ORIGATE: DIAGNOSING PROVENANCE INVARIANCE IN LEARNED ACTION REVIEWERS

Anonymous authors
Paper under double-blind review

ABSTRACT

An invariant pooling operator need not preserve the semantics of the objects placed inside it. ORIGATE diagnoses this distinction in action review by deriving paired tests from a declared provenance policy, inspecting intermediate representations, and intervening on a frozen encoding path. Neutral empty-slot handling removes sensitivity to repeating one source but does not equate a compound source with its partition across carriers. A local join intervention eliminates that partition discrepancy while preserving all trainable parameters, yet substantial policy errors remain. Matched-support fits further separate exposure to a source union's labels from exposure to its carrier encoding. When only split mixtures were trained, the intervention raises compound-input accuracy from about 80% to above 99.98%; when only compound mixtures were trained, it raises allowed-request false denial from near zero to 42.91% and 13.82% in two controls. A 192-request local study exercises these mechanism pairs through genuine context, process, and state entries while authorization remains independent. The results identify when encoding repair transfers a useful decision and when it transfers an erroneous readout. They concern trusted typed metadata and a finite policy, not language understanding or autonomous-agent safety.

1 INTRODUCTION

A reviewer can receive the same resolved operation selected by a trusted host or by repository-controlled configuration. Its policy may permit the former and require a grant for the latter. Conversely, moving the same selecting origin from a value to a branch, queued event, or restored state need not change the required disposition. Which differences matter is a property of the policy, not of an aggregation operator alone.

This creates an object-definition problem. A network that is invariant to permutations of carrier vectors need not be invariant to repartitioning the origins inside them. Empty carriers can also become learned elements whose disappearance changes a representation. Correcting either behavior does not ensure that the readout has learned the policy on a previously unobserved conjunction. Ordinary accuracy leaves these explanations entangled.

ORIGATE turns them into separate, falsifiable checks: validate the policy relation; compare representations, scores, decisions, and labels; then change one encoding path while keeping trainable parameters fixed. Matched training conditions distinguish missing union-label support from a missing compound element. Crucially, the same intervention can help one trained readout and hurt another. An encoding identity is not an accuracy guarantee.

Our evidence combines controlled learning, complete per-fit paired interventions, and local requests that directly instantiate the diagnosed layouts. The classifier approximates a known policy from trusted metadata; it does not discover origins, understand commands, or authorize execution. The contribution is a reusable diagnostic sequence and its mechanism-specific evidence, including failed repairs and strong baselines that match fixed closure, rather than a new universally superior set architecture.

We distinguish guarantees from observations and stratify transferred errors: an invariant interface is not necessarily a better reviewer.

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.

2 POLICY-RELATIVE DIFFERENCES AND CONCRETE INPUT OBJECTS

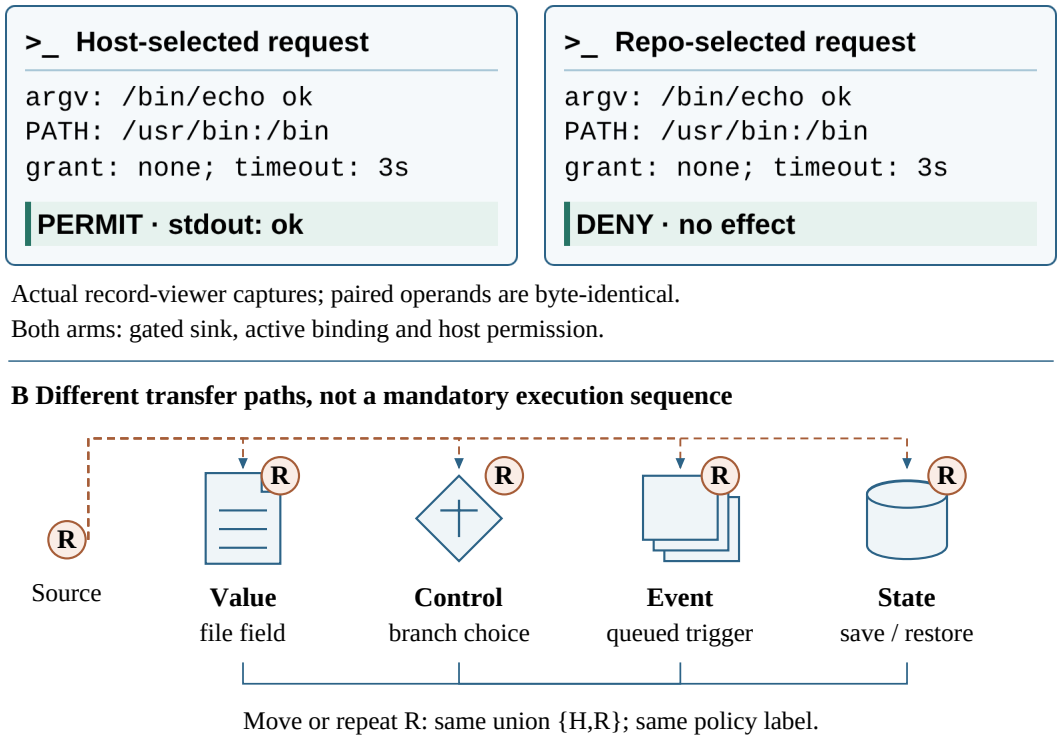


Figure 1: A: actual captures of a record viewer for a retained, operand-identical host/repository pair; these are not screenshots of a production agent. B: alternative provenance-transfer objects, not a mandatory execution sequence. Every path retains the selecting origin. Source records, viewer HTML, and the unabridged operands accompany the figure.

Let $x = (q, o, P_1, \ldots, P_4)$ contain trusted policy/grant facts q, operand descriptors o, and four carrier ancestry sets. The declared target is $y(x)$; the representation and readout give $s_\theta(x) = \sigma(\rho_\theta(q, o, z_\theta(x)))$ and $\widehat{y}_\theta = \mathbf{1}[s_\theta \geq 0.5]$. Parameters stay fixed within every intervention pair.

Preserve relevant source differences. Figure 1 holds the complete operands fixed while changing selection source. If an encoding maps an equally weighted opposite-label pair to one input, every deterministic classifier errs on at least one arm. This elementary collision check tests information loss, not statistical generalization.

Ignore only policy-irrelevant changes. For fixed q, o, union-preserving inputs satisfy $\bigcup_j P_j = \bigcup_j P'_j$. Relocation, repetition, and repartitioning are valid invariance tests only when the policy is constant on this relation. Representation equality implies equal scores for a fixed readout, but neither equality implies a correct policy prediction. A carrier-sensitive policy may legitimately require different labels; unionization would then discard necessary information.

A carrier is a container for selecting ancestry, not necessarily an atomic source. One record may package repository and unknown origins together, whereas another distributes them across value and control. Their physical fields differ but their union need not. This distinguishes a symmetry of the input objects from the policy relation the reviewer is expected to respect. We preserve exact common inputs for these tests rather than relying on similar commands or approximate record matching.

3 FROM POOLED OBJECTS TO A DIAGNOSTIC INTERVENTION

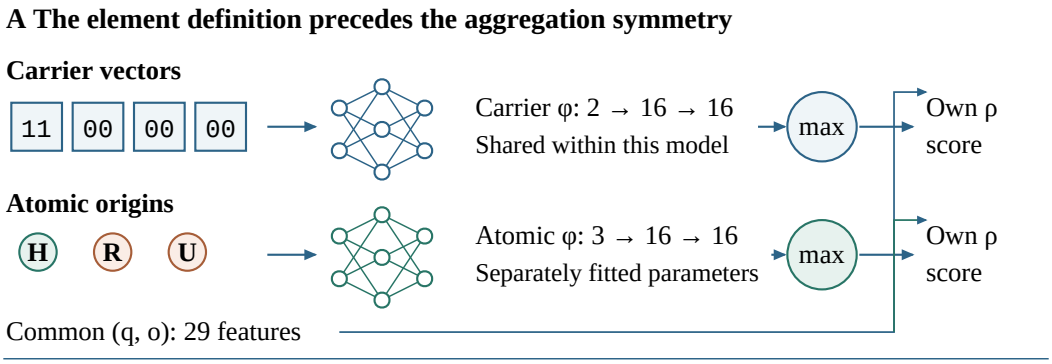


Figure 2: Separate encoders and readouts receive the same 29 common features after pooling. Witness: masked max / split-only / d0 / s7, exact-granted `ctx.read`; all 16 coordinates use one scale. The patch transfers this fit's split score. It can instead harm joint-trained fits (Figure 4). Scores never authorize effects.

Inputs to the diagnostic. A researcher supplies the policy and trustworthy labels, candidate transformations, a frozen reviewer, and access to its pooled representation. ORIGATE does not automatically infer the policy or authenticate an omitted source. Its executable checks follow four steps:

1. Verify whether the policy requires equal or different labels for each pair.
2. Record representation residual, score difference, decision flip, and both policy errors separately.
3. Test empty-element and join identities; intervene on the implicated encoding path without fitting to test labels.
4. Compare remaining errors under matched training support. Allow coexisting mechanisms or unresolved attribution.

The released tools implement pair construction, frozen inference, identity checks, interventions, and aggregation. Policy adequacy, trusted metadata, and the validity of a proposed deployment transformation remain researcher-specified premises.

A transformation is not accepted merely because a network ignores it. The policy first establishes whether its paired labels should match. A zero score difference then records response stability, while the two label comparisons separately record correctness. If an intervention removes a representation difference but leaves an incorrect decision unchanged, the diagnosis retains that failure instead of treating numerical equality as success.

4 POLICY, DIRECT CONTROLS, AND MATCHED EXPERIMENTAL SUPPORT

Write $r = \mathbf{1}[\mathsf{R} \in \bigcup_j P_j]$ and $u = \mathbf{1}[\neg\,\mathrm{Well}(\bigcup_j P_j)]$, where Well requires nonempty recognized ancestry. Active registered binding A, host permission H, and authentic, matching, fresh, unused exact grant V give

$$y(x) = A \wedge H \wedge \begin{cases} 1, & \text{context-only sink,} \\ 1, & \text{authority sink, } u = 0,\ r = 0, \\ V, & \text{gated sink, } u = 0,\ r = 1, \\ 0, & \text{otherwise.} \end{cases} \tag{1}$$

Known host authority does not universally require a new grant. Unknown ancestry rejects authority, but not the genuine context-only route. The exact kernel independently enforces these cases; no prediction enters its permit relation.

Representations. The common input has 29 features: four sink indicators, six host/active/grant predicates, thirteen binding equalities, and six non-policy operand descriptors. Four binary $(R, \mathrm{unknown})$ carrier pairs add eight coordinates. Known host ancestry is implicit; all-zero flags are not missing runtime provenance. Eight encoders compare concatenation, fixed closure, Deep Sets, slot max, masked max, zero-neutral max, atomic max, and expanded closure. Learned element maps use 16-unit layers; readouts use two 32-unit layers. Deep Sets learns both maps, not just raw summation (Zaheer et al., 2017).

Masked max omits empty slots but returns its learned $\phi(00)$ when all slots are empty in the executable baseline. Zero-neutral max uses bias-free ReLU layers: $\phi(00) = 0$ and all coordinates are nonnegative. Atomic max deduplicates supplied host/repository/unknown atoms. Expanded closure maps its two union flags through the same-size learned expansion as the carrier models. Table 2 gives parameter counts and semantics.

Fitted models and controlled training. We re-evaluate 240 full-input baseline fits: eight encoders, partial-pair or mixed-origin support, three data realizations, and five initializations. Partial support has only one origin at a time; mixed support inserts both in 10% of sampled rows. The targeted extension adds only masked and zero-neutral models under split-only and joint-only mixture supports, yielding $2 \times 2 \times 3 \times 5 = 60$ new fits. Both see the same mixed contexts, labels, and 10% selection stream. One exposes separate 10/01 carriers; the other exposes 11. Each fit uses 1,200 AdamW updates, batches of 256, learning rate 0.003, and weight decay 10^{-4} (Loshchilov & Hutter, 2019). Validation, never test performance, selects checkpoints.

Fresh evaluation and uncertainty. The frozen test has 4,096 contexts in 32 new generator groups. Each related view retains its context and group; mixed views contain 300 allowed and 3,796 denied contexts. False allow (FA) and false deny (FD) use these separate denominators. We retain all 15 crossed fits per condition, pointwise 95% intervals from 10,000 paired group resamples with fits fixed, and separate crossed-training intervals with the test fixed. These are different uncertainty targets. Appendix B distinguishes this test, historical results, and the source-version reconciliation.

A single evidence release. All current results use the pinned executable semantics and the 61000-series test, with a result-to-checkpoint manifest. Audits reproduce 300 main-input validation losses and six predeclared clean training runs; those reruns are checks, not additional study fits. Appendix J distinguishes history verification, training replication, and prior manuscript-only counts that are not quantitative evidence for this release.

Matching is explicit, not inferred from accuracy. Within each data draw, the two mixture supports consume the same sampled row identifiers and random choices at every update. Only the placement of the unknown flag differs on the selected mixed rows. Single-origin examples remain in both streams. This establishes equal context and label exposure without pretending that their occupied-carrier counts, gradients, or learned element geometry are equal. The full matching check and each selected validation step are retained with the corresponding checkpoint.

5 JOIN CONSISTENCY TRANSFERS A READOUT, NOT A CORRECTNESS GUARANTEE

For four binary carriers, unmasked coordinatewise max factors through their origin union exactly when

$$\phi(00) \le \phi(10), \quad \phi(00) \le \phi(01), \qquad \phi(11) = \max\{\phi(10), \phi(01)\}. \tag{2}$$

The inequalities require empty-element neutrality; the equality requires partition consistency. These are representation conditions, not necessary conditions for a particular readout to ignore differences. Appendix A proves the elementary join-homomorphism characterization. Masked and zero-neutral models eliminate single-origin repetition residuals, yet their partial-pair fits still flip 16.34% and 20.87% of joint/split decisions: a compound element remains a different learned object.

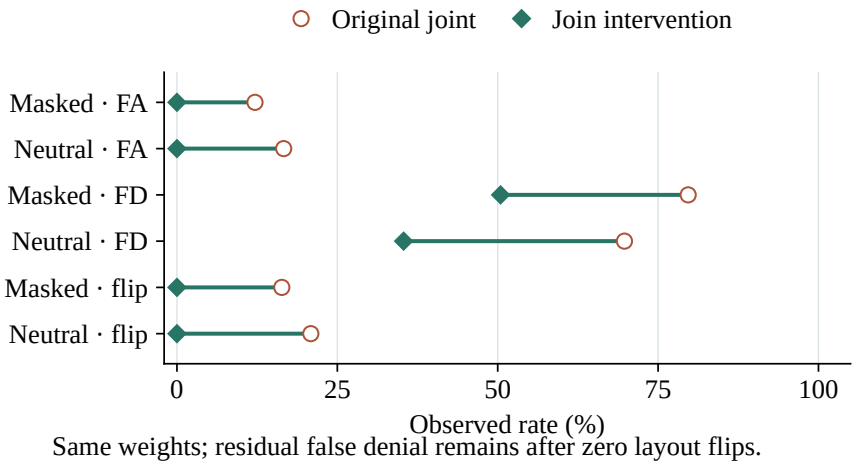


Figure 3: Original and intervened joint errors for partial-pair fits; flip compares joint with split. Rates average 15 fixed fits. FA conditions on 3,796 denied contexts; FD on 300 allowed contexts. The intervention removes layout flips, not every policy error.

A fixed-parameter path intervention. Replace only the inference output $\phi(11)$ by $\max\{\phi(10), \phi(01)\}$. Other elements, the empty fallback, common facts, readout, and threshold remain fixed. This changes the inference function without training or using targets. In the finite domain, patched joint must equal original split. All 120 configurations meet this identity to float32 tolerance; the maximum equal-union coordinate range over 30,720 assignments is 4.77×10^{-7}, with zero decision mismatches. Split and nonmixed outputs are unchanged.

The risk-transfer identity is a consequence, not a discovery. Let J, S denote paired original decisions, J^* the patched decision, and C_J, C_S their correctness indicators. Then $s_{J^*} = s_S$ and $C_{J^*} = C_S$, so

$$\Delta\mathrm{Acc}_J = \Pr(C_J = 0, C_S = 1) - \Pr(C_J = 1, C_S = 0). \tag{3}$$

The same identity holds within allowed and denied strata. It is not a new generalization theorem. The empirical questions concern which path was learned correctly, how often its errors are transferred, and how this depends on matched training support. In Figure 3, joint FA becomes zero but FD remains 50.44% and 35.33%: those errors already exist on the retained split path. Layout agreement alone would conceal them.

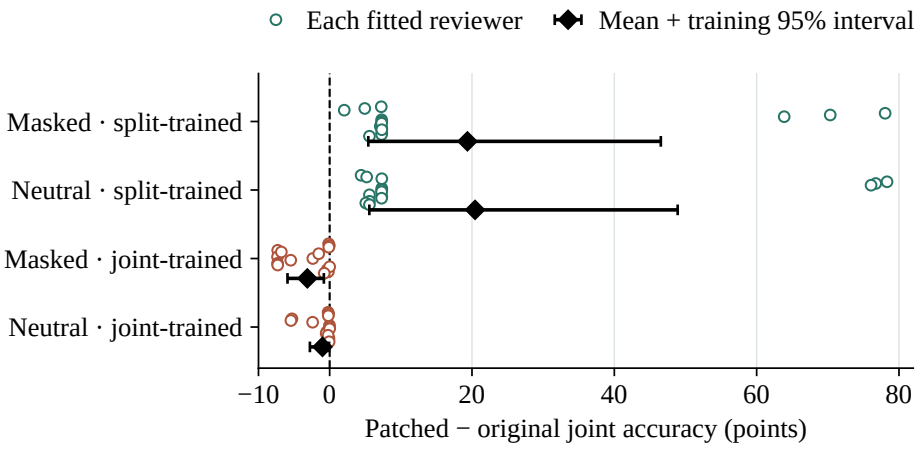
Which guarantee is being tested? Saturated sum and Boolean closure are equivalent controls in the binary domain; they are not competitors that this intervention must beat. Likewise, replacing the compound map with a fixed split reference makes score equality a consequence of the construction. The learning evidence is the original paths' different correctness under matched supervision, and the errors that survive or appear when that reference is imposed.

6 MATCHED LABEL SUPPORT CAN PRODUCE OPPOSITE TRANSFER DIRECTIONS

Without mixed origins, y and $y_{\text{alt}} = y \wedge \neg(r \wedge u)$ agree on training support but differ on allowed mixed contexts. To move beyond this unidentifiability example, both matched conditions expose the same mixed unions, labels, contexts, and 10% selection stream. Their only intended support difference is whether a mixture is packaged into one compound carrier or two separate carriers.

Table 1: Matched mixture supports. Both retain single-origin examples and the same update and validation protocol. Occupancy intentionally differs.

Training	Mixed origins during training	Held-out mixture view
Split-only	R and unknown in different carriers; no 11 element	Joint
Joint-only	R and unknown share one carrier; no split mixture	Split



Fifteen paired fits per row; zero means no change in accuracy.

Figure 4: All 15 paired fits in each matched-support condition. Horizontal position is joint accuracy change; small vertical offsets expose coincident fits. Diamonds and whiskers show mean and crossed-training 95% intervals, holding the 4,096-context test fixed.

Seen labels, unseen compound encoding. Split-trained masked and zero-neutral models reach 99.982% and 99.995% on split mixtures, but 80.62% and 79.56% on joint mixtures. Transferring the existing split path improves every fit. Mean gains are 19.36 and 20.43 points, but both medians are 7.30 points; full ranges are [2.05,78.05] and [4.42,78.32]. Large initialization-dependent gains should not be read as a typical twenty-point effect. The matched labels distinguish missing element/partition coverage from simply lacking union-label exposure.

The same intervention can transfer harm. Joint-trained models already achieve 99.992% and 99.998% on joint inputs. Mapping them to the weaker split path lowers accuracy by 3.13 and 1.01 points; each family has 14 worsened fits and one tie. More importantly, allowed-request FD rises from 0.11% to 42.91% and from 0.02% to 13.82%. Only 300 of 4,096 contexts are allowed, so overall accuracy dilutes this large conditional harm. The intervention enforces a representation identity, not an improvement recipe.

7 CORRECTED, HARMED, AND PERSISTENT ERRORS MUST BE SEPARATED

The transfer identity partitions every paired context into four disjoint states: both paths correct, only joint wrong, only split wrong, or both wrong. Under the patch, only-joint errors are corrected; only-split errors are introduced into joint; errors shared by both paths persist. We recount these cells for every fit and separately for allowed and denied requests, keeping the complete pairing and original thresholds.

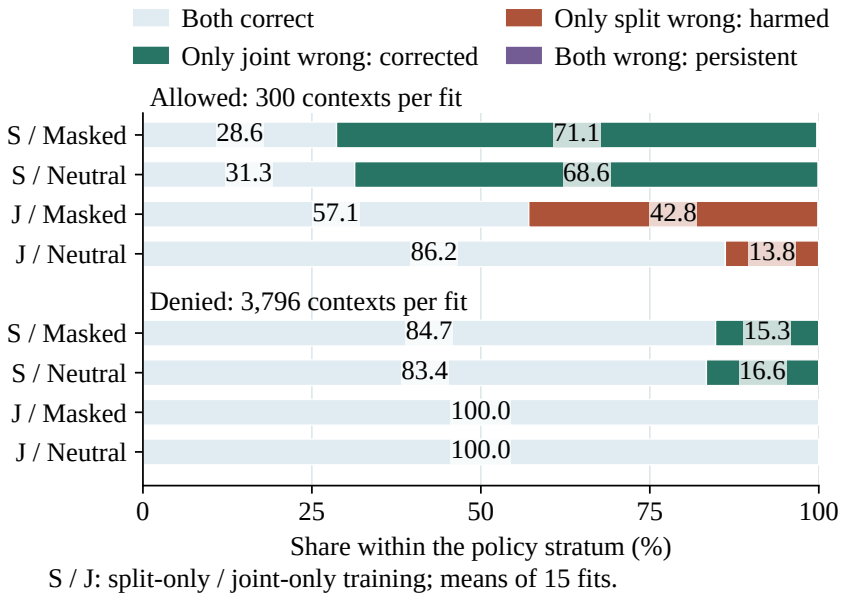


Figure 5: Four-way correctness decomposition for matched-support models. Each bar sums to 100% within its policy stratum and averages 15 fits. Small segments remain in the source data and Table 8; values are annotated only when a segment has room. Color distinguishes corrected, harmed, and persistent outcomes, not new executions.

A conditional loss can dominate a small aggregate decrement. For joint-trained masked max, 42.80% of allowed pairs are originally correct in joint and wrong in split; another 0.11% are already wrong in both. This yields 42.91% patched FD. The corresponding zero-neutral cells are 13.80% newly harmed and 0.02% persistent. Their denied pairs are all correct in both layouts. The apparent modest overall decrement is therefore a substantial selective loss, not broadly distributed mild degradation.

Aggregate improvement can also hide individual harm. Partial-pair masked max corrects 49.24% of allowed contexts but newly harms 20.00%, leaving 30.44% wrong in both paths. Its net allowed-request improvement coexists with a large damaged subset. Even in the split-trained conditions where every fit improves overall, one fit per architecture becomes slightly worse within the allowed stratum. Reporting only the number of improved aggregate fits would omit that distinction.

The reference path is the object of diagnosis. These cells are descriptive counts, not mutually exclusive explanations of hidden neurons. A policy-sufficient encoding can support a wrong readout, and several support shifts may coexist. A deployment would need evidence that its chosen reference is acceptable for the relevant policy strata; equality to that reference cannot supply such evidence by itself. Full per-fit cells, denominators, and source hashes are released rather than inferred from mean logits.

8 ACTUAL REQUESTS THAT INSTANTIATE THE DIAGNOSED MECHANISMS

We exercise the existing local reference adapter on 96 paired families: repetition or joint/split, four bounded payload variants, absent/exact/expired/spent grants, and three registered entries. Two views per family give 192 attempted requests. Complete operands, policy facts, origin union, and correct labels are identical within a pair. Provenance is deliberately injected through the test interface; this does not evaluate a production origin extractor.

Context means context, not relabeled authority. The existing `ctx.read` entry returns the bounded supplied value without a process launch or state mutation. The process entry runs a fixed host-owned echo program; the state entry writes the private SQLite store. We do not relabel these authority operations as context-only to obtain allowed mixed-source samples. All 180 frozen reviewers score before the kernel, alongside 90 fixed-parameter join paths. None of the 51,840 advisory scores enters authorization.

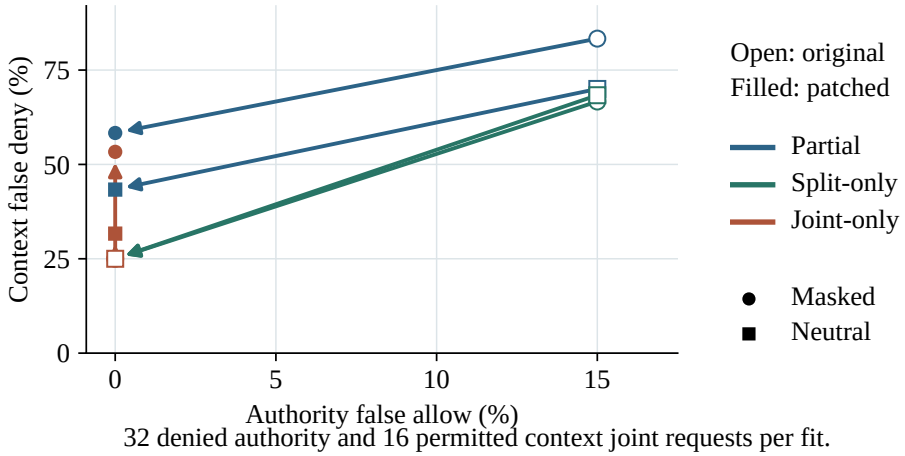


Figure 6: Error-plane trajectories on actual joint requests: open original scores to filled patched scores. Colors identify training support; shapes identify encoders. FA uses 32 denied authority attempts and FD 16 permitted context calls per fit. Lines are paired interventions, not time trends.

Mechanism transfer, with remaining errors visible. Partial-pair masked and zero-neutral models have 15% joint authority FA, which becomes zero after the join patch. Their genuine context FD falls from 83.33% to 58.33% and from 70% to 43.33%. Split-trained models retain 25% context FD after patching: every such error is in the absent-grant subset, whose recorded absent-grant verification fields are all zero. That common-fact pattern differs from training's independently sampled predicates. It is not remaining partition sensitivity. Joint-trained models again worsen under the same patch (Figure 6). Appendix G exposes grant strata and individual fits.

Execution is a separate outcome. The kernel admits 80 attempts: 64 context returns, eight process completions, and eight state writes. It denies 112 and no denied attempt produces an effect. All windows close and paired labels agree. There are 48 unique reviewer vectors and 32 policy/source patterns, not 192 independent tasks. The experiment connects the proposed mechanism to controlled current requests; it does not establish natural-attack prevalence, autonomous-agent success, or a neural security guarantee.

The residual context failures identify a recorded fact block, not one uniquely causal grant bit or neuron. Supplying the absent-grant pattern changes several common coordinates together. Separately, in 256 earlier controlled requests, only two of fifteen Deep Sets fits make false allows (16/144 and 32/144 denied requests); the others make none. Such concentration is why aggregate advisory rates, fitted variation, and observed effects are reported separately.

9 Related work and implications

Invariant operators require the right objects. Deep Sets learns element maps and invariant aggregation; PointNet uses shared point encoders and symmetric max pooling (Zaheer et al., 2017; Qi et al., 2017). Neither asserts that arbitrary carrier embeddings implement the union of the sources inside them. Our direct controls are small learned set models, not full reproductions of PointNet's geometric components. Fixed Boolean closure is a policy-specific reference and saturated sum is mathematically equivalent in this binary encoding. Retained decision-tree and adequate-support results match or nearly match it; no unique architectural advantage is claimed.

Canonicalization and a chosen reference. Ma et al. (2024) connect canonical forms with frame averaging for invariant and equivariant learning. Our binary-source patch likewise maps a compound input to the existing split representation, but union and repetition need not be invertible group actions. We do not transfer their group-theoretic conclusions to this domain. The contribution is not that equalized inputs yield equal outputs; it is the matched-support experiment and the conditional error transfer this normalization exposes.

Behavioral diagnosis and enforcement. CheckList organizes tests beyond ordinary accuracy (Ribeiro et al., 2020); causal confusion in imitation learning separates observational fit from intervention behavior (de Haan et al., 2019). Our procedure additionally validates policy relations, checks a specified join condition, and decomposes a fixed-parameter transfer. CaMeL, Progent, and Fides enforce explicit trust or privilege boundaries (Debenedetti et al., 2025; Shi et al., 2025; Costa et al., 2025). AgentDojo evaluates language agents in natural-language environments (Debenedetti et al., 2024). Their results are not reused as our baselines. Our kernel follows independent reference-monitor premises (Saltzer & Schroeder, 1975); advisory FA reduction is not a newly prevented execution.

10 Limitations and conclusion

The empirical domain is a declared Boolean policy with trusted grant facts, two origin flags, four carriers, and implicit host ancestry. The join characterization generalizes algebraically, not empirically, to larger source systems. Three training draws and five initializations expose considerable fitted variation but do not sample a population of algorithms. Conditional test intervals and crossed-training intervals address different targets, neither a security bound. Fresh generator groups are not external semantic validation.

Native admitted records contain only one policy/source pattern; their descriptor and grant variants are offline counterfactuals. The 192 mechanism requests use a local reference adapter, injected provenance, and bounded effects. Invariant split-trained scores still falsely deny 25% of context calls, localized to an absent-grant fact pattern, not uniquely to a bit or neuron. We evaluate no frozen language-model reviewer, production provenance extractor, natural-attack population, human study, or new inference-latency benchmark. A documented artifact correction excludes incompatible historical manuscript counts rather than treating them as additional evidence.

The representation lesson is to specify the policy equivalence and aggregated object before interpreting symmetry. Empty neutrality, partition consistency, and correct policy learning are distinct. Local join normalization inherits a reference readout; algebra guarantees that inheritance, not its benefit. Matched labels still yield unequal transfer across carrier encodings, and conditional paired counts reveal both corrected and newly harmed requests. ORIGATE makes these distinctions inspectable in frozen models and controlled actual requests while leaving authorization and unsupported causal attribution outside the learned score.

A reusable outcome is therefore a conditional diagnosis: retain policy-relevant distinctions, test a proposed equivalence, expose the encoded objects, and inspect the reference path before enforcing it. A failed equivalence test can justify an interface change, but the direction of its accuracy effect remains an empirical question. When labels or provenance are untrusted, that premise must be resolved independently rather than inferred from a model's confidence.

ETHICS STATEMENT

The study uses constructed metadata, retained controlled records, and harmless local requests in disposable private stores. It collects no human-subject data or production credentials. Metadata injection and counterfactuals are identified as such; passing a finite test is not a safety certification.

REPRODUCIBILITY STATEMENT

The artifact provides retained numeric checkpoints, 60 matched-support fits, frozen test arrays, full predictions, encoding interventions, actual request/effect records, and independent numerical checks. Editable vector sources and record-viewer materials support the concept figures. All experimental figures share one numerical specification with the self-contained MATLAB file; Python rendering was executed, native MATLAB was not. Appendices distinguish original evidence, reanalysis, six training-replication checks, and document checks.

AI USE STATEMENT

Generative AI assisted research framing, proofs, experimental design and implementation, analysis, writing, and figure preparation. Measurements derive from retained records or executed code; the authors remain responsible for verifying every claim and artifact before submission.

REFERENCES

Manuel Costa, Boris Köpf, Aashish Kolluri, Andrew Paverd, Mark Russinovich, Ahmed Salem, Shruti Tople, Lukas Wutschitz, and Santiago Zanella-Béguelin. Securing AI agents with information-flow control, 2025. URL https://arxiv.org/abs/2505.23643v2. Version 2.

Pim de Haan, Dinesh Jayaraman, and Sergey Levine. Causal confusion in imitation learning. In *Advances in Neural Information Processing Systems*, 2019. URL https://arxiv.org/abs/1905.11979.

Edoardo Debenedetti, Jie Zhang, Mislav Balunović, Luca Beurer-Kellner, Marc Fischer, and Florian Tramèr. AgentDojo: A dynamic environment to evaluate prompt injection attacks and defenses for LLM agents. In *Advances in Neural Information Processing Systems*, 2024. URL https://arxiv.org/abs/2406.13352.

Edoardo Debenedetti, Ilia Shumailov, Tianqi Fan, Jamie Hayes, Nicholas Carlini, Daniel Fabian, Christoph Kern, Chongyang Shi, Andreas Terzis, and Florian Tramèr. Defeating prompt injections by design. arXiv:2503.18813v2, 2025. URL https://arxiv.org/abs/2503.18813v2.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://arxiv.org/abs/1711.05101.

George Ma, Yifei Wang, Derek Lim, Stefanie Jegelka, and Yisen Wang. A canonicalization perspective on invariant and equivariant learning. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/702b67152ec4435795f681865b67999c-Paper-Conference.pdf.

Charles R. Qi, Hao Su, Kaichun Mo, and Leonidas J. Guibas. PointNet: Deep learning on point sets for 3d classification and segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017. URL https://arxiv.org/abs/1612.00593.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. Beyond accuracy: Behavioral testing of nlp models with checklist. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4902–4912. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.acl-main.442. URL https://aclanthology.org/2020.acl-main.442/.

Jerome H. Saltzer and Michael D. Schroeder. The protection of information in computer systems. *Proceedings of the IEEE*, 63(9):1278–1308, 1975. doi: 10.1109/PROC.1975.9939. URL https://ieeexplore.ieee.org/document/1451869.

Tianneng Shi, Jingxuan He, Zhun Wang, Hongwei Li, Linyu Wu, Wenbo Guo, and Dawn Song. Progent: Securing AI agents with privilege control, 2025. URL https://arxiv.org/abs/2504.11703v3. Version 3.

Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Ruslan Salakhutdinov, and Alexander Smola. Deep sets. In *Advances in Neural Information Processing Systems*, 2017. URL https://arxiv.org/abs/1703.06114.

A Formal conditions and interpretation of local interventions

Information loss before learning. For a finite sample, collect identical encodings $z = \psi(x)$ into cells, with n_{z0} and n_{z1} labels. Every deterministic predictor incurs at least $N^{-1}\sum_z \min(n_{z0}, n_{z1})$ error. Choosing either label errs on the other count; minimizing in each cell proves the statement. For continuous inputs the population analogue is $\mathbb{E}[\min(\eta(Z), 1-\eta(Z))]$, where $\eta(z) = \Pr(Y = 1 \mid Z = z)$. This elementary bound diagnoses an erased distinction. It is not a new statistical lower-bound technique or a test of near collisions.

Proposition 1 (Policy sufficiency)**.** *Under the trusted predicates of Equation 1, the target factors through (q, o, r, u). Every fixed readout of that representation is invariant to changes preserving the origin union and common input.*

Proof. The common guards and sink class depend on q. Context-only disposition is independent of origins after the guards. Every authority branch depends only on well-formedness, repository membership, and the exact-grant facts. These are the retained predicates. Equal inputs produce equal readout outputs regardless of training quality. □

This is sufficiency for the stated policy, not for policies that distinguish origin role, path, temporal order, or carrier identity. It also does not establish that the learned readout agrees with the target.

Proposition 2 (Carrier-union factorization)**.** *Let ϕ map subsets of an origin universe to $\mathbb{R}^d$, and let $k \geq 2$ be fixed. The coordinatewise representation $h(P_1, \ldots, P_k) = \max_j \phi(P_j)$ depends only on $\bigcup_j P_j$ for all assignments exactly when*

$$\phi(A \cup B) = \max\{\phi(A), \phi(B)\} \qquad \textit{for every } A, B. \tag{4}$$

Proof. For necessity, compare $(A, \varnothing, \ldots, \varnothing)$ with $(A, A, \ldots, A)$. Equal unions require $\phi(\varnothing) \leq \phi(A)$. Comparing $(A, B, \varnothing, \ldots)$ with $(A \cup B, \ldots, A \cup B)$ then gives the binary identity. For sufficiency, repeated application gives $h = \phi(\bigcup_j P_j)$. □

This elementary join-homomorphism characterization concerns representations, not a necessary condition for a particular readout to ignore representation differences. The empirical domain has only two flags and four carriers. Equation 2 is its finite reduction; all $4^4 = 256$ carrier assignments are enumerated.

Why the local patch has a restricted guarantee. Masked pooling removes empty elements whenever a flag-bearing carrier exists. In the all-empty case, the executable model's original learned $\phi(00)$ fallback is retained. Zero-neutral pooling makes the empty element zero and all other coordinates nonnegative. Replacing $\phi(11)$ by $\max\{\phi(10), \phi(01)\}$ then makes every nonempty assignment depend only on its flag union. Single-origin and empty-only assignments are unchanged. Compound inputs become the same pooled vector as their original split form. Float32 matrix kernels can introduce numerical roundoff, so implementation checks separately record residual tolerances and exact thresholded decisions.

This construction never reads targets. It does alter the inference function, and thus is accurately described as a fixed-parameter representation-path intervention rather than an unchanged model.

The guarantee is equality to a specified existing encoding, not an increase in accuracy. Joint-only training supplies a concrete counterexample to a universal performance benefit. Raw residual magnitudes remain parameterization-dependent; multiplying the representation by a positive constant and inversely scaling its incoming readout weights preserves predictions.

B DATA, IMPLEMENTED CONTROLS, AND MATCHED TRAINING

Generator and trusted facts. The retained generator uses seed $700000 + 1000i + j$ for group i and record j. Each group contains 128 contexts. Sink probabilities are 0.08 context, 0.22 blocked, 0.66 gated, and 0.04 unregistered; host/active flags have probabilities 0.94/0.96. Grant presence, authenticity, freshness, and unused status have probabilities 0.45, 0.92, 0.88, and 0.88. Equality bits are initially all one, with probability 0.20 of replacing one uniformly chosen coordinate by zero. The base source is host/repository/unknown with probabilities 0.40/0.50/0.10. Six descriptors are independent uniform $[0, 1]$. This is a synthetic policy-context generator, not a model of natural attack prevalence.

The 37-coordinate input contains: sink indicators (1–4); host, active, grant-present/authentic/fresh/unused predicates (5–10); thirteen binding equalities (11–23); six non-policy descriptors (24–29); repository flags (30–33); and unknown flags (34–37). Host ancestry is implicit. The thirteen equalities concern principal, repository, revision, workspace, release, contract, canonicalization, policy, sink, field, value, scope, and operands. Common inputs are provided verification facts, not inferred command semantics. Identifiers, target labels, and recorded outcomes are excluded from reviewer features.

Table 2: Actual full-input architectures. Learned element encoders have two sixteen-unit ReLU layers; readouts have two thirty-two-unit layers.

Reviewer	Parameters	Encoding
Concatenation	2,305	Four ordered flag pairs
Fixed closure	2,113	Two Boolean union flags
Deep Sets / slot max	2,881	Learned carrier map; sum / unmasked max
Masked max	2,881	Occupied carriers; learned all-empty fallback
Zero-neutral max	2,849	Bias-free nonnegative element map
Atomic max	2,897	Deduplicated one-hot host/repo/unknown
Expanded closure	2,881	Learned 2–16–16 expansion of union flags

The masked model returns a learned empty embedding only when every flag slot is empty, and otherwise ignores all empty slots. That fallback differs from the zero-vector description in the earlier manuscript snapshot and is recorded explicitly in Appendix J. It does not affect the single-origin repetition identity or the local join intervention: those views have at least one occupied carrier. Zero-neutral layers have no biases, yielding a zero empty encoding and nonnegative occupied encodings. Atomic pooling receives one-hot host/repository/unknown identities, includes host, and deduplicates atoms before learning; this object definition is supplied, not discovered.

Splits. Three retained training arrays have 8,192 contexts from groups $11000 + 1000d$ through $11063 + 1000d$, $d \in \{0, 1, 2\}$. Validation uses 2,048 contexts from $21000 + 1000d$ through $21015 + 1000d$. The new frozen test uses groups 61000–61031. It is disjoint from training, validation, the retained executable test 51000–51031, and the older discovery test 3000–3031. New generator groups confirm behavior within one synthetic distribution, not external semantic generalization.

B.1 SUPPORTS, UPDATE BUDGET, AND EXACT MATCHING

Every baseline partial-pair row places its origin into a uniformly selected mask from $\{1, 2, 4, 8, 3, 12\}$. Bits correspond to value, control, event, and state. Mixed support additionally selects 10% of sampled rows and independently places repository and unknown masks from that family, recomputing the target. The targeted split-only/joint-only extension keeps the ordinary augmentation and mixed-row selection matched. It draws one repository slot and an offset in $\{1, 2, 3\}$

in both regimes; unknown uses either the same slot or the offset slot. Thus both consume identical random streams while deliberately differing in mixed occupancy (one versus two).

We verify identical sampled row IDs, mixed selection, common inputs, and target labels at all 1,200 updates for each of the three data draws before fitting. Split-only never supplies an 11 element; joint-only never supplies a split mixture. Both retain single-origin examples so 10 and 01 are observed. The two regimes do not claim identical input geometry, only matched contexts, labels, mixture rate, update count, and random stream.

Initializations are 7, 17, 27, 37, and 47. Minibatch/augmentation seed is $81007 + 100d$, independent of model and initialization within each draw. Validation augmentation uses $90100 + d$, held fixed. AdamW uses 1,200 replacement batches of 256, learning rate 0.003, weight decay 10^{-4}. Validation BCE is recorded at update 1 and every multiple of 50; its minimum selects the checkpoint. No test condition or error selects a threshold, checkpoint, or support fraction. Numeric checkpoint arrays contain no executable pickles.

Table 3: Fresh evaluation populations per fit. Related views share the same 32 base groups and are not independent tasks.

Population	Contexts	Allowed	Denied
Ordinary	4,096	1,929	2,167
Same union, four carriers	4,096	1,929	2,167
Mixed joint	4,096	300	3,796
Mixed split	4,096	300	3,796

The primary fresh analysis recomputes four populations for 240 retained full-input fits and 60 new fits. Local join inference applies to 60 retained masked/zero-neutral fits (both supports) and 60 new fits. Each of these 120 configurations also evaluates all 256 carrier assignments at fixed common input. Fixed-context source probes vary 00/10/01/11 while holding every other coordinate fixed. All examples and views remain linked to their original context and group.

The artifact separately retains 120 no-descriptor and 45 exploratory missing-grant-support fits. They are not silently added as matched competitors to the 60 new mixture-support fits. Likewise, archived ordinary/test metrics are not pooled into the fresh 61000-series estimates.

C Neutral empty controls and retained direct repairs

The direct alternative succeeds. In an unmasked slot model, one repository-bearing carrier exposes $a = \phi(00)$ and $b = \phi(10)$, while four copies expose only b. The maximum can change because an element disappears, not because max violates idempotence. Appending that model's own a restores its original singleton representation with no weight change. From-scratch masking or a zero-neutral element map removes this cause entirely for nonempty single-origin repetition.

Fresh four-carrier accuracy of masked max is 99.993% under partial support and 100% under mixed support; zero-neutral gives 99.953% and 99.932%. Their remaining errors are already present on the corresponding singleton state. The low-dimensional fixed closure and atomic interfaces are also structurally invariant, while Deep Sets learns an extension that need not be invariant to cardinality. High accuracy is not evidence that all these models encode the same object.

The original empty-element repair remains retained on the retained executable test, with every fitted outcome shown in Figure 11. That earlier population and its training draws must not be substituted for the fresh experiment's errors or intervals. The targeted join repair below addresses a different identity even after empty neutrality succeeds.

D Every local join intervention and its uncertainty targets

Conditional test uncertainty. For fit m, group g, and state i, define event numerator a_{mgi} and eligibility d_{gi}. The reported rate is $\sum_{g,i}(15^{-1}\sum_m a_{mgi})/\sum_{g,i} d_{gi}$. A 10,000-draw bootstrap (seed

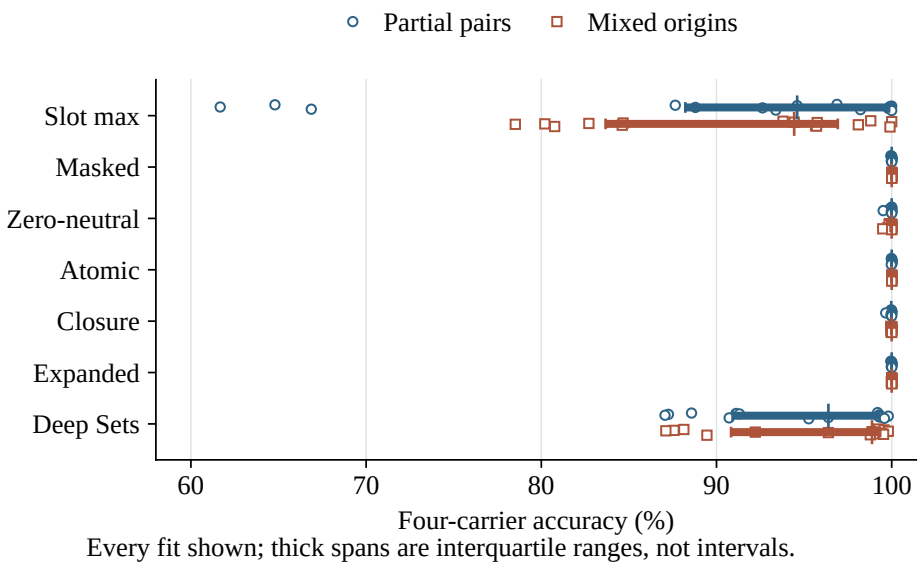


Figure 7: Four-carrier outcomes on the current test. Every point is one fit; thick spans are interquartile ranges and central marks are medians, not confidence intervals. Both supports retain all fifteen data/initialization combinations. The near-ceiling points are intentionally not separated by modifying their values.

Table 4: Fresh four-carrier accuracy (%) and complete fitted ranges. All estimates use the current 61000-series test.

Reviewer	Partial	Fit range	Mixed	Fit range
Concat.	88.31	87.65–91.31	89.30	87.65–91.41
Deep Sets	94.90	87.06–99.80	95.61	87.11–99.80
Slot max	89.68	61.67–100.00	90.91	78.52–100.00
Masked max	99.99	99.95–100.00	100.00	100.00–100.00
Zero-neutral	99.95	99.51–100.00	99.93	99.49–100.00
Atom max	100.00	99.98–100.00	100.00	100.00–100.00
Closure	99.96	99.66–100.00	99.99	99.95–100.00
Expanded	99.99	99.95–100.00	100.00	99.98–100.00

772310) samples the 32 test groups with replacement, preserving all paired states and model predictions for each selected group. Percentile 2.5/97.5 limits are pointwise conditional intervals, not simultaneous bounds or training-stability estimates. A zero observed event rate or degenerate interval does not establish zero population risk.

Crossed training variation. For paired accuracy gain Δ_{ds}, a separate bootstrap (seed 772311) samples three data indices and five initialization indices independently with replacement, then averages the Cartesian submatrix. The fresh test is held fixed. These 10,000 crossed draws estimate variation over this limited training design, not all future architectures or data distributions. They retain the same original/patched parameter pair. Both interval types are provided rather than treating narrow test uncertainty as robustness to training.

D.1 Policy errors that remain after exact partition repair

The intervention was specified without labels: the compound element output is set to the maximum of the two single-atom outputs. All original common coordinates, scores on nonmixed states, and split inputs are retained. The implementation's float32 batching makes the largest checked equal-union residual 4.77×10^{-7} rather than bit-exact zero. Every post-intervention joint/split thresholded

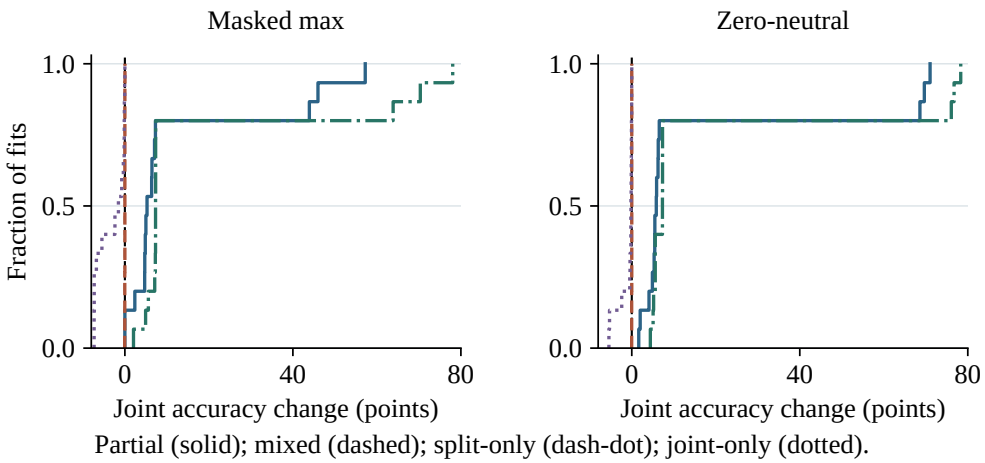


Figure 8: Empirical distributions of all fifteen paired accuracy changes for each architecture/support. Negative and near-zero changes remain in the numeric coordinates, avoiding rounded negative-zero annotations. Each step is an observed fit; the curves do not interpolate an unobserved training population.

Table 5: Joint accuracy gains and 95% intervals in percentage points. Conditional test and crossed-training intervals hold different factors fixed.

Reviewer	Support	Gain	Test interval	Training interval
Masked max	Joint-only	-3.13	[-3.45, -2.83]	[-5.92, -0.81]
Masked max	Mixed	-0.01	[-0.02, 0.00]	[-0.02, 0.01]
Masked max	Partial	13.41	[13.11, 13.72]	[3.00, 31.84]
Masked max	Split-only	19.36	[18.89, 19.80]	[5.43, 46.53]
Zero-neutral	Joint-only	-1.01	[-1.15, -0.87]	[-2.78, -0.08]
Zero-neutral	Mixed	-0.005	[-0.01, 0.00]	[-0.01, 0.00]
Zero-neutral	Partial	17.94	[17.69, 18.17]	[4.17, 44.03]
Zero-neutral	Split-only	20.43	[20.00, 20.85]	[5.58, 48.90]

decision agrees. These numerical checks are distinct from the algebraic identity and from policy accuracy.

Partial-trained models have little mixed-label support; their original split path still contains errors. Split-only mixture training teaches this path while leaving the compound element unseen, so transferring it is useful. Joint-only training instead teaches the compound path; replacing it by the original split path transfers that path's errors. The support intervention does not uniquely identify every gradient or neuron responsible for the difference. It demonstrates that enforcing one representation identity has a directional relationship to the trained readout, not an unconditional accuracy effect.

D.2 Risk transfer, four-way counts, and conditional harm

For an equal-label pair at fixed weights, the prescribed join path inherits the original split prediction. Equation 3 consequently holds for accuracy and for each policy stratum. This is an algebraic consequence of the mapping, not a new empirical success criterion. Four counts distinguish newly corrected errors, newly introduced errors, persistent errors, and unaffected correct predictions. They are computed from each individual thresholded output, never from average logits.

Partial-trained models illustrate why net improvement need not be safe for every subset. Both masked and zero-neutral families newly harm 20% of allowed states on average even as they correct more. Three fits in each family worsen on the allowed stratum. In matched split-only training, every

Table 6: Joint-view rates (%). Accuracy uses 4,096 contexts per fit, FA 3,796, FD 300. After-patch joint predictions agree with the original split decisions.

Reviewer	Training	Path	Accuracy	FA	FD
Masked max	Partial	Before	82.89	12.16	79.69
Masked max	Partial	After	96.31	0.00	50.44
Zero-neutral	Partial	Before	79.47	16.64	69.76
Zero-neutral	Partial	After	97.41	0.00	35.33
Masked max	Mixed	Before	100.00	0.00	0.02
Masked max	Mixed	After	99.99	0.00	0.11
Zero-neutral	Mixed	Before	100.00	0.00	0.00
Zero-neutral	Mixed	After	100.00	0.00	0.07
Masked max	Split	Before	80.62	15.27	71.31
Masked max	Split	After	99.98	0.00	0.24
Zero-neutral	Split	Before	79.56	16.63	68.64
Zero-neutral	Split	After	100.00	0.00	0.07
Masked max	Joint	Before	99.99	0.00	0.11
Masked max	Joint	After	96.86	0.00	42.91
Zero-neutral	Joint	Before	100.00	0.00	0.02
Zero-neutral	Joint	After	98.99	0.00	13.82

Table 7: All-fit bounds for the matched mixture conditions (%). Means do not describe a single deployment model.

Model	Train	J mean	J min	J max	S mean	S min	S max
Masked max	split	80.62	21.95	97.92	99.98	99.93	100.00
Zero-neutral	split	79.56	21.68	95.58	100.00	99.98	100.00
Masked max	joint	99.99	99.98	100.00	96.86	92.68	100.00
Zero-neutral	joint	100.00	99.98	100.00	98.99	94.56	100.00

Table 8: Four-way paired correctness (%). Each row averages 15 fits and sums to 100% up to rounding. Allowed/denied denominators per fit are 300/3,796. BC: both correct; JC: only joint wrong (corrected); SC: only split wrong (harmed); BW: both wrong (persistent).

Stratum	Support	Encoder	BC	JC	SC	BW
allowed	Partial	Masked	0.311	49.244	20.000	30.444
allowed	Partial	Neutral	10.244	54.422	20.000	15.333
allowed	Split	Masked	28.644	71.111	0.044	0.200
allowed	Split	Neutral	31.333	68.600	0.022	0.044
allowed	Joint	Masked	57.089	0.000	42.800	0.111
allowed	Joint	Neutral	86.178	0.000	13.800	0.022
denied	Partial	Masked	87.836	12.164	0.000	0.000
denied	Partial	Neutral	83.365	16.635	0.000	0.000
denied	Split	Masked	84.728	15.272	0.000	0.000
denied	Split	Neutral	83.372	16.628	0.000	0.000
denied	Joint	Masked	100.000	0.000	0.000	0.000
denied	Joint	Neutral	100.000	0.000	0.000	0.000

fit improves overall, yet one fit per architecture slightly worsens on allowed states. This descriptive distinction is lost by reporting only the aggregate improved-fit count. Complete 900 fit-by-stratum rows include all 300 reviewers; the 120 explicit join paths separately verify decision equality to the chosen reference.

Within this binary domain, explicit atom-splitting of the input provides the same pooled reference for occupied sources under masked or zero-neutral aggregation. This is an equivalent normalization control, not a competing method that must be outperformed. A larger domain would require an

explicit element/union specification and new checks. Policies sensitive to carrier roles invalidate the label-preserving premise before any such normalization can be recommended.

E READOUT PROBES AND OTHER REPRESENTATION CONTROLS

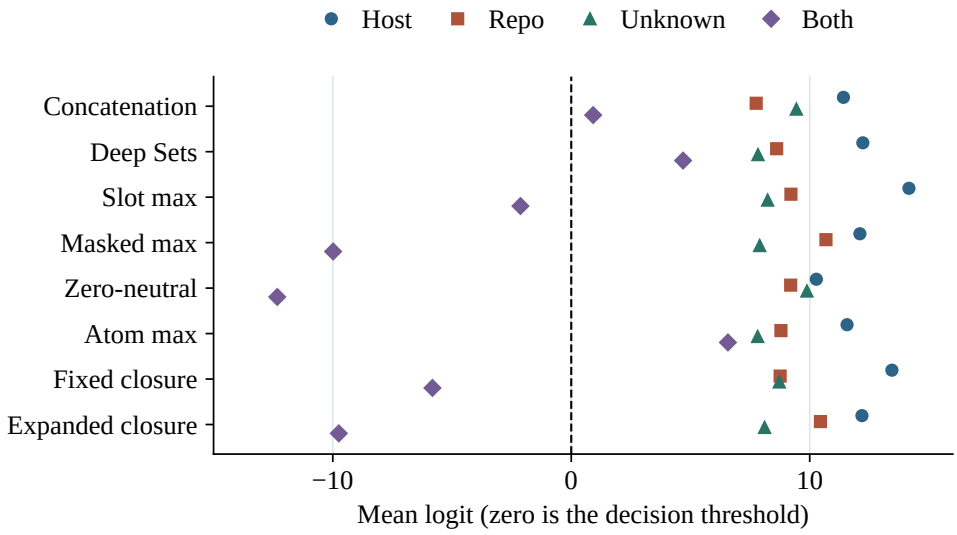


Figure 9: Actual mean logits on the same 300 allowed context-only inputs for four source states. Markers show distinct categories, not a continuous source axis. All fifteen partial-pair fits contribute; zero is the decision boundary. Errors are computed from individual predictions, not the mean logit.

The base generator does not place repository and unknown ancestry together. Consequently the correct policy and one that additionally vetoes every mixed union agree on the base training support. A readout can extend incorrectly even when its representation retains sufficient facts and ignores all policy-irrelevant layouts. Fixed closure and expanded closure are exactly invariant yet make substantial mixed-label errors; atomic pooling achieves a different, but not perfect, extension under the same data support.

Expanded closure uses a learned 2–16–16 map on its closed flags and the same readout width as the learned carrier/atomic models. Its persistent failure excludes a simple claim that nonlinear width alone resolves the issue. It does not prove that atomic encoding is universally optimal under other optimizers, policies, or source domains. Matched support later repairs many of these errors without changing structural invariance.

Raw pooled-coordinate norms are not comparable semantic severities across models. The fresh output files include both within-model representation residuals and observed score/decision changes. A rescaling can change the former while preserving the latter. Interpretation is therefore restricted to same-model identities and interventions, together with separately measured policy errors.

Wrong invariance. The retained discovery negative control changes the policy to distinguish repository influence in the value and control carriers. Balanced paired states can then have the same source union and opposite correct labels. The collision bound forces at least one error per pair for a union-only encoding. This is a scope extension, not a contradiction of Equation 1. The diagnostic must reject the proposed invariant relation before recommending a union repair.

Faulty metadata. If an input omits repository origin while retaining implicit host, even exact evaluation of the corrupted facts can permit a request that the true source policy would deny. Reporting an unknown origin instead invokes the fail-closed authority branch, subject to the unchanged con-

	Joint FA	Joint FD	Split FA	Split FD	Decision flip
Concatenation	0.0	42.5	0.0	32.0	3.8
Deep Sets	4.3	11.6	0.0	1.7	4.9
Slot max	17.0	54.4	0.0	65.5	19.9
Masked max	12.2	79.7	0.0	50.4	16.3
Zero-neutral	16.6	69.8	0.0	35.3	20.9
Atom max	0.0	2.8	0.0	2.8	0.0
Fixed closure	0.0	86.4	0.0	86.4	0.0
Expanded closure	12.3	78.8	12.3	78.8	0.0

Joint/split FA: 3,796 denied; FD: 300 allowed; all rates in percent.

Figure 10: Fresh mixed joint/split error decomposition for all eight partial-pair model families. FA and FD condition on 3,796 denied and 300 allowed contexts respectively; flips use all 4,096 paired contexts. Each cell is a mean over fifteen fitted reviewers.

text exception. These are extraction/authentication distinctions, not conclusions obtained from a classification error alone. No pooling identity recovers an unobserved source.

A useful diagnostic output is therefore conditional: a verified mapping mismatch contributes to a changed score, the remaining invariant path is wrong under the supplied label, or the input relation is not justified by the policy. More than one can be true. When labels, origin construction, or policy invariance are uncertain, the procedure reports that limitation instead of assigning a unique mechanism.

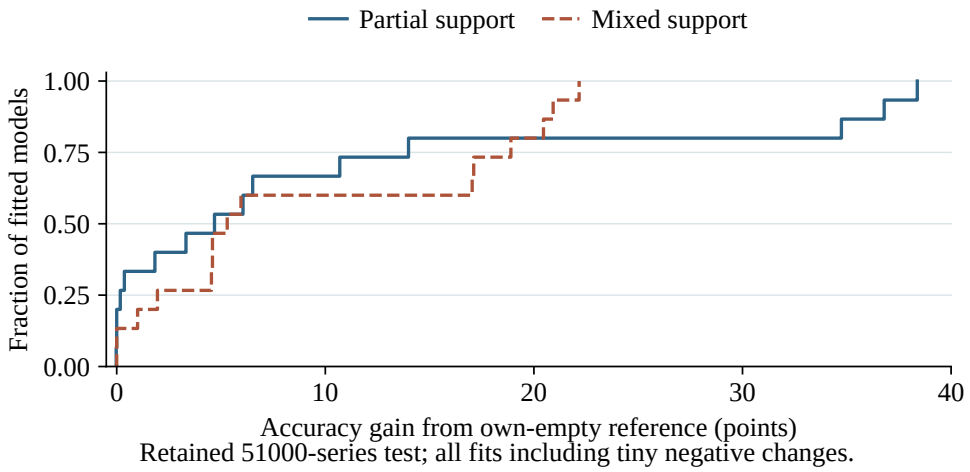
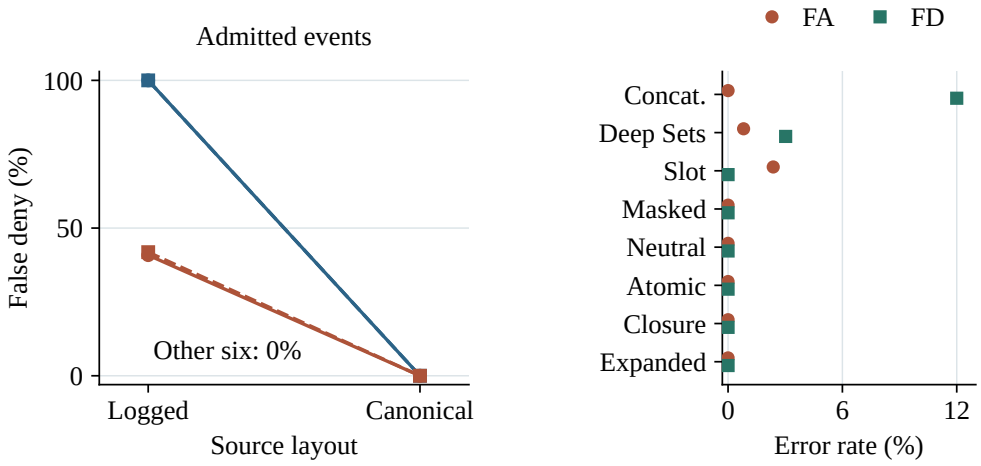


Figure 11: Retained own-empty-reference gains on the 51000-series test. Each empirical distribution contains all fifteen fitted gains. Tiny negative changes are retained; no integer rounding is used to label them. The test population is separate from current 61000-series estimates.

The empty repair reintroduces an element already seen on every singleton. It neither changes the latent compound map nor enforces $\phi(11) = \max(\phi(10), \phi(01))$. Its larger gains in a subset of fits are not a universal ten-point effect. Reusing an exact fitted checkpoint is important: training a second architecture could also change the optimization path and readout, whereas this particular comparison holds them fixed.

The complete numeric checkpoint arrays, pre/post probabilities, and intervention semantics remain archived alongside their original record identifiers. Retained numerical checks are separated from new training and from local request execution. This keeps an explanation of a representation identity from being misreported as a new end-to-end safety outcome.

F RETAINED NATIVE INPUTS, COUNTERFACTUALS, AND FIT CONCENTRATION



Left: blue concat., rust Deep Sets; solid actual, dashed matched descriptors.

Figure 12: Paired factor isolation and separate counterfactual errors. Left: logged/canonical layouts under actual (solid) or training-marginal (dashed) descriptors; blue is concatenation and rust is Deep Sets. Right: FA/FD on 28,800 offline source/grant variants, not new executions. The admitted cohort cannot estimate FA.

The retained maintenance cohort has 240 admitted events: thirty tasks in packaging, Jinja2, and attrs, with four process and four persistence events per task. The upstream profile is mini-SWE-agent 2.4.6 with prescribed actions and host review plans. Source files remain identified by their archive hashes. This population is supporting interface grounding, not a new autonomous rollout.

The 240 events have 118 distinct typed reviewer inputs but one policy/source pattern: gated, exact-granted repository ancestry in three carriers. Logged and canonicalized source layouts are crossed with actual descriptors or eight fixed training-marginal draws per event in the retained executable study. Policy labels and other common facts remain unchanged. Matched draws are repeated consistently in the actual-descriptor arrays to preserve pairing; the resulting 1,920 rows per condition are still 240 events, not new executions. Removing the six descriptors and retraining retains the concatenation layout contrast.

Separately, source identity, grant state, and nonempty source masks produce 120 counterfactuals per event, or 28,800 views (18,000 allowed; 10,800 denied). These explicitly altered facts receive independently recomputed targets. They are not naturally discovered attacks or newly executed effects. Success on the positive-only cohort cannot estimate FA; the counterfactual columns provide a different, restricted diagnostic.

The retained final controlled native collection contains 256 attempts, with 112 allowed and 144 denied. Two of fifteen Deep Sets partial-pair fits make false allows. Draw 0/seed 47 errs on 16 stale-grant persistence attempts; draw 1/seed 17 errs on 32 stale/spent persistence attempts. Every other fit has zero FA. The average 2.22% and maximum 22.22% therefore describe a concentrated fitted failure, not an identical risk for each reviewer. Development attempts remain separately archived and excluded from this final collection.

G MECHANISM-PAIRED ACTUAL REQUESTS AND TRUE CONTEXT HANDLING

The 192 new attempts use the existing local `FixtureAdapter`, not a modified label on an authority sink. Its registry has context `ctx.read`, gated `process.exec`/`state.write`, and separately available blocked selectors. The tested context entry returns the canonical bounded input value after host/active checks. The process entry executes a host-owned fixed Python echo program with its argument treated as data; state writes use the private store. No arbitrary repository program is executed.

The design crosses two mechanisms, four values (`ok-0` through `ok-3`), four grant states, and three entries. Each tuple supplies a two-view pair. Repetition places repository ancestry in one or four carriers. Partition places repository and unknown jointly in value or separately in value/control. An explicit trusted-test constructor supplies these origins. Both views retain the same common vector, complete canonical operands, operand digest, union, target, and effect semantics. This tests downstream interpretation, not production provenance extraction.

A private grant store is initialized per attempt and reset is verified. Grants authenticate the declared binding with an ephemeral host key; expired intervals are in the past, and spent grants are consumed before the attempted use. The current host validation time, equality facts, status, and body are recorded; keys are not exported. The frozen scorer runs before the monitor's execute call. Scores never enter the monitor. Process completion, read-only response, state contents, and observation-window closure are checked independently of the declared policy label.

Table 9: Actual local mechanism requests. Each mechanism/entry cell contains four values, four grant states, and two views.

Mechanism	Entry	Attempts	Permit	Deny	Completed
Repetition	Context read	32	32	0	32
Repetition	Process	32	8	24	8
Repetition	State write	32	8	24	8
Partition	Context read	32	32	0	32
Partition	Process	32	0	32	0
Partition	State write	32	0	32	0
Total		192	80	112	80

The 80 completions are not all authority-changing effects: 64 are context responses, eight processes, and eight state writes. Every rejected authority attempt has no effect. All 96 pairs preserve policy and outcome. The cohort has 48 unique full vectors and 32 policy/source patterns. Replicated values and repeated model scores are not independent tasks or new mechanisms.

The scorer set contains 120 retained full-input partial-pair models and 60 newly trained models, plus local join paths for the 90 supported masked/zero-neutral models. These 270 scoring configurations produce 51,840 pre-dispatch scores. New model training was completed before collection. No online fitting, threshold adaptation, language-model call, or human approval timing occurs.

G.1 RESIDUAL CONTEXT ERRORS BY RECORDED GRANT FACTS

For split-trained models, all patched joint errors in this context subset occur when no grant exists. In that case the real serializer sets presence/authentication/freshness/unused/equality fields to zero. The synthetic generator instead samples predicates independently and does not normally supply that full combination. The exact context policy ignores those grant details after host/active checks, but these learned reviewers need not. This is additional common-fact support shift, not residual source-partition sensitivity.

In the joint-trained models the patch can introduce errors on grant-present cases as well, again following their unchanged split predictions. Per-request and grant-stratified records retain this distinction. These findings condition on trusted fixture metadata and do not certify broad safety.

Table 10: Actual joint requests: mean advisory FA/FD (%). Every authority request in this mixed-union subset is denied; every true context request is allowed.

Model	Support	Path	Authority FA	Context FD
Masked max	partial	Before	15.00	83.33
Masked max	partial	After	0.00	58.33
Zero-neutral	partial	Before	15.00	70.00
Zero-neutral	partial	After	0.00	43.33
Masked max	split	Before	15.00	66.67
Masked max	split	After	0.00	25.00
Zero-neutral	split	Before	15.00	68.33
Zero-neutral	split	After	0.00	25.00
Masked max	joint	Before	0.00	25.00
Masked max	joint	After	0.00	53.33
Zero-neutral	joint	Before	0.00	25.00
Zero-neutral	joint	After	0.00	31.67

Table 11: Patched joint context FD (%) by grant facts. Each column has four actual requests per fit. The absent-grant subset is a separate common-fact support shift.

Model	Training	Absent	Exact	Expired	Spent
Masked max	partial	100.00	33.33	46.67	53.33
Zero-neutral	partial	100.00	20.00	26.67	26.67
Masked max	split	100.00	0.00	0.00	0.00
Zero-neutral	split	100.00	0.00	0.00	0.00
Masked max	joint	100.00	33.33	40.00	40.00
Zero-neutral	joint	100.00	0.00	13.33	13.33

G.2 Figure materials and source integrity

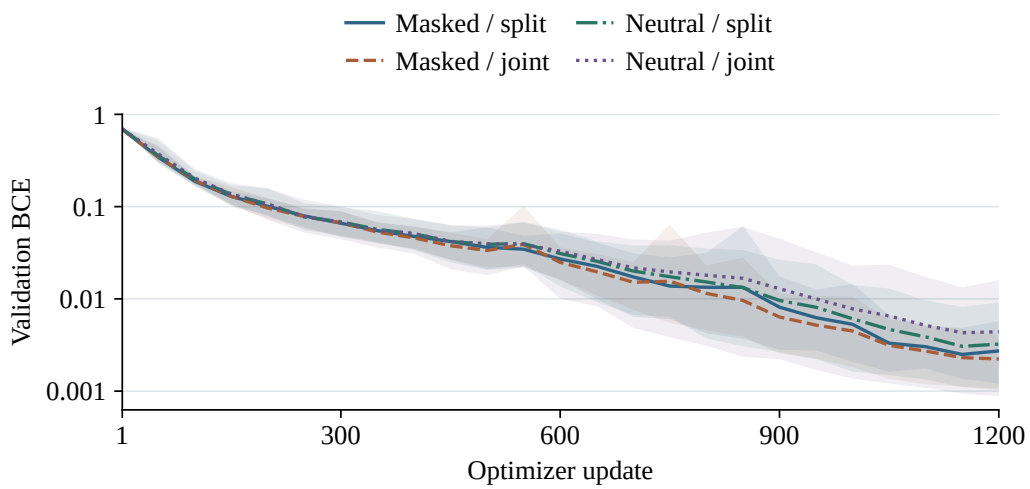
Figure 1 embeds actual headless-browser captures of a read-only HTML viewer populated from the two retained source-pair records. It is not a screenshot of an upstream native UI. Command, environment, timeout, and outcome are directly sourced; the complete identical operands and source JSON are supplied. The lower file/branch/queue/store objects are explicit schematics of alternative transfers, not records of one mandatory sequence.

Figure 2 uses the fixed checkpoint `masked_max__split_only__d0__s7` and the fixed `partition-v0-exact-ctx.read` pair. All sixteen coordinates are shown on one common scale up to 1.67598. Joint score is 0.222223, original split 0.999685, and patched joint 0.999685; both actual context calls return `ok-0`. The patched embedding is a deterministic calculation from saved parameters, not a retrained activation. The artifact contains full arrays, request IDs, weight hashes, HTML, browser captures, and editable SVGs. No external decorative photos or invented experiment screenshots are used.

H Matched validation trajectories and strong retained controls

Training-support matching was checked before fitting. Both split-only and joint-only runs sample the same contexts and target labels at every update; they differ in how a selected mixed origin is packed into one or two carriers. Equivalent labels do not imply identical gradients, element supports, or learned latent geometry. Both encoders have the declared finite parameterizations. The comparison is not equal to adding a loss that enforces the join identity.

No curve is inferred from terminal accuracy. Each recorded step contains actual batch and validation losses, and the chosen checkpoint is tied to its recorded validation minimum. The final artifact stores numeric weights, all selected-step records, and new test probabilities separately from retained

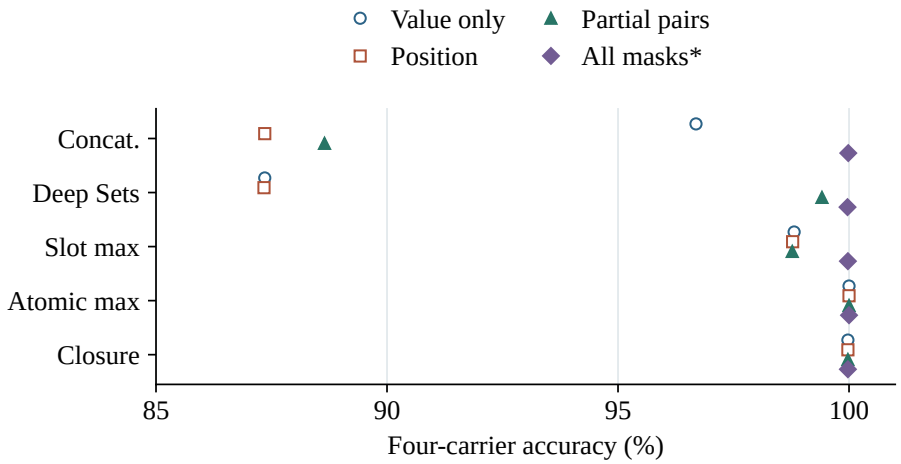


Band: full 15-fit range, not a confidence interval.

Figure 13: Validation BCE during the 60 new fits. Lines average all fifteen data/initialization fits in each condition; bands span the full fitted range. Checkpoints minimize the corresponding fixed-support validation BCE. Test interventions never select a step.

evidence. A low training-support validation loss therefore remains compatible with failure on a deliberately held-out carrier element or partition.

The new runs use CPU PyTorch 2.10.0, NumPy 2.3.5, and Python 3.13.5, with deterministic framework operations and one computation thread per worker. Recorded wall times support provenance, not a comparison of reviewer inference speed. The retained runtime figures below belong to a separate exact-enforcement experiment.



Discovery test; *all-mask training exposes all fifteen source masks.

Figure 14: Discovery support comparison, with all four training regimes displayed as distinct markers. Each value averages three data draws at initialization 7. All-mask support exposes all fifteen masks; it is not an unseen-mask test. No point is replaced by a current confirmation result.

The older study retains complete learned Deep Sets, slot max, atomic max, concatenation, and fixed closure. Adequate support lets standard learners match the fixed reference: Deep Sets reaches 99.97% and concatenation 99.98% with all-mask support, versus closure at 99.98%. Atomic pooling can match or exceed closure. These results constrain a claim of unique architectural superiority rather than undermining a mechanism diagnostic.

Saturated/clipped sum equals Boolean closure for binary flags. The retained equivalence check compares the same fitted readout on the same resulting input; it is an identity control, not a competitor to be beaten. Validation-selected decision trees and logistic classifiers are also retained. A tree on closed inputs reaches 100% on the original ordinary, relocated, and composed conditions. The original same-weight raw-sum control is not renamed Deep Sets: a full set model learns its element map and readout.

No old discovery score is substituted for a fresh confirmation number. The original five-initialization MLP study, the one-initialization set-model extension, the later crossed executable branch, and the sixty new split/joint fits are separately indexed. Their sample counts and test IDs do not form one pooled independent population.

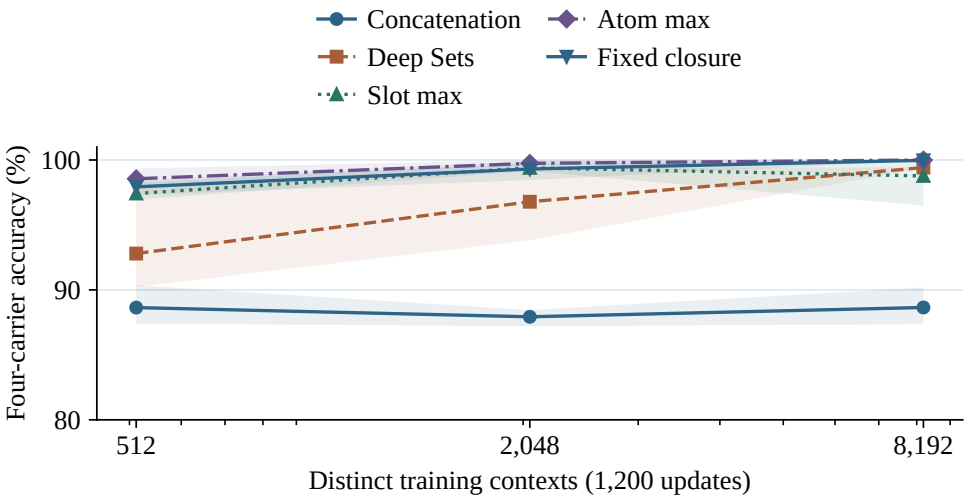


Figure 15: Retained discovery budget study under partial support. All budgets have 1,200 updates; bands span three data-realization fits at one initialization and are not confidence intervals. Context counts measure unique base records rather than augmented views.

The fixed-update budget comparison revisits smaller datasets more often. It is an investigation of support and available distinct contexts, not an equal-epoch or equal-task-diversity guarantee. General learners can close some representation gaps through exposure without acquiring a structural invariance. Conversely, exact invariance does not fix labels outside training support. The current matched split/joint conditions focus on that distinction while retaining these prior controls rather than recreating a weak-baseline advantage.

I Retained exact-enforcement timing and its scope

The retained mini-SWE-agent 2.4.6 profile mediates process and trajectory persistence. A prescribed-action trace provider supplies model actions, and project YAML is explicitly passed by the harness. Both task arms finish 90/90 runs; the protected arm has no false block among 720 intended effects under its fixed trusted plan. The ratio of summed protected to baseline time minus one is 1.12%, with retained task-cluster 95% interval $[-0.33, 2.68]\%$. It includes zero and does not establish a strictly positive overhead.

The kernel's correctness premises include complete mediation of declared authority paths, sound trusted source transfer, active bindings, protected keys and grant state, and dispatch of the same operands that were checked. Under those premises a repository-selected gated authority effect needs a valid exact grant; unknown authority ancestry is rejected. This is a conditional reference-monitor argument, not a premise supplied by classifier accuracy. Local shared-state transactions and consume-before-process semantics remain separately described in the underlying implementation.

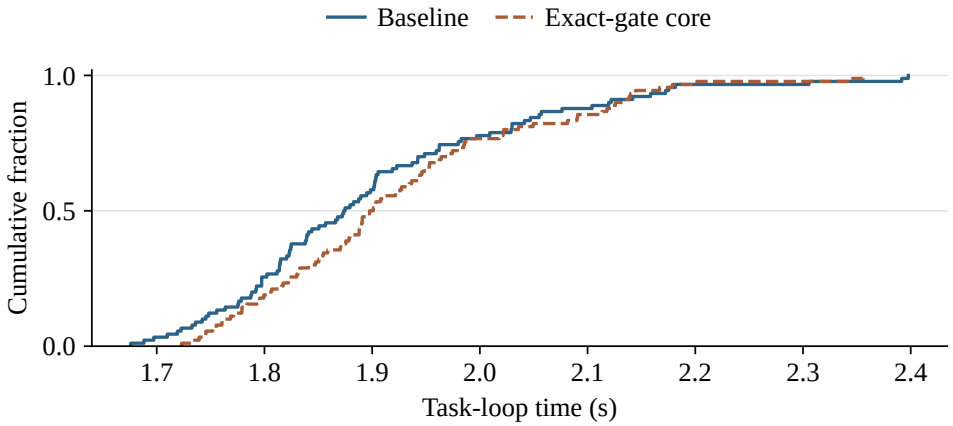


Figure 16: Retained task-loop empirical distributions: ninety matched pairs from thirty prescribed maintenance traces and three repetitions. Both arms share the instrumented isolation boundary. These are not newly run tasks and contain no learned-reviewer inference or human review inside the timer.

Both full-binding approval and the origin-aware kernel block all retained unauthorized opportunities. Host/no-new-grant behavior differs because of the permission policy, not because full binding is unsafe or measured human effort is lower. The new mechanism-pair experiment likewise records learned advice and actual authorization separately.

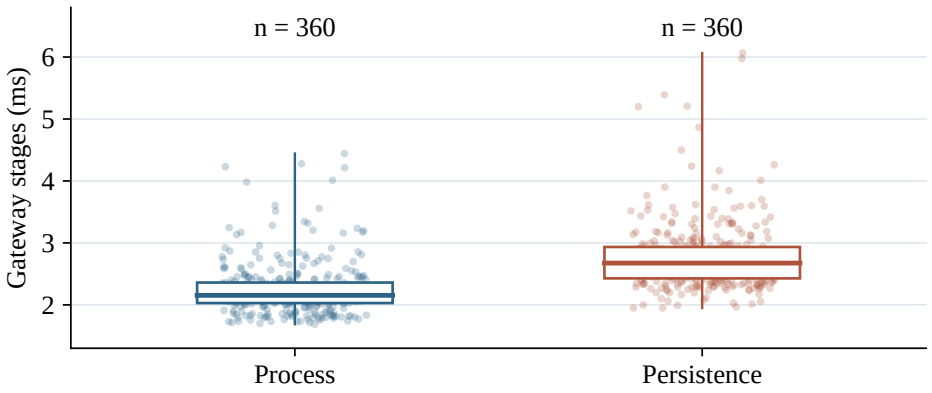


Figure 17: Retained gateway distributions, 360 observations per sink. Dots retain every timing; boxes show quartiles, lines medians, and whiskers the full range. Horizontal displacement only separates points. These stages exclude effect execution, learned inference, and human review.

These distributions are supporting compatibility/cost records. They do not measure the cost of the new join intervention or autonomous coding-agent performance. Retained endpoint tests, isolated configuration cases, and local grant-consumption checks remain separately indexed rather than enlarging the learning test denominator. A failed effect after valid consumption does not restore permission automatically or establish exactly-once execution.

Table 12: Retained complete timing summaries. Quantiles use the original observations and linear interpolation. Different rows use different units.

Measurement	n	Median	p95	p99
Baseline task (s)	90	1.874	2.174	2.392
Exact-gate task (s)	90	1.899	2.157	2.345
Process gateway (ms)	360	2.154	3.160	4.095
Persistence gateway (ms)	360	2.674	3.602	5.203

J PINNED EVIDENCE RELEASE AND TRAINING-SOURCE VERIFICATION

This submission uses evidence release `ORIGATE-Final-R1`. Its current main-analysis population is the 61000–61031 test with 4,096 contexts. The executable masked model uses its learned $\phi(00)$ when every flag slot is empty. No current result assumes the zero fallback described in an earlier manuscript draft. The result manifest binds every figure/table to its generating script, source arrays, checkpoints, predictions, and summary; package manifests identify the exact delivered files.

Historical correction, not an extra experiment. An earlier manuscript snapshot described 285 fits, groups 41000–41031, a zero all-empty masked fallback, and 384 local requests. The archived executable release instead contains 405 fitted models, groups 51000–51031, a learned fallback, and 256 final native requests plus an independent source pair. These differences are substantive. The unsupported 285-fit/384-request snapshot is not reconstructed or used as quantitative evidence. Original files and the correction record are retained, while the present paper consistently uses the executable semantics and separately identified populations. A different empty fallback could change training; its irrelevance to the nonempty join identity does not make trained outcomes interchangeable.

Table 13: Executable evidence inventory. Fit counts, probability rows, input contexts, and actual attempts are distinct units.

Evidence	Scope
405 retained fitted models	240 full-input primary, 120 no-descriptor, 45 exploratory absent-facts-support fits
60 matched-support fits	Two models, two matched supports, three draws, five seeds
300 fresh-test reviewers	240 full-input retained plus all 60 new fits
120 join paths	60 retained and 60 new; no additional training
6,881,280 reconstructed probabilities	Current ordinary/quad/joint/split originals and interventions
30,720 finite assignments	256 layouts for each of 120 join configurations
192 mechanism attempts	96 matched mechanism pairs; 80 completions, 112 denials
270 shadow scoring configurations	180 frozen original models plus 90 join paths; 51,840 scores
240 retained admitted native events	Positive-only interface grounding; counterfactuals kept separate

Three levels of verification. First, 465 retained fit histories have exactly the declared checkpoint-selection grid; every manifest's selected step equals its recorded validation minimum and every checkpoint hash matches. Eighteen copies of six distinct training/validation realizations regenerate exactly, including inputs, targets, and record identifiers. Second, the selected weights of all 300 main-input reviewers reproduce their fixed-support validation BCE with zero difference in this environment. These checks link recorded inputs, selection, and inference; they do not by themselves replay the optimizer.

Third, six configurations were fixed before replication: masked and zero-neutral, each under partial, split-only, and joint-only support, data draw 0 and initialization 7. Clean 1,200-update runs reproduce every parameter array, selected step, and recorded batch/validation loss exactly. Their inputs,

RNG streams, curves, and checkpoints are retained in a separate audit directory. They are verification reruns, not six additional study fits. We do not claim that all 465 optimization paths were independently replayed. The audit distinguishes this limited training replication from the broader independent NumPy inference and metric checks.

J.1 Clean replay outcomes and what they establish

All six replications start from their declared initialization and execute 1,200 updates, using the original fixed validation criterion. Cases were selected before the audit. The table compares the selected checkpoint and recorded losses; paired correctness aggregation never selects or modifies a checkpoint.

Table 14: Predeclared clean training replays, data draw 0 and initialization 7. The last two columns are maximum absolute differences across all checkpoint entries and all recorded batch/validation losses, respectively. Every selected step agrees with the original.

Encoder	Support	Step	Validation BCE	Weights	Losses
Masked	Partial	1200	0.0017937	0	0
Masked	Split	1200	0.0016729	0	0
Masked	Joint	1200	0.0019625	0	0
Zero-neutral	Partial	1200	0.0027555	0	0
Zero-neutral	Split	1200	0.0029766	0	0
Zero-neutral	Joint	1200	0.0086789	0	0

Equality in this environment does not imply bitwise agreement on arbitrary hardware or numerical backends. Stream digests bind sampled inputs and targets in update order; array checks independently bind generated records and identifiers. This is stronger evidence of training provenance than matching saved-model predictions alone. Only these six optimizer paths were freshly replayed; the other histories received record, input, and checkpoint checks.

J.2 Reproduction and visual-source boundaries

From the model root, the final evidence/figure route is:

```
python tools/analyze_final.py
python tools/plot_final.py
python tools/make_matlab_final.py
python tools/verify_final.py
```

The retained `verify_v4.py` independently reconstructs the current 6,881,280 main-study probabilities and 51,840 actual-request advisory scores. Earlier native populations are not substituted into these denominators. Fresh optimization is explicit and writes a separate directory; original weights and execution records are read-only during verification. The Final paper, plot specification, numerical tables, and model package share one release identifier rather than relying on a filename alone.

The standalone MATLAB function renders experimental Figures 3–17 from the same embedded specification as the executed Python renderer. Native MATLAB was not available: data serialization and static checks do not establish MATLAB execution. Figures 1–2 retain editable SVGs, record-viewer HTML/captures, fixed checkpoint identities, and all sixteen displayed embedding coordinates. Their provenance differs from a production-agent screenshot and is labeled accordingly. The common input path in Figure 2 enters each separately fitted readout after pooling.

The supplied ICLR 2027 style remains unmodified; retrieving the official ZIP for byte comparison was unsuccessful in this environment. The main text occupies nine pages without altering typography or margins. Ordinary vector figure text is checked at its final 5.5-inch width; screenshot CSS scaling is recorded separately and is not inferred from the PDF font table. No fonts, private keys, or production credentials are distributed. Final author verification of scientific claims and submission details remains necessary.