, if anywhere, learned inference earns decision authority in a system that keeps a deterministic symbolic floor as the sole author of facts — measured, one pre-registered boundary at a time, by first attempting to eliminate inference.*

Perslis Research · kist × Perslis · 2026-09-21

↓ Download the paper (PDF)

Moat-scrubbed · every figure traces to a frozen result artifact · a program of principled rejections

Preprint · working draft | PILOT / RESEARCH · small pools | Moat-scrubbed

ABSTRACT

Most AI systems answer factual questions by generating text and hoping it is correct. An alternative — a typed symbolic store that is the *only* author of facts, with any neural component demoted to *proposing* what the store must independently verify — prevents hallucination structurally rather than statistically. This report asks the question that design forces: **where, if anywhere, does learned inference earn decision authority in such a system?** The method is a single refrain: at every boundary, first attempt to eliminate

inference — build the strongest deterministic (and relational) opponent, hand the learned lane its *theoretical upper bound* (an oracle with the true model), pre-register the deployment bar before any outcome is seen (≥ 15% improvement, paired 90% bootstrap CI > 0, no worse correctness), and keep the floor as the sole truth authority. Across three tracks — inference placement over a deterministic query policy (five boundaries), cost-weighted diagnostic uncertainty over 160 human kinases and 2,775 verified features, and a generalization domain of sequential Bayesian fault diagnosis with a formal noisy-OR generative model and an information-gain planner — the deterministic baseline holds. In the generalization domain a theoretical seat finally appears in exactly one corner (strong latent structure and a priced-error objective clear the bar at the *oracle ceiling*: 16.79% and 19.79%), but a realizable finite-sample learner captures only ~1.8% of that edge and a learning curve to 24,000 incidents plateaus below the bar, relocating the limit to model class. We formalize the outcome as a conjunction of three measured gates — theoretical value, realizable recovery at a supportable data price, and adequate model class — not met anywhere tested. The program also caught two of its own scoreboard failures (a gameable fitness that rewarded doing less; a blind measurement that produced a false conclusion), both retracted. The contribution is a measured map of where each mechanism earns authority, and the falsifiable conditions under which learned inference finally would.

***Preprint / research report — not peer-reviewed. Moat-scrubbed public version.***
*Evidence grade is PILOT / RESEARCH: a single target per stress class and small candidate pools. Every empirical figure below is transcribed from a frozen, hash-recorded result artifact and cross-checked against the live code at the time of writing; where a superseded result and its correction disagree, the authoritative*

*correction is followed. Floor implementation internals beyond the concepts needed to read the method are withheld.*

## 1 · Introduction

A large and growing class of AI systems answers a factual question by generating text and hoping it is correct. In scientific domains that is dangerous: a fabricated structure, a mis-attributed association, or a silently dropped fact is not a stylistic defect but a wrong answer with a confident surface. The dominant mitigation — retrieval-augmented generation — reduces the problem without eliminating it, because the model remains the *author* of the final claim and can still add, drop, or distort what retrieval returned [6]. A different design point, pursued across a companion line of Perslis work under the banner of *the inversion*, makes a typed symbolic store the **only author of facts** and forbids any neural component from writing a fact the store cannot ground [1]. Under that design, hallucination is prevented *structurally* — as a control-flow property of what may enter the graph — rather than *statistically*, by making a generator less likely to invent.

This report does not re-argue that thesis. It asks the question the thesis forces, and answers it empirically: **not "is ML good," but where, if anywhere, does learned inference earn decision authority in a provenance-constrained symbolic system?** A field that generates-and-hopes assumes the model belongs in the answer path by default. We invert the burden of proof and place it on the model at every boundary, then measure whether it can carry it.

The method is stated up front, because the method *is* the contribution. (i) **Build the strongest deterministic method first** — never compare against a strawman; strengthen the non-learned baseline until it can no longer be improved cheaply. (ii) **Give ML every fair chance**, up to and including an *oracle upper bound* — a model handed the true generative process, so that if even the oracle cannot clear the bar, no learned model can. (iii) **Pre-register the bar before each run**: the deployment

gate, the eligibility frame, and the materiality threshold are committed to version control before any outcome is seen, with a commit-reveal seed so there is nothing to shop. (iv) **Keep the floor as the sole truth authority**: a learned lane may only ever *propose* or *rank*; the deterministic floor authors every admitted fact and reveals every true outcome, so a model can never certify itself.

We report a program of principled *rejections*. That framing is deliberate: the finding is that sufficiently capable deterministic and relational alternatives keep winning, and that map — not a win count — is the deliverable.

### 1.1 · Contributions

1. **A burden-of-proof method for inference placement.** An adversarial, pre-registered protocol — strongest deterministic opponent first, an oracle ceiling that settles the value question without building a learner, a fixed deployment bar, and the floor holding all truth authority — that turns "where does inference belong?" into a measurable decision at each boundary (§2).
2. **A five-boundary placement result and a cost-weighted uncertainty result.** Op-selection, novelty depth, goal-relevance, category composition, and relational composition each resolve deterministic (§3); and on inference's most favorable terrain — cost-weighted, partially-observed identification — an accurate learned prior's only material contribution is inseparable from truth-corruption, while its truth-preserving value is negligible, and the entire "fill missing state" value belongs to verified retrieval (§4).
3. **A formal generalization model and the program's first cleared ceiling.** A noisy-OR generative family with a tunable latent-structure knob and a cost-aware expected-information-gain planner, exported to legacy fault diagnosis, where a theoretical seat finally appears at the *oracle ceiling* under a priced-error objective — but a realizable learner does not turn it into an earned one, and a learning curve locates the limit at model class (§5).
4. **The three-gate deployment framework** — theoretical value, realizable recovery at a supportable data price, and adequate model class — a conjunction not met anywhere tested (§6); plus a meta-result: the method caught two of its own

scoreboard failures by independent verification, turning the thesis on the program itself (§7).

## 2 · Method — an adversarial, pre-registered protocol

A comparison is trustworthy only if the instrument cannot be shopped. Six controls make the program hostile to its own thesis.

**Pre-registration, committed before outcomes.** Each experiment's design, eligibility frame, deployment gate, and materiality threshold are committed to version control before the target or pool is drawn. There is no operator-chosen seed: the selection seed is derived as `int(sha256(<frozen candidate set>)[:8], 16)`, so the frozen data determines the walk order and the first accession meeting the pre-declared criterion is the target. The full walk — every skipped candidate and why — is recorded.

**Cold draws and kept wrecks.** Targets are drawn cold from a pre-declared frame; development targets are permanently barred as optimization material. Every run, pass or fail, is frozen with a per-artifact SHA-256 digest and a git tag; losses and crashes are preserved unedited as first-class evidence, not footnotes. This integrity is *tamper-evident, not tamper-proof*: unsigned digests detect in-place edits by a party lacking the trusted digest set, but do not provide authenticity or an external anchor — cryptographic signing and anchoring are named future work.

**A shared, independent scorer and the floor as truth authority.** Every lane is scored by one shared module against an independent, neutral ground truth built by code that is neither lane's. The deterministic symbolic floor authors every admitted fact and reveals every true outcome; a learned lane may only propose a candidate (re-grounded by the floor before admission) or rank a measurement — it can never certify what is true.

**Paired-bootstrap confidence intervals.** Comparisons are paired per instance and bootstrapped (1,000×) to a 90% confidence interval, so a headline mean is never

reported without the interval that says whether it is distinguishable from zero.

> **Definition (the deployment gate).** A learned lane **clears** at a boundary iff its mean improvement over the strongest deterministic opponent is at least **15%**, with a **paired bootstrap 90% CI strictly > 0**, and **no worse correctness** (no worse ground-truth recall, resolution, or correct-identification, and unsupported-final claims remaining 0). The bar is identical across all three tracks; each experiment's threshold is fixed in its pre-registration before outcomes are seen, and is never relaxed to manufacture a win.

The refrain that ties every section together: **at every boundary, we first attempt to eliminate inference.** A boundary hires inference only when a sufficiently capable cheap deterministic (and relational) alternative has first been given the chance to make it unnecessary — and failed.

## 3 · Track 1 – Placement across five boundaries

Track 1 evolves a research *policy* — which deterministic question the floor should ask next — never the answers, rewarded only by what the frozen floor can independently verify and promoted only on a sealed held-out pool. It asks, at five successively sharper boundaries, whether learned operation-selection earns a seat. It does not. Each hold moved the frontier to a more precise place and forced the placement hypothesis to evolve through three falsifiable forms.

**Boundary 1 — operation selection (no computational break).** On eight cold-drawn proteins (48 operation executions, zero errors), a real-cost instrument measured the whole prize any selector could win over the dumbest "run every operation" policy. Running all operations costs 71.9 s of wall, 75 API calls, 2.12 MB, and $0.00 to recover 401 verified novel objects; a free, *perfect* skip-oracle — the unreachable upper bound — costs 58.0 s, 63 calls, 2.11 MB, $0.00 for the same 401. The entire prize is therefore **$0.00, ~9 KB, 12 API calls, and 13.9 s of wall (19.4%)**, and it fails the pre-registered materiality bar (a saving must clear $0.01, or ≥ 50% and ≥ 2 s of

wall, or relieve a real rate-limit breach). A real 7B selector deciding six times per protein could burn the whole 13.9 s on its own inference. There is no cost problem worth intelligence: **run everything.**

**Boundary 2 — depth by novelty (a break, but no selective signal).** Research depth is genuinely expensive: a breadth-first walk over the curated protein-interaction graph, with effective branching 16.75 at the seeds then 37.9 and 34.9 at depths 1–2, projecting to ≈ 14,496 API calls (~3.4 h) at depth 1 and ≈ 27,072 (~6.3 h) at depth 2 — flatly infeasible. But verified novelty is *uniform and distinct*: an 18-node sample yields 785 unique objects, per-node 29–132 (median 41.5), cross-node redundancy only 14%, and a perfect free oracle needs **10 of 18 nodes (56%)** to cover 70% of the evidence — far above the 30% exploitable bar. Explosion true, concentration false: a yield-maximizing strategist has no signal to exploit. An undirected walk drowns in equally-true, equally-novel facts — the retrieval-is-not-memory boundary [2].

**Boundary 3 — goal-relevance (selective, but a cheap category class).** A concrete goal restores selectivity: for the densest disease goal, 9 of 40 branches are relevant (fraction 0.225); for a zinc-finger structural domain, 5 of 40 (0.125); a router keeping only relevant branches would skip 78–88% of cost at zero goal-loss. But the selective goals that exist are *cheap-annotation classes* a deterministic UniProt-annotation filter likely captures, and the genuinely expensive goals (specific pathways, drugs) are needles — no anchor reaches even 3 of 40 branches. A selective signal exists but is dominated by cheap-filter structure; inference is not yet hired.

**Boundary 4 — compositional goal (a declined interview).** A compositional goal — relevant iff the floor's verified evidence shows disease-association *and* druggability *and* pathway-embedding, a conjunction across three sources so no single annotation defines it — yields 7 of 40 positives. The strongest fair deterministic rule reaches F1 0.667 (a third feature adds nothing), and cheap features carry association with peak ≈ 0.49. Taken literally, the pre-registered rule fired ("grant the model an interview"). We *declined* it and disclosed two conservative deviations: we

strengthened the deterministic baseline to three-feature rules (no change, so the ceiling is credible and the baseline fair), and we added a small-sample guard — at 7 positives an F1 of 0.667 is a split a single branch flips, and no predictor generalizes from seven examples. Inconclusive, underpowered; the discipline is the result.

**Boundary 5 — relational composition (a cheap relational rule wins; powered ML ties, no reproducible edge).** The powered rerun repairs every DGM-004C weakness: a genuinely relational, multi-hop goal (disease-associated via the ClinVar chain *and* sharing a Reactome pathway with the hub — a property no single annotation reveals), 7 cold-drawn hubs expanded by 500 branches to **735 branches with 33 positives** (base rate 0.045), a stratified 50/50 held-out split with 16 training positives, fit-on-train/score-on-test, 1000× bootstrap. Cheap *category* rules top out near F1 0.40 — the relational goal defeats them, as designed. But a cheap *relational* rule rescues it: `has_disease_comment AND go_ov3` (≥ 3 shared GO terms with the hub) reaches TEST F1 0.345, and the relational feature alone (`go_ov3`) scores F1 0.283 at recall 0.941. The logistic-regression candidate, no longer starved, climbs from its underpowered F1 0.154 to **0.372** — essentially tied with the best rule — but the paired 90% CI on (ML − deterministic) is **[−0.064, +0.113]**, straddling zero. Against the three-condition gate the deterministic ceiling is inadequate (0.345 < 0.70), but the model neither routes economically (it keeps only 47.1% of true positives, where the cheap feature keeps 94.1%) nor beats the solver by margin — rejected on two of three conditions, now at fair statistical power.

| Boundary | Deterministic alternative | Break? | Selective? | Verdict |
|---|---|---|---|---|
| **1 · Operation selection** | run-all on a cheap free-API menu | no ($0, prize 19.4% wall, unreachable) | — | no ML — no cost problem |
| **2 · Depth by novelty** | exhaustive traversal; dedup + budget cutoff | yes (~14,496 calls @ depth 1) | no (56% of nodes for 70% of evidence) | no ML — no signal |

| Boundary | Deterministic alternative | Break? | Selective? | Verdict |
|---|---|---|---|---|
| **3 · Goal-relevance** | cheap UniProt-annotation filter | yes | yes, but cheap-category | no ML — cheap-filterable |
| **4 · Compositional goal** | 3-feature boolean rule (F1 0.667) | yes | yes (7 positives) | inconclusive — declined (underpowered) |
| **5 · Relational composition** | cheap relational rule `go_ov3` (F1 0.345, recall 0.941) | yes | yes (33 positives, powered) | no ML — ties, CI [−0.064, +0.113], no reproducible edge |

Five boundaries measured; the deterministic baseline held five times — the fourth only provisionally (underpowered), the fifth confirmed at fair statistical power. The sharpened, falsifiable thesis the sequence arrives at: *inference is justified only when an economically consequential, selective boundary resists sufficiently capable cheap deterministic AND relational/graph-query computation, AND learned inference demonstrates a reproducible advantage (bootstrapped, held-out) over those alternatives.* The sequence is frozen.

## 4 · Track 2 — Uncertainty on inference's most favorable terrain

Track 1 asked where inference does not belong. Track 2 drives onto the terrain that favors deterministic systems least: genuine, cost-weighted, partially-observed decision-making, where the objective is not "which fact exists" but "which future measurement most reduces uncertainty over competing hypotheses." A *schema-boundedness* insight frames it: because the floor authors only typed edges from a fixed relation set, the set of verifiable hypotheses is enumerable, so a model can add value only by escaping that enumeration — through combinatorial hypothesis spaces, prior-dependent hypotheses, or genuine belief uncertainty. Information-

gain planning is the last of these, and the one where a learned prior over unseen evidence could plausibly repair an ill-specified planner.

**The task.** Diagnostic identification of a hidden target among **160 human kinases** over **2,775 floor-verified features** (median 41 per candidate), under a real per-feature cost model (interpro 1 / pathway 1 / disease 2 / drug 2 / kegg 3). A measurement asks "does the target have feature *f*?"; the floor returns the true answer and eliminates inconsistent candidates. The deterministic opponent is a strengthened **cost-aware expected-information-gain (EIG)** planner — `argmax_f IG(f) / cost(type(f))` — reaching, well-specified, 7.64 queries against the $\log_2(160) = 7.32$ information-theoretic optimum; on a 40%-masked reference it needs 18.7 queries / 19.66 cost at correct-ID 1.000.

**Prior injection into admission (rejected).** A frontier model with *zero truth authority* predicted the 2,875 masked entries of the 44 discriminating features at **85.8% accuracy** against sealed truth — a genuinely good prior, matching the true base rate. Injected into the planner's *admission* (elimination) step it cut total measurement cost from 19.66 to **12.10 (−38.4%)**, paired 90% CI [6.65, 8.51] — but correct-identification **collapsed from 1.000 to 0.753**. A diagnostic showed **100% of the 39 mis-identifications were caused by a wrong imputation of the true target's own masked entry**: a single wrong cell eliminates the true target, and the floor had no fact to check the guess against, because a database entry is prior knowledge, not a measurement. Authority is binary, not an efficiency knob — **an 86% model in the truth path makes the system ~24% wrong** — and the gate rejects.

**The verified-card A/B (the decisive attribution).** The illegitimate move was filling gaps with the model's parametric memory instead of retrieving verified cards. Re-running the identical experiment and swapping *only* the data source attributes the "fill missing state" value to whichever source owns it. Verified cards beat the model's guess on **both** axes — cheaper *and* correct.

| Lane (fills the same 2,875 masked cells) | Mean cost | Correct-ID | Verdict |
|---|---|---|---|
| Baseline cost-aware EIG (no fill) | 19.66 | 1.000 | reference |
| Claude memory guess (86% accurate) | 12.10 | 0.753 | REJECT — corrupts truth |
| ML search-only ranking (admission protected) | 18.86 | 1.000 | REJECT — −4.06%, below 15% bar |
| **Verified-card retrieval** | **11.50** | **1.000** | −41.5%, CI [7.10, 9.20], at perfect correctness |

Confined to *search* only — the prior ranks the next measurement while elimination uses only floor-verified entries, so correct-ID is 1.000 by construction — the legitimate benefit is **4.06%** (CI [0.114, 1.456]), real but below the bar. The decomposition is airtight: of the 38.4-point apparent benefit in the admission variant, ~34 points were bought by truth-corruption and only ~4 are the prior's legitimate contribution to *which question to ask next*. The entire "fill missing state" value belongs to verified retrieval; the model that manufactured state was a strictly-worse, truth-corrupting substitute for cards already held — a verified card is 100% or absent, never 86%.

**The coverage sweep.** Sweeping verified-card coverage 100% → 10% (nested seeded masks), three lanes per level over 160 leave-one-out instances, tests whether ML ranking earns a seat as knowledge thins.

| Finding | Result |
|---|---|
| **Verified-card retrieval** | flat at **9.48 cost, 100% correct-ID at every coverage**; dominates by up to 6.6× at 20% (63.15 vs 9.48) |
| **ML-ranked EIG** | stays within ±5–8% of deterministic EIG at every level (+5.4% @ 50%, −7.6% @ 20%, +4.6% @ 70%, −5.1% @ 90%); never clears 15%, never with a robust CI — **crossover = none** |

| Finding | Result |
|---|---|
| **Resolution vs coverage** | 100% resolved at ≥ 70%; 33% at 20%; **0% at 10%** — as knowledge thins, resolution collapses, not ranking |

The measured boundary is cleaner than "ML helps when data is sparse": sparse verified knowledge is a *data* problem (fetch cards), not a *planning* problem. The seat for each component is now measured, not argued — established knowledge → verified cards; new evidence → measurement; truth/admission → the floor alone; search/ranking → the only ML candidate, worth ~0 here. The falsifiable open frontier is a domain whose *ranking* decision carries latent structure a deterministic planner cannot model — which Track 4 supplies.

## 5 · Track 4 — Generalization, with the formal model

Tracks 1–2 lived in molecular biology. Track 4 exports the entire method to a structurally different domain — sequential Bayesian fault diagnosis of legacy computing systems — to test whether Inference Placement is a general systems principle, and to state the formal model that makes the boundary measurable. The domain is engineered to *favor* learned inference: a single tunable knob dials it from "naive Bayes is exactly correct" to "the fault label cannot explain the correlations."

### 5.1 · The formal model

The task is sequential Bayesian diagnosis among **K = 12** fault classes, using **M = 20** cost-weighted binary diagnostics, with **L = 4** latent factors. Let $f$ be the true fault with prior $P(f)$, and let $z \in \{0,1\}^L$ be latent factors, each independently active with probability $\pi_l$.

**Generative model (noisy-OR with a latent-structure knob ρ).** The probability that diagnostic $m$ returns positive given the fault and the latent factors is

$$P(o_m = 1 \mid f, z) = 1 - (1 - \theta_{f,m}) \cdot \prod_{l\,:\,z_l=1,\, w_{l,m}=1} (1 - \varrho),$$

where $\theta_{f,m} \in [0,1]$ is the fault's per-test *signature* (the marginal a naive model estimates), $z$ are the latent factors, and $w_{l,m} \in \{0,1\}$ is the factor–test incidence matrix. The knob **ρ** governs how strongly an active latent factor drags a wired test toward positive: at **ρ = 0** the product is 1, so $P(o_m=1 \mid f, z) = \theta_{f,m}$ — outcomes are conditionally independent given $f$, and naive Bayes is *exactly* correct (the built-in control the harness reproduces to a 0.0% gap); as **ρ → 1** the latent factors induce cross-test correlations the fault label cannot explain.

**Planner (identical across all lanes).** Every lane drives the same cost-aware expected-information-gain policy. With belief $b$ over the K faults and accumulated observations $O$, the expected information gain of test $m$ is

$$\mathrm{EIG}(m) = H(b) - \mathbb{E}_{o_m}[\, H(b \mid O, o_m) \,],$$

where H is Shannon entropy over the belief. The planner selects

$$m^{\star} = \operatorname{argmax}_m \ \ \mathrm{EIG}(m) / \mathrm{cost}(m),$$

observes the machine's *true* outcome, updates $b$ by Bayes, and repeats until $\max_f b(f) \geq 0.90$, then diagnoses $\operatorname{argmax}_f b(f)$. A lane only ever picks the next test; the floor reveals the true outcome and the true fault, so a lane cannot fabricate a diagnosis — only choose a (possibly sub-optimal) measurement order. The lanes differ only in the likelihood each uses: naive-Bayes with estimated marginals (the realistic opponent), naive-Bayes with exact marginals (the clean control isolating the value of modeling the joint), the **oracle-joint** (the true generative model — the upper bound on any learned model), and a learned Bernoulli mixture (§5.4).

**Operational objective.** For the structure sweep the objective is diagnostic cost alone. From the priced-error experiment onward the pre-declared objective *prices misdiagnosis*:

$$\mathrm{loss} = \mathrm{diagnostic\_cost} + \lambda \cdot \mathbb{1}[\mathrm{misdiagnosis}],$$

where λ is the operational cost of a wrong diagnosis in units of one diagnostic test.

### 5.2 · The structure-isolated sweep

The clean test compares oracle-joint against naive-Bayes with exact marginals — both know the exact marginals, so the entire gap is the value of modeling the latent joint. The ρ = 0 tie is the harness's unbiasedness certificate.

| ρ | naive-exact cost (acc) | oracle-joint cost (acc) | structure Δ (cost) | paired CI90 (naive−oracle) | clears 15%? |
|---|---|---|---|---|---|
| 0.0 | 8.42 (0.935) | 8.42 (0.935) | 0.0% | [0.0, 0.0] | no (control tie) |
| 0.2 | 11.99 (0.915) | 11.88 (0.915) | 0.96% | [−0.26, 0.52] | no |
| 0.4 | 15.06 (0.875) | 15.88 (0.890) | −5.44% | [−1.29, −0.35] | no |
| 0.6 | 17.64 (0.840) | 17.57 (0.855) | 0.43% | [−0.88, 1.00] | no |
| 0.8 | 21.04 (0.770) | 18.82 (0.835) | **10.53%** | [1.33, 3.14] | no |

Crossover: none. Even a perfect joint model never clears the 15% cost bar at any ρ. At ρ = 0.4 the oracle is transiently *worse* on cost (it spends to disambiguate correlated symptoms) while slightly more accurate — mechanistically sensible, not noise. At ρ = 0.8 the first CI-significant structure signal in the whole program appears: the oracle is both cheaper (18.82 vs 21.04, CI [1.33, 3.14]) and more accurate (0.835 vs 0.770 — a ~28% reduction in diagnostic error) — real and growing, but sub-material by the cost bar.

### 5.3 · The priced-error objective – the first cleared ceiling

The ρ = 0.8 signal is real but sub-material on the cost-only objective because its value is in *being correct*, not spending less. Under the pre-declared priced-error objective, the pre-registered λ grid produces a clearing region of exactly two cells — both at ρ = 0.8 — the first time in the whole program any learned model, at its theoretical ceiling, clears a strict pre-registered bar. Pre-registration held exactly: at ρ = 0.8, λ = 25 gives 14.33%, below the bar; it is not crossed until the declared λ = 50.

| ρ \ λ | 0 | 5 | 10 | 25 | 50 | 100 | acc (naive→oracle) |
|---|---|---|---|---|---|---|---|
| 0.2 | 0.96% | 0.93% | 0.90% | 0.81% | 0.71% | 0.56% | .915 → .915 |
| 0.4 | −5.44% | −4.75% | −4.11% | −2.45% | −0.33% | 2.47% | .875 → .890 |
| 0.6 | 0.43% | 0.81% | 1.17% | 2.08% | 3.22% | 4.68% | .840 → .855 |
| **0.8** | 10.53% | 11.45% | 12.28% | 14.33% | **16.79%★** | **19.79%★** | **.770 → .835** |

A starred cell clears the 15% bar with paired CI > 0 ([2.905, 8.03] and [4.3, 13.715]). The seat is a corner, not a region, and it maps the division of labor exactly: the floor owns truth; the learner would model the joint structure the floor's independence assumption lacks. But the oracle is a *perfect* model — the ceiling clearing is **necessary, not sufficient**.

### 5.4 · The finite learner and the learning curve

Is the corner occupiable by a *realizable* learner? A finite-sample joint model (per-fault Bernoulli mixture, C = 8 latent classes, 40 EM iterations, Laplace smoothing, seeded init; 1,500 training + 200 held-out incidents, no oracle leakage) at the frozen corner (ρ = 0.8, λ = 50) does **not** clear. A sanity check confirms the harness: the oracle-as-mixture reproduces the 16.79% clear exactly, so the null result is the learner's, not the instrument's.

| Lane (corner ρ = 0.8, λ = 50) | accuracy | mean cost | op-loss @ λ=50 |
|---|---|---|---|
| naive-Bayes (estimated) | 0.755 | 21.4 | 33.62 |
| naive-Bayes (exact marginals) | 0.770 | 21.0 | 32.54 |
| **learned Bernoulli mixture (EM)** | **0.770** | 21.0 | **33.51** |
| oracle-joint (ceiling) | 0.835 | 18.8 | 27.08 |

Against the primary opponent the operational-loss reduction is **+0.34%** (CI [−2.28, 2.49], straddling zero) — the learner captures only ~1.8% of the oracle's edge. A

learning curve then asks whether this is data starvation. It is not: adding data helps but plateaus below the bar.

| N (incidents) | learned acc | reduction vs naive | paired CI90 | % oracle captured | clears 15%? |
|---|---|---|---|---|---|
| 1,500 | 0.770 | 0.34% | [−2.31, 2.30] | 1.8% | no |
| 3,000 | 0.765 | 1.47% | [−2.03, 3.02] | 8.5% | no |
| 6,000 | 0.770 | 4.34% | [−1.20, 3.79] | 25.7% | no |
| 12,000 | 0.785 | 7.04% | [−0.15, 4.61] | 42.8% | no |
| 24,000 | 0.805 | 7.0% | [−0.44, 4.82] | **47.8%** | no |

The captured-percentage gains decelerate hard (+6.7, +17.2, +17.1, then only +5.0 for the final doubling); the loss reduction is flat at ~7% by N = 12k–24k (the last doubling moved it 7.04% → 7.0%); it never clears 15% and every paired CI includes zero. Accuracy climbs monotonically 0.770 → 0.805 (oracle 0.835): the learner is data-responsive but plateaus below the bar. The marginal value of data has collapsed to near zero by 24k while the learner still recovers only ~48% of the opportunity — so the binding constraint is **model class, not data budget**. The verdict is a *theoretical seat only*: the door the ceiling opened exists, but this realizable learner does not turn the key. The next gate — a structure-matched latent-factor learner — is named as open and has not been run; no claim rests on its outcome.

## 6 · The three-gate framework

Across all three tracks the durable contribution is not "symbolic beats ML" but a reformulation of Inference Placement: the decision to grant a learned layer authority is not binary but a **conjunction of three empirically measured gates.**

> **Framework (earned deployment of learned inference).** Deploy learned inference at a boundary only when **all three** gates are met.
> **Gate 1 — Theoretical value.** Does an inference opportunity exist at all? Measured by the *oracle ceiling* — the true generative model, the upper bound on any learned model. *Answer: yes, in exactly one corner* (strong latent structure and a priced-error objective: (ρ=0.8, λ=50)=16.79%, (ρ=0.8, λ=100)=19.79%). Everywhere else across all three tracks, the ceiling itself does not clear the bar.
> **Gate 2 — Realizable recovery + data price.** Can a real learner recover it, and at what evidence budget? Measured by the finite learner and its learning curve. *Answer: partially, plateauing below the bar* (~1.8% of the oracle edge at N=1,500, rising to 47.8% of the opportunity by N=24,000 but flat at ~7% loss reduction, never significant). Data is *not* the binding constraint.
> **Gate 3 — Model class.** Is the limiter the representation? Measured by a structure-matched learner. *Answer: open* — named and pre-registerable but not run; the learning curve identifies representation as the live hypothesis without claiming its outcome.
> Across everything tested, this conjunction is **not** fully met: Gate 1 is met in one corner; Gates 2–3 are not met by any learner tried, and Gate 3's frontier is still open.

*The door exists; the real key does not turn it.*

The full map places every measured boundary against the gates. Everything above the last row resolves deterministic; the last row is the one place a seat provably exists in theory — and it is unearned in practice.

| Boundary (track) | Deterministic alternative | Learned-inference seat |
|---|---|---|
| Operation selection (Placement) | run-all on a cheap free-API menu | none — no cost problem to solve |
| Combinatorial depth (Placement) | exhaustive traversal; uniform novelty | none — yield is not concentrated |
| Selectivity / goal-relevance (Placement) | cheap category rules | none — cheap-filterable |
| Relational composition (Placement, powered) | cheap relational graph-query rule | none — ML ties, no reproducible edge |
| Ordinary uncertainty (Uncertainty) | verified-card retrieval + deterministic EIG | none material — ranking worth ~0; a model in the truth path corrupts at its error rate |
| **Latent structure + priced error (Generalization)** | naive-Bayes EIG | **a seat exists at the oracle ceiling but remains unearned by a realizable learner** |

The framework yields an epistemic-status rule for *allocating computation by the epistemic status of the information, not by which technology is available*: information already **known** is retrieved from verified cards, not inferred; information **unknown but measurable** is measured (a card is fetched), not hallucinated; information requiring **verification** is admitted only by the deterministic floor — the model never decides what becomes true; and only information requiring **search over genuine uncertainty** is even a candidate for learned inference, which must then earn its place empirically, one measured boundary at a time.

*Known → retrieve. Unknown-but-measurable → measure. Needs-verification → the floor. Genuine uncertainty → test whether learned inference improves the search.*

## 7 · Two scoreboard failures the method caught in its own instruments

The program's thesis — that neither an optimizer maximizing a number nor a scoreboard reporting a ratio can certify that the number tracks the goal — was turned on the program itself. Twice, an instrument lied, and both times the lie was caught by inspection and independent connector testing, not by trusting the number. Both are retracted.

**A gameable fitness that rewarded doing less.** An early evolutionary phase optimized a pure ratio — verified-discoveries per operation-run. Greedy search on that ratio produced a policy that "won" the ratio on the sealed pool (24.25 vs the incumbent's ≈3.0) *by doing less*: it skipped productive operations and whole proteins, recovering only **97 verified facts on the sealed pool versus the incumbent's 148**. A ratio with no discovery floor rewards recovering *fewer* real facts while raising the average — the reward-hacking failure mode that makes self-improving systems dangerous [5], reproduced in miniature on a symbolic policy. The fix was a discovery-preserving fitness (reward total verified discovery; penalize only operations that yield nothing).

**A blind measurement that produced a false conclusion.** The reward oracle counted a beyond-default discovery only if *both* subject and object were absent from the default extract. But every beyond-default fact is `protein → X`, and the protein is always in the default set, so the `subject not in A` clause silently zeroed every such discovery. It was caught when a pathway query reported OK yet *zero* new discoveries, while dumping the edges showed three provenance-backed edges the oracle had filtered out. Under the corrected oracle all operations are productive and vary by protein; the false "build new connectors" conclusion — the connectors already worked — was explicitly retracted, and the corrected run-all baseline recovers 323 verified facts on the sealed pool versus the 148 the bug had induced (≈ 2.2×).

An optimizer cannot define what its number should mean; a scoreboard cannot certify that a ratio tracks the goal. Only an independent verifier can — the thesis, applied to the program's own instruments. The superseded artifacts are preserved unedited as wrecks under correction banners, with the correction record authoritative.

## 8 · Limitations and scope

These are stated so the result cannot be over-read; the honesty is part of the result.

- **Small N; single target per stress class.** This is a characterization curve being built, not a peer-reviewed result and not a general proof. Placement targets are single proteins per run; pools in the evolutionary and uncertainty phases are 4–5 proteins each (the diagnostic-identification task uses 160 candidates but a single frozen instance). No population or statistical-superiority claim is made.
- **Results are bounded by the pre-registered tasks, datasets, and protocols.** Track 2's negative result is precisely: under this pre-registered task, dataset, masking procedure, learned prior, planner implementation, and evaluation protocol, no material ML-ranking advantage was observed — not that learned inference cannot improve search over uncertainty in general.
- **Generalization is shown by one further domain, not proven in general.** Track 4 is a single synthetic generative family (noisy-OR latent factors), one simulator, one cost model, one learner class per experiment, K=12 / M=20 / L=4, 200 held-out incidents per evaluation. The domain is engineered to *favor* learned inference and the learned lane is handed its theoretical ceiling for the value question — a deliberately generous setup. Further domains would strengthen or refute the claim.
- **The seat found is theoretical (oracle) and unearned by a realizable learner.** Gate 1 is met only at the oracle ceiling, in one corner. The finite-learner and learning-curve results are about *this* learner, *this* data budget, and *this* protocol — not a proof that no learner ever could occupy the seat.

- **The open next gate is model class.** A structure-matched latent-factor learner (Gate 3) is named and pre-registerable but has not been run; no claim in this report depends on its outcome. If it clears where the mixture plateaued, the seat is earnable; if it too plateaus, the deterministic win is total even at the one boundary the oracle opened.
- **Self-estimated ML costs and a model-tier note.** Where frontier-model token and dollar figures appear they are self-estimated lower bounds (input tokens were not separately metered); the decisive cost claim is qualitative (model-calls > 0 and $ > 0 versus 0 / $0), not the precise magnitude.
- **Integrity is tamper-evident, not tamper-proof.** Frozen artifacts carry unsigned SHA-256 digests that detect in-place edits by a party lacking the trusted digest set; they do not provide authenticity, non-repudiation, or an external anchor. Signing and anchoring are future work.

## 9 · Conclusion

We asked not whether machine learning is good but where, if anywhere, learned inference earns decision authority in a system that keeps a deterministic symbolic floor as the sole author of facts — and we answered it the only way a verification-first program can: by building the strongest deterministic method first, handing ML its oracle upper bound, pre-registering the bar before each run, and refusing to let any lane certify its own truth. Across five placement boundaries, an uncertainty task on inference's most favorable terrain, and a generalization domain built to favor the model, the deterministic (and retrieval) baseline held — until a theoretical seat finally appeared at the oracle ceiling that no realizable learner turned into an earned one. The result is a measured map, not an adjudication: a conjunction of three gates — theoretical value, realizable recovery at a supportable data price, and adequate model class — that has not been met anywhere tested, and the specific, falsifiable conditions under which it finally would be. The program even turned its thesis on its own instruments, catching and retracting two scoreboard failures. The

door exists; the real key does not turn it — and the map says exactly where to look for one.

## References


1. Perslis Research. "Peel: Structural Hallucination Prevention for Offline AAC Through Symbolic Fact Authorship." Preprint, 2026. research.perslis.com/peel.html
2. Perslis Research. "Retrieval Is Not Memory: Memory as a Governance Function over Experience." Preprint, 2026. research.perslis.com/memory.html
3. Perslis Research. "Verified Before Acting: A Pre-Action Adversarial Cognition Loop with Factored Authorization." Preprint, 2026. research.perslis.com/adversarial-loop.html
4. Perslis Research. "Traversing Data in Symbolic Systems: Typed-Relation Traversal as a First-Class Retrieval Primitive." Preprint, 2026. research.perslis.com/traversal.html
5. Perslis Research. "The Orchestration Gap: Why Model-Level Alignment Cannot Survive Multi-Model Runtimes." Preprint, 2026. research.perslis.com/orchestration-gap.html
6. Lewis, P. et al. "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks." *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. arXiv:2005.11401. `Establishes the retrieval-augmented-generation baseline in which the model remains the author of the final claim.`


## How to cite

Perslis Research. "Inference Placement: Where Learned Inference Earns Authority in a Symbolic System." Preprint, 2026. https://research.perslis.com/inference-placement.html

```
@techreport{perslis_inference_placement_2026,
  title       = {Inference Placement: Where Learned Inference Earns
                 Authority in a Symbolic System},
```

```
  author      = {{Perslis Research}},
  institution = {Perslis Research},
  type        = {Preprint (moat-scrubbed)},
  year        = {2026},
  url         = {https://research.perslis.com/inference-placement.html}
}
```

Preprint · not peer-reviewed · moat-scrubbed · Perslis Research · 2026-09-21 · every figure traces to a frozen, hash-recorded result artifact.

## ◈ Perslis Research

The AI operational runtime for your existing systems — model-agnostic, verification-first, bring-your-own-key.

```
Beta testing running — sign up today
```

RESEARCH

- Inference Placement
- Traversing Symbolic Systems
- Verified Before Acting
- Peel
- Retrieval Is Not Memory
- The Orchestration Gap

PERSLIS

- Research library
- Symbolic AI
- Evolution
- Liberate
- Home

COMPANY

- Contact
- Careers

LANGUAGE

- English
- 中文