IDENTITY IS NOT THE DECISION: CERTIFYING REPAIRS FOR FROZEN COMMUNICATION INTERFACES

Anonymous authors
Paper under double-blind review

ABSTRACT

Frozen communication interfaces can change task meaning without changing message width. Public probes may leave multiple consistent interfaces that permit the same repair, making full identity recovery spend queries on action-irrelevant distinctions. We show how to certify a repair from the shared acceptable-action set across probe-consistent interfaces. We characterize exact nonadaptive query cost by the smallest probe set separating every minimal interface group lacking a common repair, and give a noisy stopping rule with a finite-sample bound on unsafe certification. Across 225 learned-agent pairs, anchor-consistent messages induce a common receiver action on 95.7 percent of held-out states; uniform selection among them has a 2.2 percent empirical upper bound on mismatch with the frozen receiver. Intervals resample ordered pairs, 10,000 replicates: unanimity 95.7 percent $[94.7, 96.7]$ and the conditional reference-action error 2.2 percent $[1.7, 2.7]$. On calibrated noisy channels, stopping uses fewer probes than target-symbol or full-interface identification and abstains when probes are uninformative. A complete automata catalog shows query savings against standard and advice-enhanced learning. On a message-dependent task, repair restores reference-level return, and released-protocol audits evaluate held-out workflows.

1 INTRODUCTION

A frozen sender-receiver pair can disagree about a message's meaning even when its width, schema, and version label match. These checks miss the mismatch, which appears at deployment when neither endpoint can be retrained. Whether a message reproduces the intended action also depends on receiver context: the same message can match one context and induce a different action in another.

The receiver maps multiple messages to the same action, so matching an observed action does not reveal which message convention produced it. We call this over-determination. The gap between observed behavior and a unique interface identity is the identifiability gap.

The practical question is which repair the evidence supports. Let $\Theta(D)$ be the interface states consistent with a public transcript D, and let $\mathcal{S}_\epsilon(\theta)$ contain repair actions within tolerance ϵ for state θ. The repair certificate is

$$\mathcal{S}_\epsilon(D) = \bigcap_{\theta \in \Theta(D)} \mathcal{S}_\epsilon(\theta).$$

Any action in this intersection is acceptable for every interface consistent with the evidence, even when several interfaces remain possible. If the intersection is empty, the audit gathers more evidence or abstains. Thus a repair can be certified before the interface is identified.

Probe cost depends on the repair target. Recovering full identity must separate every behaviorally distinct pair; identifying a single-valued target only requires separating pairs that assign it different values. For a shared-action repair, exact nonadaptive cost is the smallest probe set that separates every inclusion-minimal group of interfaces with no common acceptable action. Pairwise conflicts determine this cost when pairwise-overlapping safe-action sets always share an action; otherwise, larger conflicts can dominate. The query burden follows the repair target, not message width alone.

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

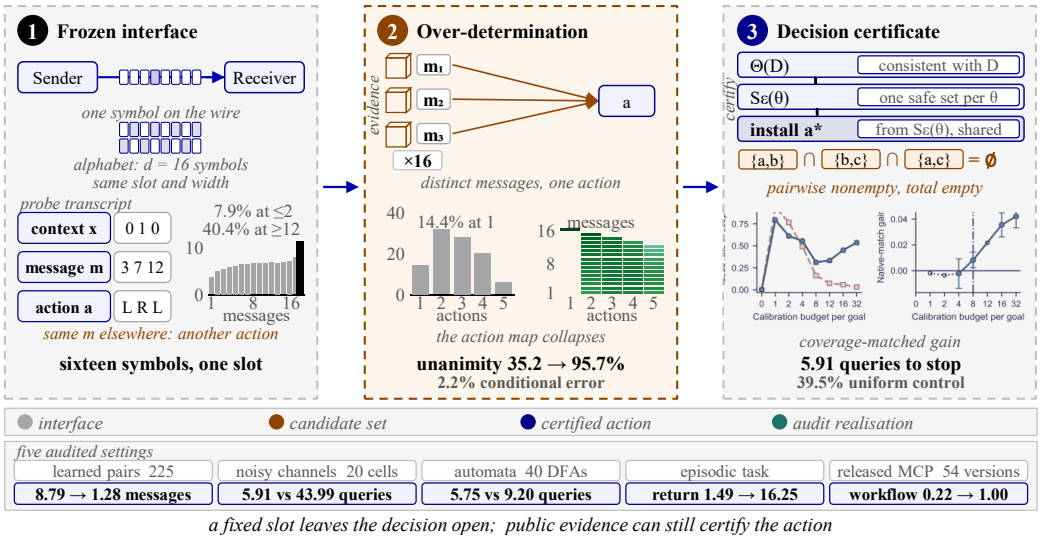

Figure 1: Fixed-width messages can hide over-determination: public probes narrow candidates to a shared repair. At thirty-two anchors, 95.7 percent of held-out states are action-unanimous; among them, conditional reference-action mismatch has a 2.2 percent empirical upper bound. Lower panels compare action agreement, probe costs, and repair outcomes.

Across 225 learned-agent pairs, thirty-two public anchors yield one receiver action on 95.7 percent of held-out states; choosing uniformly among consistent messages gives an empirical 2.2 percent upper bound on mismatch within this subset. Noisy channels and a complete automata catalog measure query costs for different repair targets. A message-dependent task and released MCP workflows test downstream repair outcomes.

Contributions. We give an exact query characterization for shared-action repair and a noisy-probe rule that certifies a shared action, continues when interfaces disagree on the repair, or abstains.

2 Related Work

Probing a frozen system. Public-query methods provide distinct query models for behavioral learning, conformance testing, and permutation identification. Active automata learning identifies a minimal-state behavioral quotient through membership and equivalence queries Angluin (1987); Kozen (1997). Finite-state conformance testing constructs test suites that compare an implementation with a specification Chow (1978); Lee & Yannakakis (1996); Dorofeeva et al. (2010). Ko and Teng study recovery of an unknown permutation Ko & Teng (1986); Sarkar and Dey study coding-theoretic message identification over noisy permutation channels Sarkar & Dey (2025; 2026). Advice-enhanced automata learning uses rewriting advice to reduce membership-query demands Fica & Otop (2025). These works make the identification objective explicit within their respective models. Our objective is the repair shared by all frozen interfaces consistent with public transcripts: an actionable decision can be certified before the hidden interface is fully identified.

Decisions under partial identification. Selective prediction and risk-controlling methods construct sets of labels or predictions with calibrated coverage and loss guarantees Chow (1970); El-Yaniv & Wiener (2010); Geifman & El-Yaniv (2017); Lei et al. (2018); Bates et al. (2021). Bayesian experimental design selects probes to optimize prior-weighted information or decision utility Chaloner & Verdinelli (1995), while formal-methods approaches to safe reinforcement learning combine verification with runtime monitoring under an explicit system model Fulton & Platzer (2018). Our audit computes the repair actions shared by every interface consistent with public transcripts. A nonempty intersection certifies an installable repair; decision-conflict probes guide further queries when candidate interfaces require incompatible actions, and the algorithm abstains when no such probe remains.

Frozen interfaces in deployed systems. Studies of deployed tool and model-context interfaces characterize lifecycle, compatibility, and aggregate reliability Hasan et al. (2026); Hou et al. (2026); Wang et al. (2025); Monce et al. (2025). Our released-protocol study evaluates behavior-class identification, class-conditioned adapters, and held-out workflow outcomes. The frozen sender-receiver analysis makes the repair target and query cost explicit, while released-protocol audits test behavior-class repair on held-out workflows.

Repairing a fixed model against a specification. Model-repair methods change a model to satisfy an explicit specification. They use convex correction at counterexamples, counterexample-guided retraining with final certification, SMT search over selected weights, reachability analysis before retraining, robust-optimisation criteria for termination, or behaviour imitation as the objective Cruz et al. (2022); Bauer-Marquart et al. (2021); Cohen & Strichman (2022); Yang et al. (2021); Boetius et al. (2023); Liang et al. (2024b). REGLO targets globally specified network repair, while ProRepair uses preimage synthesis and property refinement Fu et al. (2024); Ma et al. (2025); our method keeps the network fixed and certifies a communication action shared by evidence-consistent interfaces. Work on feedback through frozen language models studies learned messages Akyürek et al. (2023); output-level methods use critique-guided revision Ma et al. (2023) and failure-signal diagnosis Hashemifar et al. (2026). We measure the public queries needed to identify the target-compatible action (Section 3). Certification from observable events is a live line. A protocol that tests two agents on shared events and certifies the terms whose observed disagreement stays under a statistical threshold bounds the disagreement of decisions restricted to that certified core Schoenegger et al. (2026), and a repair channel opened when a group's plans are predicted to violate a cooperation constraint raises task success at a measured communication cost Yang et al. (2026). The object certified here is the repair action the evidence leaves, and its price is the size of the set that evidence fixes.

Communication learning and non-stationarity. Communication learning optimizes a resource during training and evaluates it by task return, while here the channel is fixed and the question is what public behavior certifies. Centralized training with decentralized execution, counterfactual credit assignment and value factorization form the standard toolkit for partially observed coordination Lowe et al. (2017); Foerster et al. (2017); Rashid et al. (2018); Yu et al. (2021), and a parallel literature learns what to send, whom to address, when to communicate and how wide a message should be, including under bandwidth limits and unreliable channels Foerster et al. (2016); Jiang & Lu (2018); Das et al. (2018); Kim et al. (2019); Sun et al. (2023); Weil et al. (2023); Yuan et al. (2023). The misalignment is relational. What matters is the pair of codebooks and the decoding table between them, so non-stationarity of the marginals leaves the problem untouched Emami et al. (2023); Xiong & Singh (2026); Zhang et al. (2023); Hayashi (2026). Public probes leave some codebook distinctions unresolved while identifying the repair actions that remain available to the operator (Section 5).

3 What public probes establish

A public probe leaves a set of interfaces consistent with the observed responses. A repair is certifiable when every interface in that set supports a common action.

What behavior can identify. Let Θ be the set of frozen protocol states and P_θ the distribution of every finite adaptive behavioral experiment. Two states are behaviorally equivalent, $\theta \sim_B \theta'$, when $P_\theta(\tau) = P_{\theta'}(\tau)$ for every policy and transcript τ. Two states in the same class cannot be separated by any public probe, estimator, or randomized decision rule. The quotient $\Theta/\!\sim_B$ is therefore the finest interface description that public behavior can identify (Appendices E.5 and F).

The repair shared by compatible interfaces. Let D denote the public probe transcript, $\Theta(D)$ the set of interface states whose observable behavior is consistent with D, and $\ell(\theta, a)$ the loss of repair action a. The context is part of the state, so every statement below is read at a fixed decision context x, and $\mathcal{S}_\epsilon(\theta)$ is written for $\mathcal{S}_\epsilon(\theta; x)$. For a tolerance $\epsilon \geq 0$, let $\mathcal{S}_\epsilon(\theta)$ collect the actions within ϵ of the best one under ℓ. The repair certificate is the intersection of these action sets over every interface the evidence leaves open, $\mathcal{S}_\epsilon(D) = \bigcap_{\theta \in \Theta(D)} \mathcal{S}_\epsilon(\theta)$. An action in $\mathcal{S}_\epsilon(D)$ is safe for every interface state that remains compatible with the evidence. A nonempty intersection certifies a shared repair; an empty one calls for more evidence or abstention.

Interface identity and repair are different targets. A version label, a repair action, and a loss induce different partitions of the same behavioral family. For an exact target $q : \Theta \to \mathcal{Z}$, identification requires

$$\theta \sim_B \theta' \quad \Longrightarrow \quad q(\theta) = q(\theta'),$$

which is equivalent to q factoring through the behavioral quotient. Shared-action repair asks whether every remaining interface supports a common acceptable action (Appendix E.7).

Exact probe cost for a shared repair. Let Θ be a finite interface family, $\mathcal{X}$ a finite probe set, and $\rho_\theta : \mathcal{X} \to \mathcal{Y}$ the deterministic response map of interface θ, with nonempty safe-action sets $\mathcal{S}_\epsilon(\theta)$. A query set Q identifies a single-valued target $q : \Theta \to \mathcal{Z}$ exactly when it intersects every target-disagreement set

$$D(\theta, \theta') = \{x \in \mathcal{X} : \rho_\theta(x) \neq \rho_{\theta'}(x)\} \quad \text{for each pair with } q(\theta) \neq q(\theta').$$

A shared-action repair can require separating groups of candidates, not just pairs. The transcript cell containing the true interface $\theta_\star$ is $C_Q(\theta_\star) = \{\theta \in \Theta : \rho_\theta|_Q = \rho_{\theta_\star}|_Q\}$. A query set certifies a shared action exactly when every such cell has a nonempty safe-action intersection. Let $\mathcal{M}_\epsilon$ collect the inclusion-minimal $A \subseteq \Theta$ satisfying $\bigcap_{\theta \in A} \mathcal{S}_\epsilon(\theta) = \emptyset$, and define their separating probe edges by

$$E(A) = \{x \in \mathcal{X} : \exists \theta, \theta' \in A,\ \rho_\theta(x) \neq \rho_{\theta'}(x)\}.$$

The exact nonadaptive cost is the smallest probe set that separates every minimal group with no common safe action, equivalently the hitting number of this conflict hypergraph:

$$Q^\star_{\mathrm{dec}} = \min_{Q \subseteq \mathcal{X}} \{|Q| : Q \cap E(A) \neq \emptyset \text{ for every } A \in \mathcal{M}_\epsilon\}.$$

The pairwise subfamily gives a lower bound $\tau_2 \leq Q^\star_{\mathrm{dec}}$, where τ_2 is the minimum hitting number over two-interface conflicts. This bound is exact when every pairwise-intersecting subfamily of safe-action sets has a common action (the 2-Helly case). The three safe sets $\{a, b\}$, $\{b, c\}$, and $\{a, c\}$ form a minimal three-way conflict with no pairwise conflict, exposing the higher-order structure captured by the certificate (Appendices E.9 and F).

Stopping with noisy probes. Let $\widehat{\Theta}_t$ be a confidence family after t observations satisfying

$$\Pr\left(\theta_\star \in \widehat{\Theta}_t \text{ for every } t \geq 0\right) \geq 1 - \delta,$$

and set $\widehat{\mathcal{S}}_t = \bigcap_{\theta \in \widehat{\Theta}_t} \mathcal{S}_\epsilon(\theta)$. Every predictable probe policy that returns $a_\tau \in \widehat{\mathcal{S}}_\tau$ at a stopping time τ when this set is nonempty, and abstains otherwise, satisfies

$$\Pr\left(\tau < \infty,\ a_\tau \notin \mathcal{S}_\epsilon(\theta_\star)\right) \leq \delta.$$

The row-wise likelihood-ratio construction in Appendix E instantiates the condition for the replacement channel. The conflict hitting number gives the exact cost for fixed probes; the confidence family controls stopping when probes are chosen adaptively and repeated.

A probe transcript determines the compatible interfaces; each contributes its acceptable repairs, and their intersection is the set of actions the operator can install:

$$\Theta(D) \ \longrightarrow\ \{\mathcal{S}_\epsilon(\theta) : \theta \in \Theta(D)\} \ \longrightarrow\ \mathcal{S}_\epsilon(D).$$

The operator selects the target q before applying the rule.

Algorithm 1. Certify a shared repair, gather a probe, or abstain.
input transcript D, active family $\mathcal{H}(D)$, safe-action sets, query policy π based on past observations, remaining budget
output certify(a^*), gather(x), or abstain
1 if $\mathcal{H}(D) = \emptyset$ **return** abstain; otherwise $A(D) \leftarrow \bigcap_{\theta \in \mathcal{H}(D)} \mathcal{S}_\epsilon(\theta)$
2 if $A(D) \neq \emptyset$ **return** certify(a^*) for a fixed $a^* \in A(D)$
3 if the budget remains and $\pi(D) \in \mathcal{X}_{\mathrm{conf}}(D)$ **return** gather$(\pi(D))$
4 return abstain

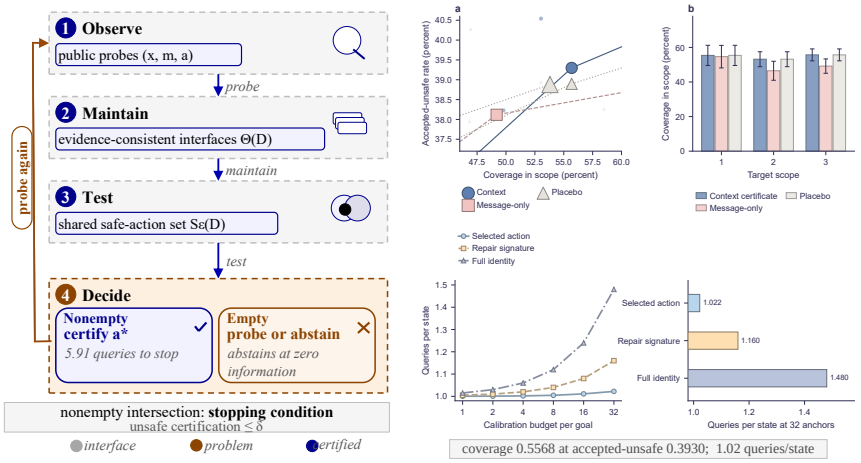

Figure 2: The audit maintains probe-consistent interfaces and tests their shared acceptable-action set (left). The right panels show coverage against accepted-unsafe rate, coverage by target scope, and query costs as calibration grows. At thirty-two anchors, target-specific repair uses 1.022 queries per state, compared with 1.160 for a repair signature and 1.480 for full identity.

Use $\mathcal{H}(D) = \Theta(D)$ for exact audits and $\widehat{\Theta}_t$ for noisy audits. Define $\mathcal{X}_{\mathrm{conf}}(D) = \bigcup_{A \in \mathcal{M}_\epsilon(\mathcal{H}(D))} E(A)$, restricting conflicts to the active family. Exact selection solves the finite minimum hitting-set problem; minimal-conflict enumeration is exponential in $|\mathcal{H}(D)|$ and minimum hitting set is NP-hard, so a lazy implementation selects a probe that hits the largest uncovered conflict family. For learned agents, each bucket $B(g, m) = \{a \in \mathcal{A} : a.\mathrm{msg} = m,\ a.\mathrm{goal} = g\}$ defines $C_D(g, m) = \{m' : \hat{a}(x_i, m') = a_i \text{ for every } (x_i, a_i) \in B(g, m)\}$; for held-out context x, these messages induce the actions $A_D(x) = \{f_R(x, z) : z \in C_D(g, m)\}$. They agree exactly when $|A_D(x)| = 1$. The selector installs the lowest-index message in $C_D(g, m)$; reference-action mismatch measures fidelity to the frozen receiver. Noisy-channel and automata results evaluate the interface-level intersection $\mathcal{S}_\epsilon(D)$.

Receiver context can make two messages that match on anchor states induce different actions on a held-out state. The audit therefore retains every message that matches the anchors and checks the actions they induce in each held-out context.

Public behavior and version labels. Public probes reveal interface behavior; version labels remain checkpoint metadata. Versions that produce the same behavior cannot be distinguished by any public transcript, a limit captured by the behavioral quotient studied in active automata learning Angluin (1987); Aichernig et al. (2022). The repair can still be certified when all remaining versions support the same action. The audit returns that repair or abstains.

When repair queries cost less than identity queries. The classical bijective case gives a tight endpoint. For a transform class whose public interface is a bijection on the alphabet, m queried sources leave exactly $(d - m)!$ consistent interfaces and $d - 1$ of them identify the interface Ko & Teng (1986). In the learned endpoint, each bucket fixes one sender message while the receiver's reply depends on context: its anchors carry 1.62 distinct chosen messages at four anchors and 3.88 at thirty-two. Query cost tracks how many distinct messages the evidence leaves in each bucket across four admissible transform families, widths through $d = 32$, and trained finite-task checkpoints. For a full permutation family, a target supported on $k < d$ source coordinates costs k queries, while full identity costs $d - 1$; parity can retain the full-interface cost. For the audited cyclic histories, the label costs one query at length $K = d$ and is unidentifiable at $K = d + 2$ (Appendix F).

4 Guarantees and Evidence for Repair Certification

Noisy channels evaluate the interface-family certificate; automata compare target-specific stopping with language-equivalence learning. The learned endpoint instantiates this finite family: $C_D(g, m)$ is the observable projection of $\Theta(D)$, $A_D(x)$ its action image, and unanimity the singleton case of

$\mathcal{S}_\epsilon(D)$; reference mismatch is deployment fidelity. The projection is exact on the class of receivers whose actions agree at a context whenever they agree on every anchor, which is the smoothness condition the local guarantee names; the endpoint audits report anchor unanimity, neighbour distance and held-out reference fidelity, the three quantities that check it. Figure 2 summarizes the certificate: certify a shared action, gather a decision-conflicting probe, or abstain.

Exact enumeration exhibits a three-way repair conflict: the safe-action sets $\{a, b\}$, $\{b, c\}$, and $\{a, c\}$ intersect pairwise, while their total intersection is empty. A pairwise test therefore misses a candidate family with no common repair.

Noisy-channel certification. With a time-uniform likelihood confidence set at $\delta = 0.05$, 200,000 trials across 20 width-reliability cells yield equal-cell mean query counts of 5.91 for shared-action certification, 7.59 for target-symbol identification, and 43.99 for full-identity recovery. Shared-action stopping uses no more queries than target-symbol identification in any cell. Minimum empirical pathwise coverage is 99.62 percent, and the maximum cellwise unsafe-certificate rate is 0.15 percent, with a maximum 95 percent Wilson upper bound of 0.25 percent (Appendix E.11).

The abstention boundary. A separate 180,000-trial signal sweep finds that all three policies abstain at the uninformative point $r = 1/d$. At normalized signal $\eta = 0.25$, where $r = 1/d + \eta(1 - 1/d)$, shared-action certification rates are 97.80, 99.23, and 99.95 percent for widths 3, 4, and 8. Paired query savings over target-symbol identification are 19.28 [18.92, 19.65], 19.17 [18.81, 19.52], and 12.06 [11.81, 12.32] queries, respectively (95% normal intervals; Appendix E.11).

A comparison with advice-enhanced learning. On the complete catalog of 40 commuting two-state DFAs, public $L^\star$ averages 9.20 teacher queries to learn a language-equivalent automaton, and the released advice-enhanced implementation averages 7.90. Target-specific stopping averages 5.75 queries without advice and 5.20 with advice; the advice-plus-stop arm certifies every language-acceptance case and 30 percent of each richer target family (Appendix G, Fica & Otop (2025)).

Proposition 1 (Unanimity guarantees message support on selected anchors)**.** *Let a state have a bucket B of calibration anchors and let $N \subseteq B$ be the k nearest. A gate that issues only when the anchors of N agree on the chosen message returns a message that every anchor of N also chose. A gate that emits its modal message from a strict subset $N' \subset N$ on a state where N is not unanimous returns a message that is not shared by all of N.*

Appendix E proves the support-agreement property directly: unanimity has zero violations, while the audited two-of-three and plurality gates incur violations on issued states whose full neighbourhoods disagree.

Proposition 2 (Same-anchor rules cannot separate indistinguishable interfaces)**.** *Let a rule read only the anchors of a set N, and let the candidate family contain two interfaces that record the same action at every anchor of N while their safe sets are disjoint. Then no rule reading only N returns an action in the intersection. The two candidates are indistinguishable to it, so it returns the same action for both, and that action lies outside the safe set of at least one.*

What a rule reads bounds what it can certify, and reading anchors that separate the two candidates is what removes the case. This is why the axis the evaluation sweeps is the budget (Appendix E).

Proposition 3 (How candidate messages determine reference-action mismatch)**.** *Fix an issued state whose candidate message set has size d, and let $r(x)$ count the candidates whose action is the reference action there. A gate that installs one candidate mismatches the reference exactly when that candidate lies outside those $r(x)$, so over a pool the reference-action mismatch rate is $1-\mathbb{E}[r]/d-\Delta$, with $1 - \mathbb{E}[r]/d$ the error a uniform draw would make and Δ the advantage the gate's own rule carries over it.*

The state fixes the first term and the rule sets the second, and the evaluation measures both (Section 5).

Proposition 4 (Local safety transfers under four conditions)**.** *Fix a state with receiver context x and sender message m. Let $a_1, \ldots, a_k$ be the target actions recorded at the k nearest anchors carrying m. Suppose the safe sets agree within a radius δ, the neighbours lie inside it, the recorded actions*

Rule	Calibration reads	Online per output	Output rate	Ref. mismatch	Abstention
local gate, 32 anchors	$\|B\|$ anchors	1	0.5381	0.3106	on disagreement
lowest in set, 32	$\|B\|\, d$ actions	1	1.0000	0.0185	on an empty bucket
Calibration reads are paid once per anchor bucket; each emitted message costs one online receiver action.					
lowest in set, 160	$\|B\|\, d$ actions	1	1.0000	0.0000	on an empty bucket
a random message	none	1	1.0000	0.3954	never

Table 1: Learned-endpoint outputs at each anchor budget. Each emitted message incurs one online receiver query; calibration reads are paid once for the anchor bucket. Output rate counts in-scope states with a selected message. Action unanimity evaluates every anchor-consistent message on held-out states and is reported separately.

agree, and each recorded action is safe for every interface the transcript leaves open. Then the action the local gate installs at x is safe for every interface the transcript leaves open.

The learned endpoint reports anchor unanimity, neighbour distance, and held-out reference-action fidelity. The noisy-channel study reports interface-level certificate coverage and false-certification under its finite candidate family (Appendices E.8 and E.11).

5 Evidence across learned and released interfaces

5.1 How repair outputs are measured

The *reference pair* is the frozen sender and receiver whose behaviour defines the target; its *reference message* is the symbol sent at a state, and *native decision match* is agreement with the reference receiver's action. For each goal-message bucket (g, m), the message-level audit forms $C_D(g, m)$ from messages matching every recorded anchor action, and evaluates $A_D(x) = \{f_R(x, z) : z \in C_D(g, m)\}$ on held-out context x. The *target scope* is the set of goals scored in an evaluation. The *output rate* is the share of in-scope states receiving a selected message, and *reference-action mismatch* is the share of output states where its action differs from the frozen receiver's action. The *action-unanimity rate* is the share of held-out states on which every message in $C_D(g, m)$ induces the same receiver action; it is computed over all candidates, independently of which message is selected. The *receiver context* is the observation with its message block zeroed; its neighbours are the three nearest anchors carrying the same sender message (Appendices K and E).

For deployment, use the set rule for action fidelity and the gate for support coverage: calibration costs $|B|d$ candidate-action reads versus $|B|$ anchor reads, followed by one online query per output.

5.2 Output rate and action fidelity diverge

At MPE/simple_speaker_listener (width 16, 15 seeds, 225 pairs), the message-map learner issues on 49.19% of states and mismatches on 38.12% of issued states. The context gate mismatches on 39.30%, against a 39.40% population base. With the same anchors, a classifier has 32.91% mismatch at full issue and 32.04% at matched issue 0.5374; the message-set selector has 1.88% at full issue on $22{,}500$ held-out states (Appendices A, B, and M.9). At a matched $|B|d$ calibration budget, the learned classifier and set rule both issue on 1.0000 of sampled states, with mismatch 0.3291 and 0.0188; the $|B|$ gate issues 0.5374 at 0.3057.

5.3 Many-to-one decoding creates over-determination

The reference message is always a candidate, so each state has at least one match. The average state carries 8.65 further matches against $[3.64, 4.18]$ for a control matched on each message's action frequency, and on 14.38 percent of states all sixteen candidates agree. The number of matching messages depends on how many distinct actions the receiver assigns to the sixteen candidates. Across all 225 pairs they induce at most 3.29 distinct actions; this count correlates with the number of reproducing candidates at -0.69 while leaving the evidence-consistent set nearly unmoved (Appendix B.7). The action tracks the context the receiver already holds, and the message changes it on the 85.6 percent of states where the sixteen candidates do not agree (Appendix B.5). The con-

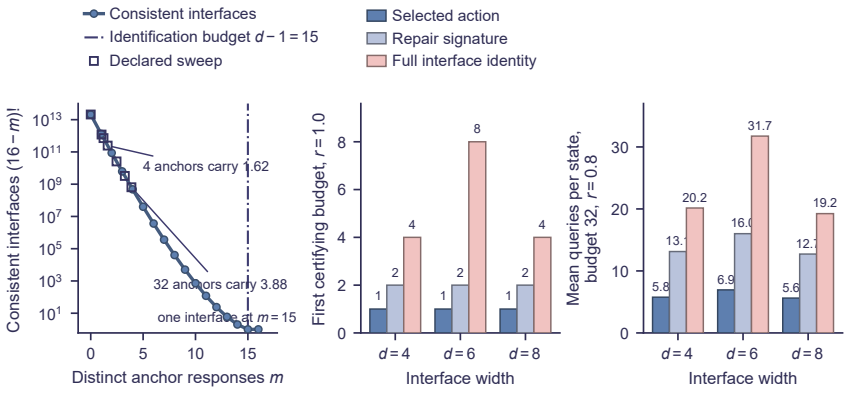

Figure 3: The target-specific budget hierarchy. The left panel tracks the number of probe-consistent interfaces as distinct anchor responses accumulate; the middle panel gives the first certifying budget for each target; the right panel reports mean queries at the largest noisy budget. A selected-action certificate activates before a repair signature and full interface identity.

text conflict rate, how often the same sender message induces different receiver actions in different contexts, is 0.4432 at three neighbours. In most classified conflicts, the target action remains locally realizable (Appendix Figure 6); additional evidence resolves which message produces it.

The local gate often acts beyond its observed support. At thirty-two anchors, 91.2 percent of accepted states lie beyond the empirical agreement radius, where mismatch is 32.61 percent. Restricting anchors to that radius, with nearest-anchor fallback, lowers mismatch to 1.85 percent (Appendix Figure 6 and Appendices B and B.4).

5.4 Evidence improves held-out action fidelity

We vary calibration evidence from 0 to 32 anchors per goal across 204 cells and 396 paired contrasts (Figure 3). An output-rate-matched random placebo isolates the effect of anchor content, while the target hierarchy in Figure 3 shows why the same evidence budget activates a selected-action certificate before a full interface identity.

What the placebo controls. The placebo reuses the gate's output events, so issue rates match exactly. When the gate issues, the placebo installs $\pi(m)$, where m is the gate's message and π is a per-pair alphabet permutation fixed before evaluation; elsewhere it forwards the sender's message. At four anchors per goal, both paired native-match intervals include zero: $[-0.0066, +0.0081]$ and $[-0.0091, +0.0138]$. Raising the budget from four anchors per goal to thirty-two increases native match by 0.0445 with interval $[+0.0339, +0.0542]$ and reduces reference-action mismatch by 0.0723 with interval $[-0.0874, -0.0574]$, at a cost of 0.025 to 0.057 extra queries per state, against 1.006 online queries per state at the declared budget (Figure 2).

Issue rate itself is non-monotonic: it falls from 0.7943 at one anchor per goal to 0.3134 at eight, then recovers to 0.5381 at thirty-two. Separate held-out comparisons retain positive point estimates across the tested neighbourhood sizes and distance metrics (Appendix M).

5.5 Public evidence narrows the candidate set

Across 225 learned-agent pairs, four to thirty-two anchors shrink the mean message set from 8.79 to 1.28 and raise action unanimity from 35.2 to 95.7 percent. At thirty-two anchors, choosing uniformly among consistent messages has an observed 2.1 percent mismatch, giving an empirical upper bound of 2.2 percent on conditional reference-action mismatch among unanimous states; uniform alphabet sampling errs on 39.5 percent (Appendix B.4).

Matched-anchor performance. With four messages realizable, the set rule approaches the anchor-imposed error floor at thirty-two anchors (0.5242 versus the 0.5209 lower bound). With

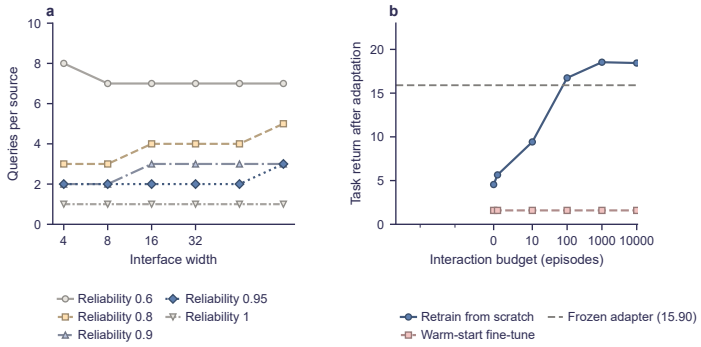

Figure 4: Per-source query cost across interface widths at four detection reliabilities (left), and episodic task return after adaptation across interaction budgets (right). Probing saturates as the interface widens.

one realizable message, every anchor-based rule has error at least 0.9025; the full comparison with local and message-alignment baselines appears in Appendix B.8.

A second task environment. Across six checkpoints and 36 pairs, its anchor-only error falls from 0.2024 at four anchors to 0.1510 at one hundred and sixty, while the same-state lower bound stays near 0.08. Receiver retraining reaches the frozen pair's operating point after one hundred episodes ($2{,}500$ steps); consistent-set repair reaches it without retraining (Appendix B).

Task success and receiver match are distinct outcomes. Receiver fidelity and task success separate under the budget sweep: fidelity rises from 0.5851 to 0.6478, while success ranges from 0.1662 to 0.1684 around the frozen-policy value of 0.1669. At equal probe budgets, a repair focused on the task's action choices improves return by 3.055 at $d = 4$ and 2.989 at $d = 8$ over learning the full interface; an oracle with access to the corrective message raises success to 0.1711 (Appendix B).

Repair restores return on a message-dependent task. On a separate episodic task where the message carries the goal, repair selected from probes raises twenty-step return from 1.4892 to 16.2497 at width four and from 0.7925 to 11.9900 at width eight, compared with 5.0 and 2.5 for controls that ignore the message. Paired gains are $+14.76$ with interval $[13.00, 16.34]$ over ten seeds and $+11.20$ with $[10.74, 11.64]$ over five. For each seed whose messages satisfy the task's one-to-one action contract, repair matches the reference return (Appendix L.3).

The distinct-message statistic gives 0.3654 reference-action mismatch at half issue rate, and gate risk predictions have expected calibration error 0.0249 on $2{,}669{,}625$ issued states. Matched-support controls and the statistic's construction appear in Appendices J, M, E.4, and B.6.

Released MCP audits improve held-out workflow success. One public probe classifies all 54 versions across four MCP servers. Held-out workflow success rises from 0.222 to 1.000 across nine sequential-thinking versions, from 0.300 to 1.000 across ten memory versions, and from 0.821 to 1.000 across twenty-eight everything versions (Appendix B.15). The protocol uses local fixtures over official stdio; rejected versions are replayed with raw JSON-RPC, with one schema-validation path disabled, and the behavior class fixed before held-out workflows are evaluated.

6 Discussion and conclusion

Over-determination explains how a repair can be settled before interface identity. At thirty-two anchors, the learned message set averages 1.28 candidates and induces one receiver action on 95.7 percent of held-out states; among these, conditional mismatch has a 2.2 percent empirical upper bound. The theory gives the exact number of fixed probes needed to separate conflicting repairs, and the noisy stopping rule certifies, gathers a probe, or abstains. On a message-dependent task, repair restores reference-level return; released MCP audits improve held-out workflow success. The audit stops when all remaining interfaces support a common repair.

REFERENCES

Rishabh Agarwal, Max Schwarzer, Pablo Samuel Castro, Aaron Courville, and Marc G. Bellemare. Deep reinforcement learning at the edge of the statistical precipice. *arXiv preprint arXiv:2108.13264*, 2021. doi: 10.48550/arXiv.2108.13264. URL https://arxiv.org/abs/2108.13264.

Bernhard K. Aichernig, Edi Muškardin, and Andrea Pferscher. Active vs. passive: A comparison of automata learning paradigms for network protocols. *arXiv preprint arXiv:2209.14031*, 2022. doi: 10.48550/arXiv.2209.14031. URL https://arxiv.org/abs/2209.14031.

Afra Feyza Akyürek, Ekin Akyürek, Ashwin Kalyan, Peter Clark, Derry Tanti Wijaya, and Niket Tandon. Rl4f: Generating natural language feedback with reinforcement learning for repairing model outputs. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2023. URL https://aclanthology.org/2023.acl-long.427/.

Aaron D. Ames, Samuel Coogan, Magnus Egerstedt, Gennaro Notomista, Koushil Sreenath, and Paulo Tabuada. Control barrier functions: Theory and applications. *arXiv preprint arXiv:1903.11199*, 2019. doi: 10.48550/arXiv.1903.11199. URL https://arxiv.org/abs/1903.11199.

Dana Angluin. Learning regular sets from queries and counterexamples. *Information and Computation*, 1987. doi: 10.1016/0890-5401(87)90052-6. URL https://doi.org/10.1016/0890-5401(87)90052-6.

Stephen Bates, Anastasios Angelopoulos, Lihua Lei, Jitendra Malik, and Michael Jordan. Distribution-free, risk-controlling prediction sets. *Journal of the ACM*, 68(6):43:1–43:34, 2021. doi: 10.1145/3478535. URL https://doi.org/10.1145/3478535.

Fabian Bauer-Marquart, David Boetius, Stefan Leue, and Christian Schilling. Specrepair: Counter-example guided safety repair of deep neural networks. *arXiv preprint arXiv:2106.01917*, 2021. doi: 10.48550/arXiv.2106.01917. URL https://arxiv.org/abs/2106.01917.

Avrim Blum, Steve Hanneke, Chirag Pabbaraju, and Donya Saless. Proofs as explanations: Short certificates for reliable predictions. In *Proceedings of the 38th Conference on Learning Theory*, volume 291 of *Proceedings of Machine Learning Research*, pp. 401–420, 2025. URL https://proceedings.mlr.press/v291/blum25a.html.

David Boetius, Stefan Leue, and Tobias Sutter. A robust optimisation perspective on counterexample-guided repair of neural networks. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 2712–2737, 2023. URL https://proceedings.mlr.press/v202/boetius23a.html.

Thomas Carmichael. Architecture determines observability of transformers. *arXiv preprint arXiv:2604.24801*, 2026. doi: 10.48550/arXiv.2604.24801. URL https://arxiv.org/abs/2604.24801.

Kathryn Chaloner and Isabella Verdinelli. Bayesian experimental design: A review. *Statistical Science*, 10(3), 1995. doi: 10.1214/ss/1177009939. URL https://doi.org/10.1214/ss/1177009939.

C. Chow. On optimum recognition error and reject tradeoff. *IEEE Transactions on Information Theory*, 16(1):41–46, 1970. doi: 10.1109/TIT.1970.1054406. URL https://doi.org/10.1109/TIT.1970.1054406.

T. S. Chow. Testing software design modeled by finite-state machines. *IEEE Transactions on Software Engineering*, SE-4(3):178–187, 1978. doi: 10.1109/TSE.1978.231496. URL https://doi.org/10.1109/TSE.1978.231496.

Dor Cohen and Ofer Strichman. Automated repair of neural networks. *arXiv preprint arXiv:2207.08157*, 2022. doi: 10.48550/arXiv.2207.08157. URL https://arxiv.org/abs/2207.08157.

Ulices Santa Cruz, James Ferlez, and Yasser Shoukry. Safe-by-repair: A convex optimization approach for repairing unsafe two-level lattice neural network controllers. In *2022 IEEE 61st Conference on Decision and Control (CDC)*, 2022. doi: 10.1109/CDC51059.2022.9993239. URL https://doi.org/10.1109/cdc51059.2022.9993239.

Abhishek Das, Théophile Gervet, Joshua Romoff, Dhruv Batra, Devi Parikh, Michael Rabbat, and Joelle Pineau. Tarmac: Targeted multi-agent communication. *arXiv preprint arXiv:1810.11187*, 2018. doi: 10.48550/arXiv.1810.11187. URL https://arxiv.org/abs/1810.11187.

Nirjhor Datta, Swakkhar Shatabda, and M. Sohel Rahman. Frozen but not always accessible: A representation analysis of genomic language models. *arXiv preprint arXiv:2608.05329*, 2026. doi: 10.48550/arXiv.2608.05329. URL https://arxiv.org/abs/2608.05329.

Rita Dorofeeva, Khaled El-Fakih, Stephane Maag, Ana R. Cavalli, and Nina Yevtushenko. Fsm-based conformance testing methods: A survey annotated with experimental evaluation. *Information and Software Technology*, 52(12):1286–1297, 2010. doi: 10.1016/j.infsof.2010.07.001. URL https://doi.org/10.1016/j.infsof.2010.07.001.

Ran El-Yaniv and Yair Wiener. On the foundations of noise-free selective classification. *Journal of Machine Learning Research*, 11:1605–1641, 2010. URL https://jmlr.org/papers/v11/el-yaniv10a.html.

Patrick Emami, Xiangyu Zhang, David Biagioni, and Ahmed S. Zamzam. Non-stationary policy learning for multi-timescale multi-agent reinforcement learning. *arXiv preprint arXiv:2307.08794*, 2023. doi: 10.48550/arXiv.2307.08794. URL https://arxiv.org/abs/2307.08794.

Michał Fica and Jan Otop. Active automata learning with advice. *arXiv preprint arXiv:2508.10535*, 2025. doi: 10.48550/arXiv.2508.10535. URL https://arxiv.org/abs/2508.10535.

Jakob Foerster, Gregory Farquhar, Triantafyllos Afouras, Nantas Nardelli, and Shimon Whiteson. Counterfactual multi-agent policy gradients. *arXiv preprint arXiv:1705.08926*, 2017. doi: 10.48550/arXiv.1705.08926. URL https://arxiv.org/abs/1705.08926.

Jakob N. Foerster, Yannis M. Assael, Nando de Freitas, and Shimon Whiteson. Learning to communicate with deep multi-agent reinforcement learning. *arXiv preprint arXiv:1605.06676*, 2016. doi: 10.48550/arXiv.1605.06676. URL https://arxiv.org/abs/1605.06676.

Feisi Fu, Zhilu Wang, Weichao Zhou, Yixuan Wang, Jiameng Fan, Chao Huang, Qi Zhu, Xin Chen, and Wenchao Li. REGLO: Provable neural network repair for global robustness properties. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(11):12061–12071, 2024. doi: 10.1609/aaai.v38i11.29094. URL https://doi.org/10.1609/aaai.v38i11.29094.

Nathan Fulton and André Platzer. Safe reinforcement learning via formal methods: Toward safe control through proof and learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018. doi: 10.1609/aaai.v32i1.12107. URL https://doi.org/10.1609/aaai.v32i1.12107.

Aditya Gangrade, Anil Kag, and Venkatesh Saligrama. Selective classification via one-sided prediction. In *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pp. 2179–2187, 2021. URL https://proceedings.mlr.press/v130/gangrade21a.html.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems 30 (NIPS 2017)*, 2017. URL https://papers.nips.cc/paper_files/paper/2017/hash/4a8423d5e91fda00bb7e46540e2b0cf1-Abstract.html.

Themis Gouleakis, Christos Tzamos, and Manolis Zampetakis. Certified computation from unreliable datasets. In *Proceedings of the 31st Conference on Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pp. 1958–1990, 2018. URL http://proceedings.mlr.press/v75/gouleakis18a.html.

Mohammed Mehedi Hasan, Hao Li, Emad Fallahzadeh, Gopi Krishnan Rajbahadur, Bram Adams, and Ahmed E. Hassan. Model context protocol (mcp) at first glance: Studying the security and maintainability of mcp servers. *ACM Transactions on Software Engineering and Methodology*, 2026. doi: 10.1145/3814959. URL https://doi.org/10.1145/3814959.

Soroush Hashemifar, Havva Alizadeh Noughabi, Fattane Zarrinkalam, and Ali Dehghantanha. Diagnosing and repairing factual errors in rag under budget constraints. *arXiv preprint arXiv:2606.29377*, 2026. doi: 10.48550/arXiv.2606.29377. URL https://arxiv.org/abs/2606.29377.

Yusuke Hayashi. Lost and found in translation: Variational diagnostics for neural codebook channels. *arXiv preprint arXiv:2605.18846*, 2026. doi: 10.48550/arXiv.2605.18846. URL https://arxiv.org/abs/2605.18846.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2733–2743, Hong Kong, China, 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1275. URL https://doi.org/10.18653/v1/D19-1275.

Xinyi Hou, Yanjie Zhao, Shenao Wang, and Haoyu Wang. Model context protocol (mcp): Landscape, security threats, and future research directions. *ACM Transactions on Software Engineering and Methodology*, 2026. doi: 10.1145/3796519. URL https://doi.org/10.1145/3796519.

Jiechuan Jiang and Zongqing Lu. Learning attentional communication for multi-agent cooperation. *arXiv preprint arXiv:1805.07733*, 2018. doi: 10.48550/arXiv.1805.07733. URL https://arxiv.org/abs/1805.07733.

Daewoo Kim, Sangwoo Moon, David Hostallero, Wan Ju Kang, Taeyoung Lee, Kyunghwan Son, and Yung Yi. Learning to schedule communication in multi-agent reinforcement learning. *arXiv preprint arXiv:1902.01554*, 2019. doi: 10.48550/arXiv.1902.01554. URL https://arxiv.org/abs/1902.01554.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. doi: 10.48550/arXiv.1412.6980. URL https://arxiv.org/abs/1412.6980.

Ker-I Ko and Shia-Chung Teng. On the number of queries necessary to identify a permutation. *Journal of Algorithms*, 1986. doi: 10.1016/0196-6774(86)90013-1. URL https://doi.org/10.1016/0196-6774(86)90013-1.

Dexter C. Kozen. Myhill–nerode relations. In *Automata and Computability*, pp. 89–94. Springer New York, 1997. doi: 10.1007/978-1-4612-1844-9_16. URL https://doi.org/10.1007/978-1-4612-1844-9_16.

Patrick Leask and Noura Al Moubayed. Probing by analogy: Decomposing probes into activations for better interpretability and inter-model generalization. *OpenReview preprint*, 2025. URL https://openreview.net/forum?id=IhI1qGyk6S.

D. Lee and M. Yannakakis. Principles and methods of testing finite state machines: a survey. *Proceedings of the IEEE*, 84(8):1090–1123, 1996. doi: 10.1109/5.533956. URL https://doi.org/10.1109/5.533956.

Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J. Tibshirani, and Larry Wasserman. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113 (523):1094–1111, 2018. doi: 10.1080/01621459.2017.1307116. URL https://doi.org/10.1080/01621459.2017.1307116.

Daphne Lenders, Andrea Pugnana, Roberto Pellungrini, Toon Calders, Dino Pedreschi, and Fosca Giannotti. Interpretable and fair mechanisms for abstaining classifiers. *arXiv preprint arXiv:2503.18826*, 2025. doi: 10.48550/arXiv.2503.18826. URL https://arxiv.org/abs/2503.18826.

Hengyue Liang, Le Peng, and Ju Sun. Selective classification under distribution shifts. *arXiv preprint arXiv:2405.05160*, 2024a. doi: 10.48550/arXiv.2405.05160. URL https://arxiv.org/abs/2405.05160.

Zhen Liang, Taoran Wu, Changyuan Zhao, Wanwei Liu, Xue Bai, Wenjing Yang, Ji Wang, and Wanrong Huang. Birdnn: Behavior-imitation based repair for deep neural networks. *Neural Networks*, 2024b. doi: 10.1016/j.neunet.2024.106949. URL https://doi.org/10.1016/j.neunet.2024.106949.

Hung Ming Liu. Probing the latent world: Emergent discrete symbols and physical structure in latent representations. *arXiv preprint arXiv:2603.20327*, 2026. doi: 10.48550/arXiv.2603.20327. URL https://arxiv.org/abs/2603.20327.

Ryan Lowe, Yi Wu, Aviv Tamar, Jean Harb, Pieter Abbeel, and Igor Mordatch. Multi-agent actor-critic for mixed cooperative-competitive environments. *arXiv preprint arXiv:1706.02275*, 2017. doi: 10.48550/arXiv.1706.02275. URL https://arxiv.org/abs/1706.02275.

Junhao Luo, Ning Huang, Ziqi Sha, Wenxuan Tang, and Wei Deng. Beyond local accuracy: A protocol-level identifiability audit for controlled llm reasoning evaluation. *arXiv preprint arXiv:2608.13326*, 2026. doi: 10.48550/arXiv.2608.13326. URL https://arxiv.org/abs/2608.13326.

Jianan Ma, Jingyi Wang, Qi Xuan, and Zhen Wang. Provable repair of deep neural network defects by preimage synthesis and property refinement. In *Proceedings of the 2025 ACM SIGSAC Conference on Computer and Communications Security*, CCS '25, pp. 4169–4183. ACM, 2025. doi: 10.1145/3719027.3765057. URL https://doi.org/10.1145/3719027.3765057.

Pingchuan Ma, Zongjie Li, Ao Sun, and Shuai Wang. "oops, did i just say that?" testing and repairing unethical suggestions of large language models with suggest-critique-reflect process. *arXiv preprint arXiv:2305.02626*, 2023. doi: 10.48550/arXiv.2305.02626. URL https://arxiv.org/abs/2305.02626.

Gustave Monce, Thomas Degueule, Jean-Rémy Falleri, and Romain Robbes. Client-library compatibility testing with api interaction snapshots. In *2025 IEEE International Conference on Software Maintenance and Evolution (ICSME)*, 2025. doi: 10.1109/ICSME64153.2025.00081. URL https://doi.org/10.1109/ICSME64153.2025.00081.

Andrea Pugnana and Salvatore Ruggieri. Auc-based selective classification. *arXiv preprint arXiv:2210.10703*, 2022. doi: 10.48550/arXiv.2210.10703. URL https://arxiv.org/abs/2210.10703.

Andrea Pugnana, Lorenzo Perini, Jesse Davis, and Salvatore Ruggieri. Deep neural network benchmarks for selective classification. *arXiv preprint arXiv:2401.12708*, 2024. doi: 10.48550/arXiv.2401.12708. URL https://arxiv.org/abs/2401.12708.

Tabish Rashid, Mikayel Samvelyan, Christian Schroeder de Witt, Gregory Farquhar, Jakob Foerster, and Shimon Whiteson. Qmix: Monotonic value function factorisation for deep multi-agent reinforcement learning. *arXiv preprint arXiv:1803.11485*, 2018. doi: 10.48550/arXiv.1803.11485. URL https://arxiv.org/abs/1803.11485.

Geoffrey Roeder, Luke Metz, and Diederik P. Kingma. On linear identifiability of learned representations. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 9030–9039, 2021. URL https://proceedings.mlr.press/v139/roeder21a.html.

Abhishek Sarkar and Bikash Kumar Dey. Message identification over binary noisy permutation channels. In *2025 IEEE International Symposium on Information Theory (ISIT)*, 2025. doi: 10.1109/ISIT63088.2025.11195449. URL https://doi.org/10.1109/ISIT63088.2025.11195449.

Abhishek Sarkar and Bikash Kumar Dey. Message identification over noisy permutation channels: Deterministic encoding and feedback. In *2026 IEEE International Symposium on Information Theory (ISIT)*, 2026. doi: 10.1109/ISIT62367.2026.11654043. URL https://doi.org/10.1109/ISIT62367.2026.11654043.

Philipp Schoenegger, Matt Carlson, Chris Schneider, and Chris Daly. Verifiable semantics for agent-to-agent communication. *arXiv preprint arXiv:2602.16424*, 2026.

Dan Schumacher, Erfan Nourbakhsh, Rocky Slavin, and Anthony Rios. Prompting underestimates llm capability for time series classification. *arXiv preprint arXiv:2601.03464*, 2026. doi: 10.48550/arXiv.2601.03464. URL https://arxiv.org/abs/2601.03464.

Vasileios Sevetlidis. A fiber criterion for representation identifiability in supervised learning. *arXiv preprint arXiv:2606.01092*, 2026. doi: 10.48550/arXiv.2606.01092. URL https://arxiv.org/abs/2606.01092.

Yakov Pyotr Shkolnikov. Do foundation models know geometry? probing frozen features for continuous physical measurement. *arXiv preprint arXiv:2603.06459*, 2026. doi: 10.48550/arXiv.2603.06459. URL https://arxiv.org/abs/2603.06459.

Qingshuang Sun, Denis Steckelmacher, Yuan Yao, Ann Nowé, and Raphaël Avalos. Dynamic size message scheduling for multi-agent communication under limited bandwidth. *arXiv preprint arXiv:2306.10134*, 2023. doi: 10.48550/arXiv.2306.10134. URL https://arxiv.org/abs/2306.10134.

Ryan J. Tibshirani, Rina Foygel Barber, Emmanuel J. Candes, and Aaditya Ramdas. Conformal prediction under covariate shift. *arXiv preprint arXiv:1904.06019*, 2019. doi: 10.48550/arXiv.1904.06019. URL https://arxiv.org/abs/1904.06019.

Zhenting Wang, Qi Chang, Hemani Patel, Shashank Biju, Cheng-En Wu, Quan Liu, Aolin Ding, Alireza Rezazadeh, Ankit Shah, Yujia Bao, and Eugene Siow. Mcp-bench: Benchmarking tool-using llm agents with complex real-world tasks via mcp servers. *arXiv preprint arXiv:2508.20453*, 2025. doi: 10.48550/arXiv.2508.20453. URL https://arxiv.org/abs/2508.20453.

Jannis Weil, Gizem Ekinci, Heinz Koeppl, and Tobias Meuser. Learning to cooperate and communicate over imperfect channels. *arXiv preprint arXiv:2311.14770*, 2023. doi: 10.48550/arXiv.2311.14770. URL https://arxiv.org/abs/2311.14770.

Cameron Wilding, Mina Shaker, and Fatemeh Ganji. Not to break, but to attest: Adversarial probes for privacy-preserving llm verification. *arXiv preprint arXiv:2608.27954*, 2026. doi: 10.48550/arXiv.2608.27954. URL https://arxiv.org/abs/2608.27954.

Nuoya Xiong and Aarti Singh. Cooperative multi-agent rl with communication constraints. *arXiv preprint arXiv:2601.12518*, 2026. doi: 10.48550/arXiv.2601.12518. URL https://arxiv.org/abs/2601.12518.

Xiaodong Yang, Tom Yamaguchi, Hoang-Dung Tran, Bardh Hoxha, Taylor T. Johnson, and Danil Prokhorov. Neural network repair with reachability analysis. *arXiv preprint arXiv:2108.04214*, 2021. doi: 10.48550/arXiv.2108.04214. URL https://arxiv.org/abs/2108.04214.

Yuxuan Yang, Yue Zhang, Lin Chen, and Wei Wang. Defining, observing, and repairing cooperation in LLM multi-agent systems. *arXiv preprint arXiv:2603.00349*, 2026.

Chao Yu, Akash Velu, Eugene Vinitsky, Jiaxuan Gao, Yu Wang, Alexandre Bayen, and Yi Wu. The surprising effectiveness of ppo in cooperative, multi-agent games. *arXiv preprint arXiv:2103.01955*, 2021. doi: 10.48550/arXiv.2103.01955. URL https://arxiv.org/abs/2103.01955.

Lei Yuan, Feng Chen, Zhongzhang Zhang, and Yang Yu. Communication-robust multi-agent learning by adaptable auxiliary multi-agent adversary generation. *arXiv preprint arXiv:2305.05116*, 2023. doi: 10.48550/arXiv.2305.05116. URL https://arxiv.org/abs/2305.05116.

Jiachen Zhang, Junnan Nie, Junyi Lao, Wei Cheng, Chenghao Liu, Jiaxin Jiang, and Songfang Huang. What frozen vlas already know about success: A probing study of value-like structure in foundation robot policies. *arXiv preprint arXiv:2605.28527*, 2026. doi: 10.48550/arXiv.2605.28527. URL https://arxiv.org/abs/2605.28527.

Ziqian Zhang, Lei Yuan, Lihe Li, Ke Xue, Chengxing Jia, Cong Guan, Chao Qian, and Yang Yu. Fast teammate adaptation in the presence of sudden policy change. *arXiv preprint arXiv:2305.05911*, 2023. doi: 10.48550/arXiv.2305.05911. URL https://arxiv.org/abs/2305.05911.

REPRODUCIBILITY STATEMENT

The protocol-server audit reads published package versions installed locally and queries each server over its own standard input and output channel. Each probe is a read-only metadata request. It asks no server to touch a file system, a network peer or a user account. No live public service is contacted: the servers run as local processes from the registry's published artifacts and stop when the probe returns. The audit issues one request at a time, with no retry loop, and excludes servers that require credentials before probing. A deployed audit specifies its endpoint scope and request rate before probing. This audit covers the published local packages and issues one sequential request per audited field. The implementation, the configuration files, the archived outputs and the evaluation scripts are provided as supplementary material. The definitions and the query-cost derivation are in Section 3 and Appendices F and D, and the provenance records that re-audit both endpoints are in Appendix K. The frozen endpoints were trained with MAPPO under the released configuration with Adam Kingma & Ba (2014), and each run records its schedule, seed and checkpoint selection. The evidence archive records fixed seeds, raw outputs and independent checkers; requirements-lock.txt fixes the dependencies and Appendix G records the advice-enhanced automata-learning reference. Every learned endpoint reports its held-out split, seed count and episode count alongside the metric, and the falsification scan records out-of-scope applications separately from certified ones. The code, the data, the compute, the seed handling and the reporting conventions are described in the Code and Data section of the appendix.

Appendix for "Identity Is Not the Decision: Certifying Repairs for Frozen Communication Interfaces"

Contents of this appendix. Thirteen sections carry the derivations, the reanalyses and the extended evidence.

Section A the controlled matrix, the natural stratification, and the provenance records.
Section B additional experiments and controls.
Section C the declared decision rule, the three splits, and the multiplicity correction.
Section D why the turning point sits where it does, and the cost of a target against an identity.
Section E the propositions and their proofs.
Section F the query-cost derivation.
Section G the public automata-learning endpoint.
Section H the per-source budget and the noise law.
Section I the probe-allocation procedure.
Section J the certificate levels, what the public statistic measures, and the statistic comparison.
Section K the audited cells behind the reported counts.
Section L the selective-prediction line and the shift-corrected guarantees, in full.
Section M the extended checks on calibration, the second endpoint and the protocol audits.

A Interaction budget and deployment evidence

The endpoint under study is MPE/simple_speaker_listener at width 16, over fifteen independently trained seeds and all 225 ordered pairs. Each pair carries 64 calibration episodes and a frozen evaluation split of 128 episodes, yielding 31,680,000 state rows across the study.

Width	Profiles	Aliased	No repair	Version-label rule	Behavior-class rule
$d = 4$	6	4	0.33	0.67	1.00
$d = 8$	12	8	0.17	0.50	1.00

Table 2: Successful held-out tool calls through official MCP stdio. The behavior-class rule matches the oracle success rate in both configurations. The version-label rule abstains on the aliased profiles.

Distinct chosen messages	1	2	3	4	5	6	7	8	9
False-safe rate	.1396	.2534	.3249	.3544	.3988	.3870	.4139	.4312	.5012

Table 3: Reference-action mismatch against the primary public statistic, computed over 5,324,775 gate-issued states. A statistic available before the decision and independent of the calibration budget orders the gate's mismatch rate by a factor of 3.6 across its range.

Width	$k = 1$	$k = 2$	$k = 3$	$k = d - 1$	$k = d$
$d = 4$	1	2	3	3	3
$d = 5$	1	2	3	4	4

Table 4: Exact noiseless nonadaptive cost for the support target $q_S(\pi) = (\pi(x))_{x \in S}$. The cost is k until the full bijection is requested, where the classical $d - 1$ budget applies. Noisy certification is evaluated separately through the finite frontier audit.

B Supporting experiments

This section carries the controls and the additional experiments behind Section 5: the gate's coverage against the support it buys, the candidate set audited against its alternatives, the content the installed message carries, the task axis, the transform families and the released protocol servers.

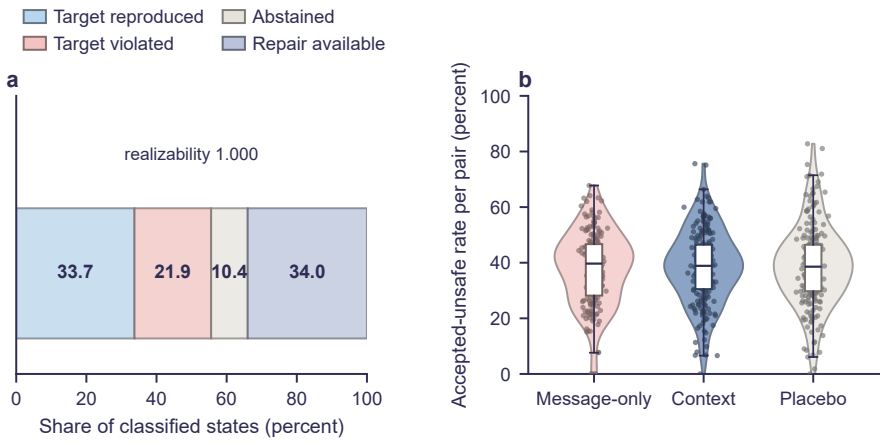

Figure 5: The natural repair issues on states where the anchor evidence calls for abstention, on the primary endpoint. Share of classified evaluation states by gate outcome (left) and per-pair reference-action mismatch for the three arms over 225 ordered pairs (right).

B.1 The gate trades coverage for support

The gate admits a state only when the neighbours it read agree. Weakening it therefore buys coverage, and the record says what it costs. The endpoint stores the receiver context, the sender message, the native action and the receiver's action on every candidate message, so a weaker gate is scored on the same 720,000 held-out states of the declared scope, after the local gate is reproduced exactly on the 22,500 states the decision trail samples: those replay with no mismatch, and the reproduction also recovers the gate's issue rate and reference-action mismatch to four decimals.

	all three agree			two of three			any plurality		
Anchors	Issue	Ref. mismatch	Support	Issue	Ref. mismatch	Support	Issue	Ref. mismatch	Support
1	0.7943	0.3959	0.0000	0.7943	0.3959	0.0000	1.0000	0.3925	0.0562
2	0.6123	0.3998	0.0000	0.8067	0.3969	0.0826	1.0000	0.3914	0.2195
4	0.5568	0.3930	0.0000	0.9427	0.3871	0.3009	1.0000	0.3887	0.3392
8	0.3134	0.3286	0.0000	0.8507	0.3643	0.4520	1.0000	0.3685	0.5219
16	0.4514	0.3013	0.0000	0.9023	0.3233	0.3938	1.0000	0.3260	0.4499
32	0.5381	0.3106	0.0000	0.9162	0.3245	0.3140	1.0000	0.3281	0.3682

Table 5: Issue rate, reference-action mismatch and support violation by gate over held-out states in the declared scope. Support is the share of issued states where the installed message is not one every neighbour the gate read also chose. The unanimity gate has zero support violations by construction, while weaker gates trade support agreement for issue coverage. At four anchors unanimity has the highest mismatch; from sixteen anchors it has the lowest mismatch of the three.

The gate therefore marks one point on a coverage-fidelity curve. At the declared budget, a two-of-three gate lowers reference-action mismatch from 0.3930 to 0.3871 while raising issue rate from 0.5568 to 0.9427, and it is the only gate whose support violation is zero at every budget. From sixteen anchors on it is also the safest of the three, at 0.3013 against 0.3233 and 0.3260. What it buys is a property of the candidate set, which is what the guarantee is stated over, and the states it declines to certify are exactly the states on which a neighbour it read chose otherwise. The weaker gates are not a better rule; they are the same statistic read at a looser threshold, and the threshold is what chooses which of the two properties the operator is buying.

B.2 The gate trade replicates across interfaces

The trade the previous subsection measures is not a property of one environment. Three further records carry the same two objects, the anchor pools and a decision trail, and the same three gates are scored on each under one implementation. On the second endpoint the replay of the recorded rule agrees on 25,035 of the 25,200 sampled decisions, and on the narrower alphabet the whole sweep reproduces exactly, 2,800 of 2,800.

	all three agree		two of three		any plurality	
Interface	Cover	Support	Cover	Support	Cover	Support
four anchors per goal						
Speaker-listener, width 16	0.5568	0.0000	0.9427	0.3009	1.0000	0.3392
Simple reference	0.2767	0.0000	0.3994	0.2094	0.5024	0.3620
Speaker-listener, width 8	0.5525	0.0000	0.9675	0.2687	1.0000	0.2925
Speaker-listener, width 32	0.4700	0.0000	0.9000	0.4778	1.0000	0.5300

Table 6: The same three gates on four records: two tasks and three alphabet widths, at four anchors per goal. Support is the share of issued states where the installed message is not one every neighbour the gate read also chose. Unanimity reaches zero on all four by construction, every weaker gate pays for its coverage, and the two narrow records reproduce the local gate's decisions exactly.

Requiring all three neighbours to agree holds its support violation at 0.0000 at every budget here too, while the two-of-three gate pays 0.2094 at four anchors, 0.4829 at sixteen and 0.5088 at thirty-two, and admitting any plurality pays 0.3620, 0.5999 and 0.6043. The trade is larger on this endpoint than on the primary one, where the two-of-three gate pays 0.3009 at four anchors and 0.3938 at sixteen. The pools differ structurally between the two: a goal on the second endpoint carries several sender messages, so its buckets hold fewer anchors. What the gate buys is the same property on both, and what changes across endpoints is how much a weaker gate gives up to buy coverage.

B.3 An endpoint that shapes its own public surface

The local gate reads evidence the operator collects, and its reference-action mismatch is predicted by a statistic that is a property of the pool. An endpoint that decides what it returns at the operator's probe points can therefore shape the pool. The attack modelled here is the weakest useful one: the endpoint answers every probe with the message its own pool already agrees on, so each bucket becomes maximally consistent, while its behaviour away from the probe points is untouched. It is applied to the recorded anchors and re-scored on the recorded states, so the attack is modelled, and the endpoint it is scored on is the recorded one.

	issue rate		reference mismatch		statistic	
Anchors	honest	shaped	honest	shaped	honest	shaped
4	0.5568	0.6711	0.3930	0.3753	1.6154	1.0000
16	0.4514	0.6110	0.3013	0.3579	3.2530	1.0000
32	0.5381	0.6604	0.3106	0.3691	3.8838	1.0000

Table 7: An endpoint that answers every probe with its own bucket's modal message, against the recorded behaviour. The statistic collapses to its lowest bin at every budget, whose recorded reference-action mismatch is 0.1475, while the realized mismatch rises to 0.3579 and 0.3691 at the two larger budgets.

The statistic collapses to one distinct chosen message at every budget, which is the bin whose held-out reference-action mismatch is 0.1475, so the operator reads a mismatch near a seventh. The realized mismatch is 0.3753, 0.3579 and 0.3691, so the reading is low by a factor of about 2.5. At sixteen and thirty-two anchors the attack also raises the realized rate, from 0.3013 to 0.3579 and from 0.3106 to 0.3691, because a gate that always issues an output acts on states the honest gate declined and issue coverage rises with it. The statistic retains the same value while its predictive calibration changes.

What bounds the exposure is the setting. The probes are the operator's own and are drawn from a seed the endpoint does not see, and the calibration is fitted on one split and judged on another, so an endpoint that shapes only the probes it can identify does not shape the states the risk is measured on. An endpoint that can identify the probe states, or that shapes its behaviour everywhere, is outside what this certificate prices, and the reported risk is therefore conditional on the endpoint not doing so.

B.4 THE CANDIDATE SET, BUILT AND AUDITED

The certificate is defined over the set of candidates the evidence leaves open, and that set is constructed here from the recorded evidence. A candidate is a message of the alphabet; installing it at a state induces one action through the frozen receiver, so its safe set is that action, and the intersection is non-empty exactly when every candidate still consistent with the anchors induces the same action at the state. A candidate is consistent when it reproduces the recorded reference action at every anchor of the state's bucket. The audit recomputes the local gate's decision from the anchors and reproduces the recorded decision with no mismatch before reporting anything.

The objects, and one worked example. Six objects carry the paper and they are not interchangeable, so they are listed together before they are used.

Object	What the probes give	Exact where	Read at deployment
the messages the evidence leaves, $C(x)$	a set narrowing from 8.79 messages to one as anchors accumulate	one message at one hundred and sixty anchors	install the lowest member, with no held-out action
the interface states the evidence leaves open	a family of several states, never collapsed to one	on the finite family (Appendix E)	the message answers of that family
the action the frozen receiver takes	8.65 further messages reproduce it on the average state	measured on the endpoint	the target a repair reproduces
the comparison, a nearest-agreement rule	its support property holds exactly	Proposition 1; its inheritance sits in Appendix E.8	abstains everywhere else

Table 8: Objects used by the audit, from the evidence-consistent message set to the comparison rule. The first row is computable without a held-out answer.

Object	Symbol	What it is
Interface hypothesis	θ	a receiver behaviour: a map from (context, message) to action
Candidate family	Θ	the hypotheses a query set is allowed to leave open
Evidence-consistent set	$\Theta(D)$	the hypotheses that answer the transcript's queries correctly
Safe-action set	$\mathcal{S}_\epsilon(\theta)$	the actions within ϵ of optimal for θ
Decision object	$\mathcal{S}_\epsilon(D)$	the common safe actions, $\bigcap_{\theta\in\Theta(D)} \mathcal{S}_\epsilon(\theta)$
Message	m	one symbol of the alphabet, an input the receiver is fed
Message-consistency set	$C_D(g, m)$	messages matching every anchor action in bucket (g, m)
Neighbour support	$N(x)$	the k nearest same-message anchors the deployed gate reads
Frozen-policy agreement	*none*	the installed message reproduces the reference action, a fidelity reading

Table 9: The objects the paper uses, and which of them is a hypothesis and which is an input. The distinction that carries the most weight is the fifth row against the seventh: a hypothesis determines what is safe, and a message is what the receiver is fed.

A worked example, on three messages and two actions. Let the evidence leave three hypotheses open, $\Theta(D) = \{\theta_1, \theta_2, \theta_3\}$, with safe sets $\{a, b\}$, $\{b, c\}$ and $\{a, c\}$. Every pair overlaps, so a pairwise-conflict test sees no edge and predicts zero probes. The three-way intersection is empty, however, so the current evidence certifies no action: the audit must abstain or seek a probe that splits this conflict and then recompute the surviving family. This is why the exact object is the family of subsets with empty intersection, which Appendix E characterizes. The learned-agent audit evaluates the action image of $C_D(g, m)$ and reports whether its anchor-consistent messages induce a single receiver action.

Two more examples, at the sizes the exact statements are computed at. For the cost separation, take the permutation family on four sources and let the target be the action of the first source. A query on a source reveals its image, so one query on the first source separates every pair of interfaces that disagree about the first source, and that is exactly the set of pairs whose target values differ: the target costs one query. The full identity needs more, because a pair of interfaces can agree on the

first source and differ elsewhere, and those pairs are not separated by that query; on this family the exact nonadaptive costs are three queries for the full interface against one for the first-source target. The same reading gives four against one at five sources. The separation is therefore not a constant factor but a consequence of which pairs the target asks to separate.

For the placebo, take an alphabet of three messages and the permutation that sends $0 \to 1$, $1 \to 2$ and $2 \to 0$. Where the certificate certifies and would install message 1, the placebo installs message 2; where the certificate abstains, the placebo installs the sender's message unchanged. Coverage matches by construction because the certification events are reused, so the two arms certify the same states and the only quantity that moves between them is the content of the message installed. That is what makes the contrast a content contrast at fixed coverage.

The message-level audit. The interface theory maintains the transcript-consistent family $\Theta(D)$ and tests the intersection of its safe-action sets. At the learned endpoint, the audit constructs $C_D(g, m)$ directly from the anchor records. Let $D_{g,m}$ contain calibration anchors with goal g and sender message m:

$$C_D(g, m) = \{z \in [M] : f_R(x_i, z) = a_i^\star \text{ for every anchor } (x_i, a_i^\star) \in D_{g,m}\}.$$

For a held-out context x in bucket (g, m), let $A_D(x) = \{f_R(x, z) : z \in C_D(g, m)\}$ and let $r(x)$ count the candidates matching the reference action $a^\star(x)$. Write $a_D(x)$ for the unique element when $|A_D(x)| = 1$. The audit measures candidate-set size, action unanimity, and reference-action agreement.

A wrong unanimous action makes every candidate message wrong, so its uniform-set loss $1 - r(x)/|C_D(g, m)|$ equals one. Averaging this pointwise implication gives

$$\widehat{\Pr}(a_D(x) \neq a^\star(x) \mid |A_D(x)| = 1) \leq \frac{\widehat{\mathbb{E}}[1 - r(x)/|C_D(g, m)|]}{\widehat{\Pr}(|A_D(x)| = 1)}.$$

At thirty-two anchors, across $45{,}000$ held-out pair-state observations, the measured set loss is 0.020573 and action unanimity is 0.957444, yielding an empirical conditional reference-action mismatch bound of 0.0215. Bootstrap resampling over the 225 ordered pairs gives 95% intervals of $[94.7, 96.7]\%$ for unanimity and $[1.7, 2.7]\%$ for the conditional mismatch.

Anchors	Mean $\lvert C_D \rvert$	Unanimous action	Rule issues	Issues outside it	Inside it
4	8.79	0.352	0.554	0.419	0.932
16	1.85	0.889	0.454	0.070	0.746
32	1.28	0.957	0.538	0.030	0.709

Table 10: The anchor-consistent message set over 225 ordered pairs and 200 held-out states each. Mean set size falls from 8.79 to 1.28, action unanimity rises from 35.2 to 95.7 percent, and rule issues outside the unanimous-action set fall from 41.9 to 3.0 percent. The last column is the share of issued states whose installed action matches the unanimous action.

At four anchors, the local gate issues on 55.4 percent of states while candidate actions are unanimous on 35.2 percent. At thirty-two anchors, unanimity reaches 95.7 percent and the gate's out-of-support issue rate falls to 3.0 percent. These measurements separate candidate-action agreement from the action selected by the deployment rule.

The floor on the message-consistency set. The per-state decomposition in Appendix E can be read on the alphabet or on the empirical set $C_D(g, m)$. The table below reports the uniform-draw error, the deployed rule's error on issued states, and the difference between them. The floor is the error a rule drawing uniformly from the state's own consistent set would make, the gate error is the share of issued states on which the message the gate installs fails to reproduce the reference action, and the last column is the gate's advantage over that draw.

At sixteen and thirty-two anchors, candidate messages induce a unanimous action on 88.9 and 95.7 percent of states. Uniformly sampled set members have 5.2 and 2.1 percent error, while the local gate covers 45.4 and 53.8 percent of states at 29.2 and 30.7 percent error. Candidate-action agreement and selector behavior are distinct parts of the measured repair process.

Anchors	$\lvert C_D \rvert$	Floor on C_D	Gate error	Advantage	Issued outside C_D
4	8.79	0.2985	0.3971	−0.0986	0.419
16	1.85	0.0517	0.2917	−0.2400	0.070
32	1.28	0.0206	0.3074	−0.2868	0.030

Table 11: The selection term read on the message-consistency set, over all 225 ordered pairs. The set narrows toward the reference message as the budget grows while the gate's own choice does not follow, and the advantage goes from -0.10 to -0.29. The reference message lies in $C_D(g, m)$ by construction, so $1 - 1/|C_D(g, m)|$ bounds the floor from above at 0.22 at thirty-two anchors, and the measured 0.0206 lies below it.

B.5 What the message carries

The information the installed message carries about the receiver's action is measured directly, in bits, on the held-out split.

Arm	I(message; action)	H(action)	Share
No adapter	0.0057	2.2807	0.0025
Message-only repair	0.0098	2.2520	0.0044
Agreement gate	0.0122	2.2446	0.0054
Alphabet permutation	0.0091	2.2489	0.0041

Table 12: What the installed message carries about the receiver's action, over all 225 pairs and 15,840,000 held-out rows. The gate's message explains 0.54 percent of the receiver's action entropy, and a random alphabet permutation explains less, because the permutation is drawn independently of the context while the gate's message is drawn to match it.

The share is the quantity that matters for a probe-based reading of the interface: if the message explains 0.54 percent of the action, then a probe that records which message reproduces the behaviour is reading the receiver's context, and the reading cannot be attributed to the code.

B.6 The instruments, stated

Two conventions decide what the numbers in the main text mean, and both are fixed here.

The receiver context, and the metric. The receiver's observation is the raw environment observation, and the context is that observation with the trailing message block cleared. The anchors are put in the same form, so the gate's distance is taken between two objects built the same way, and no further normalisation or whitening is applied: the Euclidean distance is the unweighted distance in the observation's own coordinates. The block is cleared, and the distance the neighbourhood is taken under is one of the axes the main text reports the gain surviving.

The statistic, and its bins. The statistic is the number of distinct chosen messages in a same-message bucket. It is an exact integer and it is binned at its own values, so the bin count follows the reachable range, which grows with the budget: four values at the declared budget and nine at thirty-two anchors. Adaptive binning and monotone regression are not applied, and the calibration error in the main text is computed on these bins against the held-out rates. Whether a monotone fit would sharpen the reading is open, and the bin table is the instrument an operator would threshold either way.

B.7 What task characteristic predicts over-determination

Over-determination is the number of candidates that reproduce the required action. The characteristic that should drive it is how many-to-one the receiver's action map is: if sixteen messages collapse onto few actions, many of them reproduce the required one, and a near one-to-one map leaves little room.

Both quantities are measured per ordered pair over all 225 pairs, over the whole alphabet, so they describe the task as a whole.

	min	median	max
distinct actions the sixteen candidates induce	1.84	2.75	3.29
candidates reproducing the reference action	6.92	9.66	12.57

Table 13: The task characteristic and the over-determination, per pair, over all 225 ordered pairs. The sixteen candidates induce at most 3.29 distinct actions on any pair, so no pair in this family has a near one-to-one action map.

The characteristic predicts the over-determination and leaves the candidate set to the anchors. Across pairs the number of distinct actions correlates with the number of reproducing candidates at -0.694 by Pearson and -0.626 by Spearman, so a more collapsed action map is the over-determined case. The same quantity correlates with the evidence-consistent set at only -0.158 and -0.156 at four and thirty-two anchors, so what narrows that set is the discriminative power of the anchors, while the action map's shape governs the over-determination. The two axes are therefore separate, which is the reading the main text reports: the budget moves identifiability and the task fixes the floor.

No pair in the family reaches the regime the question asks about. The largest number of distinct actions any pair produces is 3.29 of sixteen, and no pair produces eight or more, so the over-determination is structural to the endpoint family. A task with a near one-to-one action map is where this account predicts the gain concentrates, and Section 5 carries that case: every message separates its own signature there, and the repair restores the task.

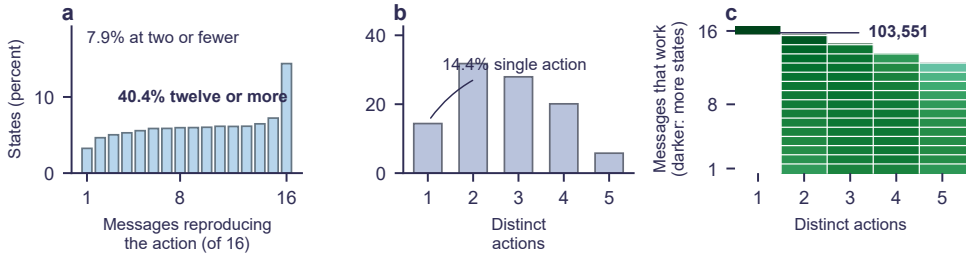

Figure 6: Across 720,000 held-out states, (a) counts messages reproducing the required action, (b) distinct actions induced by the sixteen candidates, and (c) their joint distribution. At least twelve messages reproduce the required action on 40.37% of states; all sixteen agree on 14.38%.

B.8 Where the binding constraint moves

The main text prices the repair by the evidence, and the reason is that the interface is over-determined: many messages reproduce the required action, so the channel is not what runs out. That reason has a boundary, and this section measures it by creating the opposite regime on the same endpoint.

The composition keeps the learned receiver and narrows its message channel,

$$a_K(c, m) = a_{\text{learned}}(c,\ m \bmod K), \qquad K \in \{1, 2, 4, 8, 16\},$$

so $K = 16$ is the receiver as trained and $K = 1$ is a channel that carries nothing. The target is unchanged: the repair is still asked to reproduce the action the trained receiver takes, which means that as K falls the target can leave the set of actions the channel can express. The sweep therefore varies the expressivity half of the identifiability gap, which the main text holds fixed at one.

The collapse sweep isolates the interface limit. At $K = 1$, the sender can express only 0.600 of target actions. At thirty-two anchors, the set rule's error rises from 0.019 at $K = 16$ to 0.903 at $K = 1$. The two rules converge as the channel narrows: at $K = 1$ their errors differ by less than 0.001 at thirty-two anchors. At $K = 16$, evidence quality determines the remaining error and the set rule uses the anchors more effectively; at $K = 4$, the interface limits both rules and both errors exceed 0.5.

Anchors	K	Realizable	Set error	Bound	Gate error	Set coverage
4	16	1.000	0.3210	0.0000	0.4034	1.000
4	8	0.965	0.3397	0.1384	0.4033	0.925
4	4	0.883	0.3543	0.2406	0.4067	0.800
4	2	0.748	0.3960	0.3519	0.4229	0.668
4	1	0.600	0.4297	0.4297	0.4462	0.548
32	16	1.000	0.0185	0.0000	0.3017	1.000
32	8	0.965	0.0931	0.0806	0.3087	0.508
32	4	0.883	0.5242	0.5209	0.6105	0.244
32	2	0.748	0.8683	0.8671	0.8705	0.058
32	1	0.600	0.9025	0.9025	0.9033	0.041
64	16	1.000	0.0012	0.0000	0.2824	1.000
64	8	0.965	0.0678	0.0674	0.3119	0.470
64	4	0.883	0.5883	0.5882	0.6749	0.204
64	2	0.748	0.9255	0.9255	0.9256	0.026
64	1	0.600	0.9295	0.9295	0.9295	0.025
160	16	1.000	0.0000	0.0000	0.2590	1.000
160	8	0.965	0.0667	0.0667	0.2929	0.465
160	4	0.883	0.6000	0.6000	0.6727	0.199
160	2	0.748	0.9333	0.9333	0.9333	0.023
160	1	0.600	0.9333	0.9333	0.9333	0.023
320	16	1.000	0.0000	0.0000	0.2440	1.000
320	8	0.965	0.0667	0.0667	0.2817	0.465
320	4	0.883	0.6000	0.6000	0.6701	0.199
320	2	0.748	0.9333	0.9333	0.9333	0.023
320	1	0.600	0.9333	0.9333	0.9333	0.023

Table 14: The collapse sweep over all 225 ordered pairs, with $K = 16$ the trained receiver reproducing the uncollapsed readings. Narrowing the channel takes the target out of the expressible set: the set rule's error rises from 0.019 to 0.903 at thirty-two anchors, and at $K = 1$ the interface sets it. The zero reached at one hundred and sixty anchors is a plateau, with the local rule at 0.2590 and then 0.2440.

The anchors bound every rule that reads them. A bucket's anchors are the same for every state of the bucket, so a rule reading only those anchors installs one message per bucket and cannot choose per state; a randomized rule averages such choices, which cannot help. Its error on the bucket is therefore at least the number of states the best message the anchors leave consistent gets wrong, and that number is fixed by the anchors and the recorded actions alone, so it can be reported beside any rule the operator runs. The bound column is that quantity, computed per bucket on the same states the rules are scored on. At the trained endpoint it is 0.0000 at every budget: the anchors leave a set of messages any one of which reproduces the reference action on every state, so the residual error there is message ambiguity and nothing else, and it is what the rule's error falls to as anchors are added, 0.3210 to 0.0185 from four to thirty-two. Under collapse the bound rises with the error and the rule sits on it, 0.0033 from it at $K = 4$ and thirty-two anchors and 0.0012 at $K = 2$, and the two coincide at $K = 1$. The residual error at the boundary is therefore the interface's, and no rule reading the same anchors could have avoided it.

Which member the rule installs is worth little. The recommended rule installs the lowest message the anchors leave consistent, so its advantage could be an artefact of that tie-break. It is not: a uniformly random consistent message errs on 0.0206 of states at thirty-two anchors on the trained endpoint against the rule's 0.0185 (Table 15), and on 0.2985 against 0.3210 at four anchors, so the consistent messages are close to interchangeable and what the rule contributes is consistency. The same comparison at $K = 1$, 0.9025 against 0.9025, is the case where the anchors carry no information about the action at all and no choice among them can matter.

Evidence limits and interface limits. On the main endpoint, more anchors reduce the anchor-only rule's mismatch to 0.0000 at one hundred and sixty; in the collapse sweep, narrowing the channel raises the same-state error lower bound from 0.0000 to 0.9025. On the second endpoint, across six checkpoints and 36 ordered pairs, the anchor-only error falls from 0.2024 at four anchors to 0.1510 at one hundred and sixty, while the lower bound on those same states stays near 0.08. The two

Rule	Output rate	Error
a random message	1.000	0.3954
a random consistent message	1.000	0.0206
the set the evidence leaves	1.000	0.0185
local agreement	0.441	0.3017

Table 15: Four rules on the same anchors, all 225 ordered pairs. A random message is the content-free reading; a random consistent message is the tie-break control, and the two differ by 0.0021 here against -0.0226 at four anchors, so what the rule contributes is consistency. It needs no held-out action, dominates the local-agreement rule on both axes, and costs 7.7 to 10.4 times the gate's calibration queries.

endpoints show when added evidence controls mismatch and when the interface's attainable action set sets the floor.

The global-alignment alternative. The nearest non-contextual alternative is a single global map from sender message to receiver message, fitted on the anchors by an optimal assignment on the sender-message to chosen-message confusion and applied without reading the state at all. It is fitted inside each budget on exactly the anchors that budget reads, so the comparison is at equal evidence. It errs on 0.3643 of states at four anchors and on 0.3113 at one hundred and sixty: more evidence barely moves it, against 0.3210 falling to 0.0000 for the context-conditioned rule. The reason is the one the paper measures throughout, that the same sender message needs different repairs in different contexts, so a map that cannot read the context cannot use the evidence it is given.

The budget can be read off the evidence. An operator choosing a budget would like a stopping rule that does not require knowing the answer, and the consistent set supplies one, because it is computable at decision time. On every one of the 225 pairs there is a smallest budget at which the anchor-only rule becomes exact and a smallest budget at which the consistent set becomes a singleton. They coincide on 172 pairs. On the remaining 53 the singleton arrives later and never earlier, so an operator who collects until the set holds one message never stops before the evidence has determined the answer. The mean size of the set falls from 8.768 to 3.746, 1.843, 1.283, 1.026 and 1.000 against an error of 0.3210, 0.1758, 0.0539, 0.0185, 0.0012 and 0.0000, so the error is a monotone function of a quantity the operator can read, and the budget stops being a hyperparameter and becomes a stopping condition.

What separates the states is the evidence. The quantity the operator already holds is the size of the consistent set, and it orders the states: it falls 8.768, 3.746, 1.843, 1.283, 1.026 and 1.000 against an error of 0.3210, 0.1758, 0.0539, 0.0185, 0.0012 and 0.0000. A geometric margin orders them differently: read on the states whose margin exceeds the extent of the neighbours the rule reads, which is between half and two thirds of them, it leaves an error of 0.3929, 0.2964, 0.3104, 0.2873 and 0.2681 at four, sixteen, thirty-two, sixty-four and one hundred and sixty anchors. The characterization the guarantee admits is therefore the evidence's own indeterminacy, which the operator reads off the consistent set.

The radius is the whole story. The account above explains the local rule by geometry: most gate-accepted states lie outside the radius at which the anchor actions are observed to agree, so the neighbours a local rule can reach sit outside the radius its assumption needs. That explanation is testable without new data, by keeping the rule and changing only which anchors it is allowed to read. Restricted to the anchors inside that radius, and falling back to the single nearest when none is, the local rule errs on 0.3210, 0.0185 and 0.0000 of states at four, thirty-two and one hundred and sixty anchors, which is exactly the anchor-only rule's error at all three budgets. The number of anchors it can read inside the radius rises from 1.10 to 3.90 to 18.87 over the same range. So the radius is not a correlate of the local rule's behaviour but the thing that governs it: given neighbours inside the radius the rule is optimal, and outside it the same rule errs on 0.3261 at thirty-two anchors. This is also the quantity the density measurement prices, since the same-message bucket would need about $1{,}070$ anchors per goal to place its neighbours inside the same radius. The choice an anchor records is the lowest message that works, and 63.5 percent of the anchors record the first message. Reading anchors one at a time therefore identifies very little, and the identification the paper uses comes from requiring a whole bucket at once, which narrows the set to 1.28 messages at thirty-two

anchors. The sweep's variants agree with that reading. A rule that installs the majority message of the nearest anchors, which always issues, errs on 0.3261 of states at $k = 3$ and on 0.3431, 0.3239, 0.3249 and 0.3276 at k of one, five, eight and sixteen, against 0.0185 for the rule that reads the bucket and 0.3954 for a message drawn at random from the alphabet. The geometry agrees as well: 91.2 percent of gate-accepted states lie outside the radius at which the anchor actions are observed to agree, so the neighbours a local rule can reach sit outside the radius its assumption needs.

B.9 The task axis does not move

The fidelity gain lives at sixteen to thirty-two anchors, so the same grid is read on the task. The receiver's actions solve the goal on 0.1662 to 0.1684 of states across the sweep, against 0.1669 for the frozen policy left unrepaired, while native decision match rises from 0.5851 at zero anchors to 0.6478 at thirty-two. Every paired task contrast is indistinguishable from zero at every budget, against both the permutation and message-only repair, the widest being $+0.00182$ with interval $[-0.00006, +0.00379]$.

The two axes therefore separate by 0.0627 of fidelity across the grid at no measurable task cost, which is the measured form of the claim that fidelity to a frozen policy and task success are different quantities on this endpoint. The certificate buys agreement with a decision that solves the task on 16.84 percent of states, and it buys it at every budget without moving that rate.

B.10 Soundness under falsification

A certificate that carries risk is only useful if it also refuses unsound applications. We construct five stress cases with declared expected behavior, covering a target class whose label is non-constant, a target action that is non-constant within a behavior class, a message-sensitive receiver that is context-inconsistent, an unsafe target completion applied outside its certified target, and a utility ambiguity with identical public behavior. All five return the expected outcome.

We then scan the natural states for out-of-scope use. Across the scan, $108{,}250$ applications were made outside the certified target, and none of them were counted as certified. The distinction matters because an unsafe rate computed over all applications would look worse while carrying less information. The quantity that governs safety is whether the certificate was claimed on states where it holds.

B.11 Finite-sample behavior across message widths

A certificate is useful only if the evidence it needs stays bounded as the alphabet grows. We sweep message widths from 4 to 128 and channel reliabilities from 0.6 to 1.0 against a declared success threshold of 0.95, with four thousand full-permutation trials and twenty thousand target trials per cell, giving 30 cells at twenty-two query budgets each.

Every cell reaches the declared threshold. The quantity that governs cost is the number of repeated queries per source, and it saturates in the width. At reliability 0.6 the requirement is eight queries per source at width 4 and seven for every width from 8 to 128; at 0.9 it is two at widths 4 and 8 and three from 16 to 128; at full reliability a single query per source suffices at every width. Total cost therefore grows linearly in the number of sources while the per-source requirement does not grow at all.

That is the shape a practitioner needs, because a wider alphabet does not make each source harder to read. The budget can be planned per source. The released implementation also keeps three quantities separate that are often conflated, namely the one-query channel capacity, the pairwise Chernoff information, and the finite-sample recovery probability, and it reports the pairwise union fields as loose diagnostics for the global assignment decoder.

B.12 Transform families

The mechanism should not depend on the structure of the transform that created the mismatch. We evaluate five families on the same endpoint. These are cyclic permutations, dihedral transforms, general invertible row-stochastic transforms, near-permutation stochastic transforms, and block per-

mutations. General, near-permutation and block each contribute 25 cells, five checkpoint seeds crossed with five staleness levels, and all pass both the structural and the scientific screen.

B.13 MATCHED-BUDGET TASK RETURN

We also compare a full-interface learner against a localized decision target on task return at an identical total probe budget. The full-interface baseline distributes its budget across the $d-1$ sources and solves the global permutation likelihood, while the localized target spends the same budget on the single source that governs the deployment goal. We select the candidate transform on a 50-episode calibration stream, require that its worst wrong target mapping lose at least 0.25 return against the oracle, and freeze both the transform and the target source before evaluating 200 held-out episodes on twelve seeds per width.

At the unsaturated budget the localized target improves return by 3.055 for $d = 4$ with interval $[0.483, 6.085]$ and by 2.989 for $d = 8$ with interval $[0.618, 5.793]$, at a target accuracy of 1.000 against 0.667 for the full-interface baseline at both widths. The two budgets saturate together at larger values, which is consistent with the constant per-source repetition count established above. On the primary learned endpoint the frozen policies solve the task on 16.84 percent of states, so fidelity to those policies and task success are separate axes there; the released servers in Appendix B.15 carry the task-success axis.

B.14 COMPARISON WITH A TRAINED RECEIVER

The alternative to certifying a frozen pair is to retrain the receiver. We compare the two under a declared interaction budget in which the certificate is fitted from public probes and retraining consumes environment episodes. On the declared finite task with twelve seeds and four hundred evaluation episodes per condition, a receiver trained from a fresh initialization returns 4.548 before any environment episode, 9.419 after ten, 16.740 after one hundred, and 18.535 after one thousand, against a frozen adapter that returns 15.902.

The comparison is decided by what each procedure needs. The certificate reaches the adapter's operating point from public probes on a frozen pair with no environment interaction, and retraining requires roughly one hundred episodes of it. The certificate is therefore a deployment-time procedure, and it is the appropriate tool when a deployment cannot afford a training budget. We also report a warm-started receiver that continues from the canonical policy. It stays at 1.589 across every budget we tried, so it inherits the canonical policy's operating point.

B.15 MCP PROTOCOL EVALUATION

We repeat the decision comparison across the official Python MCP SDK (mcp==2.2.0) over stdio transport.[1] The client completes initialize, tools/list, and tools/call. Public tool metadata is identical across profiles. The hidden server profile applies a cyclic behavior transform, so $K > d$ creates permanent label aliases. Each audit spends one public probe and then evaluates 100 held-out tool calls.

The MCP study provides a protocol-layer validation of the same decision rule and isolates protocol-level compatibility with a local versioned fixture. MCP-Bench evaluates tool-using language-model agents on live servers Hasan et al. (2026); Hou et al. (2026); Wang et al. (2025); the endpoint here is cross-version compatibility and repair.

Released-version history. We audit seven published versions of the MCP filesystem server from 2025.3.28 through 2026.8.31. Each package is launched over stdio, calibrated on public list and file-info calls, and evaluated on held-out files. Public behavior forms three classes, and version-label accuracy is capped at 0.429 while the behavior-class decision remains exact on held-out probes.

Released task histories. We audit nine sequential-thinking, ten memory, and 28 everything versions. In sequential-thinking, one probe identifies the class and type normalization raises a held-out

[1] Model Context Protocol specification, version 2025-06-18, https://modelcontextprotocol.io/specification/2025-06-18.

workflow from 0.222 to 1.000; wrong repair matches no repair. In memory, a read probe separates response-envelope classes and a graph adapter raises entity-relation success from 0.300 to 1.000. The memory result uses a client compatibility layer with result-schema validation disabled; its estimand is workflow compatibility. In everything, one tools/list probe maps ten public classes to the advertised add or get-sum action; sum success rises from 0.821 to 1.000, and wrong repair succeeds on none.

To test workflow composition, we freeze three topologies, two held out, including reverse order and two-sum variants, and cross them with the observed behavior classes. Class repair and class-blind fallback both succeed on all 270 units; class repair uses 1170 total interactions including probes versus 1369 for fallback, a 14.5% reduction. On the 180 held-out units the reduction is 14.8%, and the paired class-combination bootstrap interval remains below zero. The no-repair baseline succeeds on $6/270$ units, whereas the wrong repair succeeds on none; an independent audit checks all 1080 policy rows.

Public-surface census across the reference servers. The four audited servers are a subset of what an auditor can read without a credential, so we launch every published version of each official server that starts credential-free and record the public tool surface it returns: 91 versions across six servers. Seventy-one answer the installed client; 10 more answer a raw JSON-RPC reader whose response the installed client rejects because the tool list omits a schema field the current revision requires; and 10 do not run as published at all, three because the release imports a module it does not declare and seven because it exits without a credential. Among the 81 versions whose surface can be read, the surface is not constant inside a chain, but it never identifies the version either: it takes 10 distinct values for everything, 7 for filesystem, 5 for puppeteer, 4 for sequential-thinking, and 3 each for memory and github, so the most informative chain pins at most 0.63 of its versions and the least pins 0.21.

Server	Versions	Readable	Distinct surfaces	Label ceiling
everything	28	28	10	0.357
filesystem	19	16	7	0.437
github	13	6	3	0.500
memory	14	14	3	0.214
puppeteer	8	8	5	0.625
sequential-thinking	9	9	4	0.444

Table 16: The public tool surface of every published version of the six credential-free servers, read through the stdio client. Readable counts the versions whose tool list any reader could obtain; the label ceiling is the number of distinct surfaces over that count, which bounds what a version-label decision recovers from public metadata. Of the 10 unreadable versions, three do not declare an imported module and seven exit without a credential.

The full filesystem chain, and where the calibration runs out. The audited filesystem chain covers seven of the nineteen published versions. Extending it to all nineteen, read and called over raw JSON-RPC because the installed client refuses thirteen of them, gives eight observable classes and a label ceiling of 0.421 against 0.429 across the seven, and it locates a boundary the shorter chain cannot show. Six releases from 2025.1.14 through 2025.8.21 answer the three calibration calls identically and split into two classes on the held-out calls. The difference is read_text_file, which the later half advertises and implements and the earlier half does not advertise at all: the calibration never calls it, so no amount of repetition on those three calls separates the two halves, while the public tool list does, eleven or twelve entries against fourteen. Extending the chain therefore measures the evidence, which is the subject of this paper. Three of the nineteen do not run as published at all, because the release imports a module it does not declare.

The memory chain, extended. The memory audit above covers ten of the fourteen published versions. Extending it to all fourteen adds the four earliest releases, which form a fourth observable class and move the label ceiling from 0.300 to 0.286. Those four predate the entity and relation tools the workflow calls, so the class-conditioned repair cannot be exercised on them; on the ten versions that implement the workflow the class gate is exact, which is the measurement the main text reports.

A browser server, and drift one layer down. A fifth credential-free server, the puppeteer browser server, adds eight releases and a tool family none of the others has. We hold the browser release fixed across the chain, so the server is the only variable that moves, and drive it by navigating to a data URL and reading a computed value back. One probe identifies the behavior class here too. The eight releases split three ways: three answer the current client and run the workflow, four answer the navigation call correctly but inside an envelope the current client rejects, since the payload arrives nested under `toolResult`, and one predates the tool names, exposing `navigate` where the later releases expose `puppeteer_navigate`. The census above counts the list layer, where all eight of these releases are readable; this drift sits one layer down, at the call.

B.16 THE PER-SOURCE BUDGET IS WIDTH-INVARIANT

The per-source budget is a property of the interface. The same protocol runs at width 8 and width 32 with two checkpoints each, trained under the same configuration as the primary endpoint, and the online query cost is one per state at all three widths, so the budget saturates per source independently of the alphabet.

The arm whose cost does grow with the alphabet is the scan. The evaluation-only probe reads every candidate message, so it spends one query per candidate and therefore d per state, while the certificate spends one per state at every audited width: the mean online query count is 1.0000 at widths 8, 16 and 32 against 8, 16 and 32 for the probe. A wider alphabet makes the audit more expensive to read and does not make the decision more expensive to take, which is the cost reading of the same reversal: what the operator pays for is the evidence about the source, and the scan is priced by the alphabet.

B.17 IS THE INTERFACE'S REDUNDANCY STRUCTURAL?

One match is guaranteed at every state, since the reference message is itself a candidate, so the existence question is closed by the construction and cannot carry evidence either way. What can is the count beyond that guaranteed one. On the average state 8.65 further messages reproduce the target action. A control that keeps each message's action frequency over the held-out states and breaks its pairing with the state predicts $[3.64, 4.18]$, so the observed redundancy is more than twice what the per-message action frequencies alone account for: it is a property of how the two checkpoints disagree. The control is a band because the identity of the reference message is not recorded, so the quantity it predicts is $\sum_{m \neq r} p_m(a)$, which lies between $\sum_m p_m(a)$ and $\sum_m p_m(a) - \max_m p_m(a)$. The measured value falls outside both bounds, so the reading does not depend on that unknown. The marginals of this recount reproduce the recorded aggregate over all $720{,}000$ states exactly.

B.18 HOW FAR THE LOCAL GATE SITS FROM ITS OWN CONDITION

The guarantee is conditional on a locally stable radius: inside it, any two receiver contexts admit the same safe action under every evidence-consistent interface. That radius is measurable from the anchor pools. Among anchors of one goal, the largest distance at which every pair still agrees on its action is $\delta^\star = 1.692$ over $334{,}800$ anchor pairs. The local gate issues at a median neighbour radius of 2.478, so 91.2 percent of the $12{,}465$ gate-accepted states in the audited sample lie outside the radius at which the condition is observed to hold, and the reconstruction reproduces the recorded decisions on all $22{,}500$ states it reads.

B.19 WHAT THE NEIGHBOUR RULE RUNS ON

The context the neighbour rule measures is the receiver's public state. That observation is 24-dimensional and its last 16 coordinates carry the incoming message, which the context rule zeroes, so the distance is taken on the first 8 coordinates; across the audited anchors between 6 and 9 of those 24 coordinates are non-zero, and the effective dimension of the search is therefore 8 out of the observation's 24. The condition the guarantee names is a radius, so it carries no dimension of its own, and on this endpoint the reading is not carried by the metric: cosine distance and coordinate-standardised distance move the matched contrast by less than a point against unweighted Euclidean, as reported in the main text. What the dimension changes is how many anchors fall inside a given radius, which is why the radius is measured on the endpoint; the context dimension itself is not swept

here, and the condition becomes harder to satisfy as the public state grows, since a fixed radius then contains fewer anchors of the same source.

B.20 The condition does not explain the errors

If the mismatch came from issuing outputs across an unstable neighbourhood, tightening the radius should lower it. It does not. Adding a rule that abstains whenever the neighbour radius exceeds $\delta^{\star}$ moves issue coverage from 0.5793 to 0.0433 and reference-action mismatch from 0.4053 to 0.4624 on forty pairs, against the unrepaired baseline: coverage collapses while the error rate rises, and among the states that remain inside the radius anchor agreement is still wrong about 46 percent of the time. The gate's mismatch is therefore not a neighbourhood-scale effect. The anchors do not determine the target action however close the neighbours are.

C Pre-registration, splits and correction

The decision rule as declared

The decision rule for the main contrast was declared before any evaluation state was read, and the declaration is recorded with the results. It fixes the primary split, the headline target scope, the neighbour count, the contrast itself, a bootstrap with 2000 replicates at $\alpha = 0.05$ under a fixed seed, four quantified thresholds, and the order in which the branches are tested. Intervals are reported following the practice for stochastic reinforcement learning evaluation Agarwal et al. (2021): the resampling unit is the ordered pair, never the individual state, so an interval absorbs the between-pair variation that a per-state interval would hide. The run recorded which branch fired, the declaration's fall-through, and that record reproduces from the measured values. The realizability branch fails because local realizability is 1.0000 against a threshold of 0.5, the context-effect branch fails because the conflict rate is 0.4432 against 0.05, and the mechanism branch fails because the native-match gain is $+0.0007$ against a required 0.02. The reference-action mismatch increase of $+0.0118$ sits inside its declared 0.02 budget, so the rule rejects the mechanism for buying no measurable fidelity. The verification table below restates the four thresholds against the values that were measured. The second endpoint was run under the same declared rule, and its own declaration fixes the same primary split, target scope, neighbour count, contrast, bootstrap and thresholds, so the branch it lands on is read the same way.

Branch	Test	Measured	Outcome
Interface Limited	realizability < 0.5	1.0000	not triggered
No Context Effect	conflict rate < 0.05	0.4432	not triggered
Mechanism Supported	match gain ≥ 0.02	$+0.0007$	not triggered
	reference mismatch increase ≤ 0.02	$+0.0118$	satisfied
Calibration Limited	fall-through	n/a	**recorded**

Table 17: The pre-registered branch order against the measured values, on the held-out split at target scope three. The mechanism branch is selected by the match gain, while its reference-action mismatch increase stays inside the declared tolerance.

The three splits are drawn from disjoint, seeded schedules: calibration trials from $100000 + 1009\,s + 17\,r + 7919\,t$, evaluation episodes from $200000 + 1009\,s + 17\,r + e$, and held-out episodes from $300000 + 1009\,s + 17\,r + e$, with s and r the sender and receiver seeds. The held-out schedule was never used for fitting. The unit of resampling is the ordered pair, of which there are 225, so an interval over pairs already absorbs the between-pair variation a per-state interval would miss, which is the reporting practice recommended for stochastic reinforcement learning Agarwal et al. (2021). The paired tests are corrected across the six comparisons reported in the experiments using Holm's step-down procedure; the coverage contrast moves from a raw $p = 0.0004$ to $p = 0.0022$, and the four null comparisons remain at $p \geq 0.57$.

D THE ANCHOR LADDER AND THE TARGET-AWARE FRONTIER

Two exact statements explain the turning point of the budget sweep and the ordering of the three certificate levels. The first is the classical counting of permutation queries; the second is the finite frontier theorem whose consequences the audit reports.

D.1 WHY THE TURNING POINT SITS WHERE IT DOES

For a family whose public interface is a bijection on d symbols, a public anchor reveals the image of the queried source, and the unqueried source-target pairs can then be matched in any way, so m distinct anchor responses leave exactly $(d-m)!$ consistent interfaces Ko & Teng (1986). The ladder's index is the number of distinct responses, and the recorded pools put a measured count on it. Reading each calibration bucket in capture order and sweeping the anchor budget, over all 225 pairs, a bucket carries 1.03 distinct chosen receiver messages at one anchor per goal, 1.22 at two, 1.62 at four, 2.46 at eight, 3.25 at sixteen and 3.88 at thirty-two. The response count climbs at a falling rate and levels off inside the alphabet.

Anchors per goal n	1	2	4	8	16	32
Distinct responses m, measured	1.03	1.22	1.62	2.46	3.25	3.88
Consistent interfaces $(16-m)!$, at the nearest m	15!	15!	14!	14!	13!	12!

Table 18: The consistency ladder at the audited width, indexed by what the pools carry. The count is exact for a bijective interface on $d = 16$ symbols. A bucket reaches four distinct responses at the widest budget, where the arithmetic that reads the budget as the response count would have closed the candidate set.

The measured sweep agrees with this index: at four anchors per goal the paired gain over message-only repair is $+0.0007$ with interval $[-0.0066, +0.0081]$ and the difference from the coverage-matched placebo is $+0.0022$ with interval $[-0.0091, +0.0138]$, while at sixteen to thirty-two anchors the same contrast is $+0.0522$ with interval $[+0.0415, +0.0634]$. The budget at which the ladder closes is set by how fast a bucket runs through its responses, which the recorded pools fix before any evaluation state is read.

The count is an upper bound on what a full identity costs and not a description of what a deployment needs, which is the second statement.

Figure 3 separates two budgets that the ladder alone does not. The onset, between four and eight anchors per goal, is set by the target. A decision needs only the probes that hit the disagreement sets of pairs whose target values differ, and the declared target is supported on one or two coordinates, so it costs one or two probes where the full identity costs $d-1$. The saturation is what the ladder prices, and the measured response count says why the sweep has not reached it by sixteen anchors: a bucket carries 3.25 responses there against the $d - 1 = 15$ the threshold needs, so the coverage-matched gain is still climbing, from 0.0356 at sixteen anchors to 0.0421 at thirty-two.

The grid was then extended by one point to test that reading, at twelve anchors per goal, which the original declaration did not visit. The coverage-matched gain over the gate's matched permutation was predicted to fall between the eight and sixteen anchor values, and it does. Reading the pooled coverage-matched difference on the held-out split at target scope 3 gives -0.0022 at four anchors, $+0.0082$ at eight, $+0.0216$ at twelve, $+0.0356$ at sixteen and $+0.0421$ at thirty-two, monotone across the range. The reference-action mismatch difference is monotone too and then stops, at $+0.0040$, -0.0263, -0.0645, -0.0788 and -0.0782, which places the saturation in the range between sixteen and thirty-two. The extended sweep was audited the same way as the declared one, with the state replay reproduced against the audited evaluation rows ($1{,}440{,}000$ states, zero mismatches) and 226 pooled cells passing their metric checks.

The twelve-anchor point is reported as a pooled difference because the summarizer emits paired bootstrap intervals for the declared budget pairs, and twelve is not one of them; every paired interval quoted in this paper remains the one recorded on the declared grid, and the two are not interchanged.

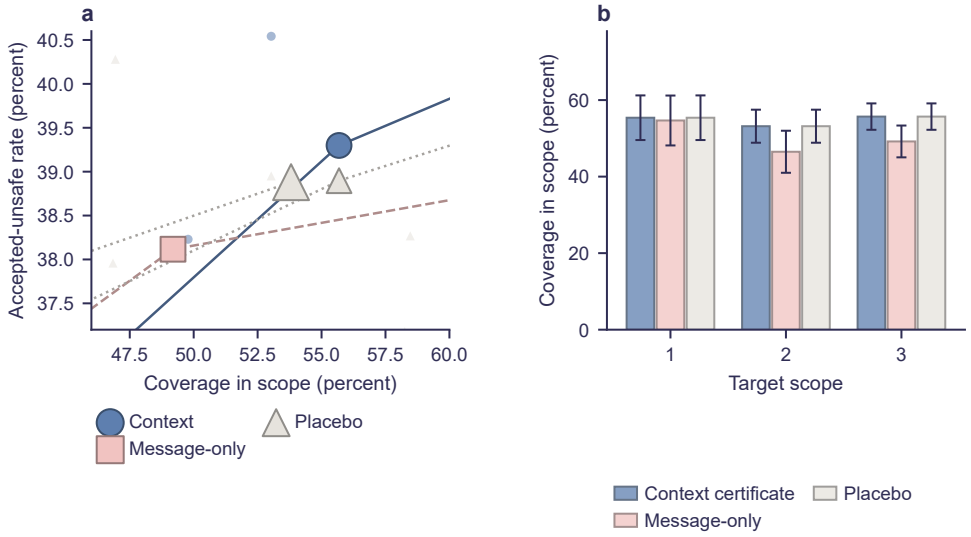

Figure 7: Issue rate and reference-action mismatch as the calibration budget grows (left), and issue rate by target scope for the three arms with 95 percent intervals over pairs (right). Issue rate moves with the budget and agreement requirement, while reference-action mismatch remains comparatively stable.

D.2 A DECISION CAN COST LESS THAN AN IDENTITY

For a single-valued target q on the candidate states, let $D(\theta, \theta')$ be the set of probes on which two candidates disagree, and let P_q be the pairs whose target values differ. A query set identifies q exactly when it intersects $D(\theta, \theta')$ for every pair in P_q, so

$$Q_q^\star \;\le\; Q_{\text{identity}}^\star.$$

In the declared family, the image of source 0 costs one query while the full permutation identity needs $d-1$ queries; the finite validator checks the $d=3$ instance exhaustively. The repair signature used by the declared experiment depends on two coordinates and therefore costs two queries. A selected-action repair is the set-valued case. Its exact cost is the hitting set over $E(A)$ for $A \in \mathcal{M}_\epsilon$, as defined in Appendix E.9; disjoint-safe pairs provide a lower bound and become exact when the safe-action family is 2-Helly. The finite non-Helly witness has exact costs 1 for a selected action, 1 for a target symbol, and $d-1$ for full identity.

D.3 THE MEASURED ACTIVATION LADDER

The same decomposition was run as a noisy audit with a time-uniform likelihood-ratio confidence set over a declared catalog, at widths $d \in \{4, 6, 8\}$, reliabilities $\{0.8, 0.9, 1.0\}$, budgets $\{1, 2, 4, 8, 16, 32\}$, 512 trials per cell and $\delta = 0.05$. The first budget at which the auditor returns anything, and the mean query cost at the largest budget at reliability 0.8, are:

The same confidence rule was then applied to a frozen learned endpoint, moving beyond the declared catalog: 270 cells over five seeds, three target sizes, three reliabilities and six budgets, with 256 trials per cell at $\delta = 0.05$. Minimum empirical coverage was 0.97265625, above the declared $1-\delta$, the maximum unsafe-return rate was 0.02734375, below δ, and an independent consistency audit over the 810 recorded rows returned no errors. That audit is bounded exactly as the declared channel is, with a finite cyclic transform family, injected replacement noise on behavioral anchors, and held-out learned action signatures.

Audited object	Width	First certifying budget	Mean query cost
Selected action	$d = 4$	1	5.76
Repair signature	$d = 4$	2	13.13
Full identity	$d = 4$	4	20.15
Selected action	$d = 6$	1	6.94
Repair signature	$d = 6$	2	16.01
Full identity	$d = 6$	8	31.75
Selected action	$d = 8$	1	5.62
Repair signature	$d = 8$	2	12.72
Full identity	$d = 8$	4	19.24

Table 19: The audited activation ladder: the first certifying budget is read at $r = 1.0$, the mean query cost at 32 probes and $r = 0.8$. The target-dependent levels activate at one and two queries at every width, and the full-identity level is catalogue dependent. Across all 162 cells no returned action was unsafe.

D.4 The evidence cost is priced per source

The finite-sample audit holds this measurement across a thirty-two-fold change in the alphabet, over six widths from $d = 4$ to $d = 128$, five reliabilities, and the declared error level, with Wilson intervals on every cell. Table 20 reports the number of queries each source needs before its interface is identified.

Reliability	$d = 4$	$d = 8$	$d = 16$	$d = 32$	$d = 64$	$d = 128$
0.60	8	7	7	7	7	7
0.80	3	3	4	4	4	5
0.90	2	2	3	3	3	3
0.95	2	2	2	2	2	3
1.00	1	1	1	1	1	1

Table 20: Queries per source needed to identify its interface, by alphabet width and anchor reliability, over 30 cells. A thirty-two-fold increase in the alphabet costs at most two extra queries.

The near-flatness is the point: an interface over 128 symbols is not 32 times more expensive to pin down than one over 4 when the queries are charged to the source that needs them. This is also why the certificate's online cost stays at roughly one query per state (Appendix J) while the calibration budget, which is spent once per source, is what moves the result.

Endpoint terms. The *target scope* is the set of landmark goals scored, taken as the first k goals in ascending goal id. The *issue rate* is the share of in-scope states where the learned gate returns a message, and *reference-action mismatch* is the share of issued states where that message changes the frozen receiver's action. This is a measure of fidelity to the reference pair. A *receiver context* is the receiver observation with its incoming message block zeroed. The local gate selects the lowest-index anchor-consistent message and issues only where its three nearest anchors of the same sender message agree on the message and target action. The noisy finite-family study separately evaluates an interface-level confidence-set certificate (Appendices E and E.11).

Worked alias class. For a cyclic transform of order four with six releases, versions v and $v+4$ are behaviorally equivalent. The repair target $q(v) = v \bmod 4$ is identifiable, while the version label has best accuracy 0.667. The $d = 4, K = 6$ experiment tests whether this distinction changes task return, over and above a label score.

E Formal statements

These formal results comprise exact finite-family characterizations of the quotient and query costs, a time-uniform noisy-channel guarantee, and a conditional local-transfer proposition. The learned-endpoint analyses report anchor unanimity, neighbour distances, and reference-action fidelity.

E.1 THE UNANIMITY GATE'S SUPPORT PROPERTY IS EXACT

The third axis the evaluation reports has a one-line proof, and the proof is why the measurement is exactly zero.

Proposition 1. Let a state have a bucket B of calibration anchors and let $N \subseteq B$ be the k nearest under the declared metric. A gate that issues only when the anchors of N agree on the chosen message returns a message that every anchor of N also chose. A gate that emits the modal message from a strict subset $N' \subset N$ on a state where N is not unanimous returns a message that is not shared by all of N.

Proof. The unanimity rule returns the common value of $a.\mathrm{chosen}$ over N, so every anchor of N carries the returned message and the support violation is empty. The reduced gate returns a value chosen by at least one anchor in N'. When N is not unanimous, at least one anchor in N chooses a different value, so the output lacks unanimous support across N. □

The violation is identically zero for the unanimity rule. For a strict-subset rule that emits a message from the anchors it reads, the violation is the share of issued states whose full neighbourhood is nonunanimous. The metric is silent when every issued neighbourhood agrees; Tables 5 and 6 report this support-agreement cost for the audited weaker gates.

E.2 WHAT NO RULE READING THE SAME ANCHORS CAN CERTIFY

Proposition 2. Let a rule read only the anchors of a set N, and let $\Theta(D)$ contain two candidates that record the same action at every anchor of N while their safe sets are disjoint. Then no rule reading only N returns an action in $\bigcap_{\theta\in\Theta(D)} S_\epsilon(\theta)$: whatever it returns lies outside the safe set of at least one of the two.

Proof. A rule that reads only N is a function of the anchors in N. The two candidates record the same action there, so the rule returns the same action for both, and since the two safe sets are disjoint an action in one lies outside the other. The returned action is therefore outside at least one safe set in the candidate family, and the intersection is not certified. □

The proposition characterizes what the anchors decide. A rule that reads a set N is a function of it, so the intersection is certifiable from N only when no two candidates the anchors treat alike disagree on safety. That is what the budget moves: reading anchors that separate the two candidates removes the case, which is why the sweep is the axis the evaluation reports. The anchor-safety condition (A4) supplies the case outright, since it says that the action an anchor records is safe for every candidate consistent with the evidence, and it is the premise that turns agreement into a certificate and the premise the inheritance proposition states. The evaluation's part is to measure the distance, which Section 5 does for the locality the inheritance also needs, and which the finite family does exactly, where the candidate set is enumerated and the intersection is computable.

E.3 REFERENCE-ACTION MISMATCH IS A SELECTION TERM AGAINST A UNIFORM DRAW

Proposition 3. Fix an issued state x with candidate message set $C(x)$, and let $r(x) = |\{m \in C(x) :$ the action of m is the reference action at $x\}|$. A gate that installs one message of $C(x)$ mismatches the reference action at x exactly when the message it installs lies outside those $r(x)$. Over the issued states of a pool the reference-action mismatch rate is therefore $1 - \mathbb{E}[r(x)/|C(x)|] - \Delta$, where the first term is the error a rule drawing uniformly from the state's own candidate set would make and Δ is the advantage the gate's own rule carries over that draw. The candidate set is the state's own, so the floor carries no assumption that the alphabet size is fixed.

Proof. The first statement is the definition of the two sets: reference-match counts the reference action, and $r(x)$ counts the candidates that reproduce it. For the second, write the mismatch at x as $1 - \mathbf{1}\{m_x \in R(x)\}$, where $R(x)$ is the set of those $r(x)$ messages and m_x is the message the gate installs. A uniform draw from $C(x)$ places m_x in $R(x)$ with probability $r(x)/|C(x)|$, so its expected error is $1 - \mathbb{E}[r(x)/|C(x)|]$; subtracting the gate's expected error from that leaves Δ by definition. □

The two terms separate what the state offers from what the gate does with it, and the budget moves the second. Taking the candidate set to be the full alphabet, the floor is 0.3902 at the declared

budget, 0.3857 at sixteen anchors and 0.3917 at thirty-two, while the reference-action mismatch rate is 0.3930, 0.3013 and 0.3106: the gate's advantage over a uniform draw runs from -0.0028 at four anchors to $+0.0844$ at sixteen, so it is the first eight anchors that make the selection informative, and the floor is nearly flat across the whole sweep. Taking it to be the empirical message-consistency set $C_D(g, m)$ instead, the floor and the advantage are recomputed in Appendix B.4. No pool in the audit has a mean r below one, so no pool's error is explained by the absence of a repairing message; every pool's error is a question of which of the candidates the gate installs.

E.4 The statistic is the set of messages the gate can install

Proposition. Fix a bucket B of calibration anchors and let $C(B) = \{c_1, \ldots, c_s\}$ be the distinct messages its anchors chose, so $s = |C(B)|$ is the public statistic. Let $\mathcal{R}(x)$ be the candidate messages that reproduce the reference action at x.

(i) *Support.* On every gate-issued state of B the gate installs a message of $C(B)$: it returns the common chosen value over $N(x) \subseteq B$, and that value lies in $C(B)$ by definition. The statistic is therefore the size of the gate's output support on the bucket, and it is the smallest pool-level quantity that describes the set of messages the gate can choose from.

(ii) *No choice at one.* If $s = 1$ the gate has a single message available, so $\Delta(B) = 0$ and the reference-action mismatch rate on B is that message's own mismatch rate. Every departure of Δ from zero requires $s \geq 2$.

(iii) *The two axes.* With p_i the probability that the gate installs c_i on a gate-issued state of B,

$$\text{reference-match}(B) = \sum_{i=1}^{s} p_i \, \Pr[c_i \in \mathcal{R}(x) \mid m_x = c_i] \,,$$

$$\Delta(B) = \text{reference-match}(B) - \frac{1}{s} \sum_{i=1}^{s} \Pr[c_i \in \mathcal{R}(x)] \,.$$

The error depends on the bucket and on the state only through $C(B) \cap \mathcal{R}(x)$: the bucket fixes that set, the state fixes how much of it is right, and no function of the bucket alone determines the error.

Proof. (i) is the gate's third step: it returns the common $a.\text{chosen}$ over $N(x)$, and $N(x) \subseteq B$, so that value is one of the messages the bucket's anchors chose. (ii) follows from (i) at $s = 1$, where the installed message is c_1 on every gate-issued state of B, so reference-match$(B) = \Pr[c_1 \in \mathcal{R}(x)] = (1/s) \sum_i \Pr[c_i \in \mathcal{R}(x)]$ and $\Delta(B) = 0$ by the expression in (iii). (iii) is the law of total probability over the message the gate installs, and its second expression is Δ of the preceding subsection read on the bucket's that set. □

The statistic and the state's over-determination are independent because they are properties of different objects, and the evaluation reports both readings. Binning by the statistic averages the state term away, which is why the bin means order the risk across nine levels while a single pool's rank correlation with its own risk is weak, and why the two statistics correlate with each other by only -0.2094, -0.1147 and $+0.0294$ across the three budgets. The lowest bin is the sharpest case of (ii): a bucket whose anchors agree offers the gate one message, the gate can do neither better nor worse than that message, and the recorded rate there is 0.0948.

E.5 The public-interface quotient

Let the message alphabet be $X = \{1, \ldots, d\}$ and let a sender-receiver transform be a row-stochastic matrix T whose rows have a unique maximum under a declared tie break. A public anchor query chooses one source x and observes the wire symbol

$$w_T(x) = \arg\max_y \, (Te_x)_y,$$

and the interior probabilities of row x are not observed. For a query set $Q \subseteq X$ write $O_Q(T) = (w_T(x))_{x \in Q}$.

Observational equivalence. For every query set Q, two transforms of the family are observationally equivalent exactly when they agree on the restricted interface $w_T|_Q = w_U|_Q$. Proof: the observation is a deterministic function of the selected rows and depends on them only through their maxima, so equal restricted interfaces give every anchor response the same law and no rule separates the two; if the restricted interfaces differ, an anchor at a differing source separates them. The identifiable object is therefore the class $[T]_Q = \{U : O_Q(U) = O_Q(T)\}$, and two transforms with a common interface on all of X cannot be separated even with unlimited anchors.

Exact criterion. Write $W(F) = \{w_T : T \in F\}$ for the family's interface maps. The interface is identifiable from Q exactly when O_Q is injective on $W(F)$, that is when $w_T|_Q \neq w_U|_Q$ for every distinct pair. When the criterion fails the audit returns the quotient $W_Q(F) = \{w_T|_Q\}$ or abstains.

Query cost is a hitting set. For two distinct interface maps let $D(w, w') = \{x : w(x) \neq w'(x)\}$ be their disagreement set. A query set identifies the interface exactly when it intersects $D(w, w')$ for every distinct pair, so the minimum nonadaptive cost is

$$q_{\mathrm{na}}(F) = \min\big\{|Q| : Q \cap D(w, w') \neq \emptyset \text{ for every distinct pair}\big\},$$

with $\lceil \log_d |W(F)| \rceil \leq q_{\mathrm{na}}(F) \leq d$, the upper bound being the query of every source. The adaptive analogue replaces the hitting set by the minimum depth of a decision tree whose leaves are quotient classes. Minimum hitting set is NP-hard in general. The finite catalogs below expose the exact target-induced query frontiers.

The sharp classical case. For the family of all bijections on X, the unobserved source-target pairs can be matched in any way, so m distinct anchor responses leave exactly $(d - m)!$ consistent interfaces; $m = d - 1$ leaves one, $m = d - 2$ leaves two, and $d - 1$ is optimal in the worst case Ko & Teng (1986). This is the ladder that prices the saturation of the budget sweep in Appendix D, and its counting core is the classical permutation query complexity.

E.6 A DECISION COSTS NO MORE THAN AN IDENTITY, AND CAN COST LESS

Let $q : \Theta \to Z$ be a target, such as a repair action, a safe-action set, or a version label, and let $P_q = \{\{\theta, \theta'\} : q(\theta) \neq q(\theta')\}$ be the pairs whose target values differ. A query set certifies q exactly when it intersects $D(\theta, \theta')$ for every pair in P_q, so the exact nonadaptive target cost is the same minimum over a smaller constraint set:

$$Q^\star_q = \min\big\{|Q| : Q \cap D(\theta, \theta') \neq \emptyset \text{ for every pair in } P_q\big\}, \qquad Q^\star_q \leq Q^\star_{\mathrm{identity}},$$

The inequality holds because P_q is a subset of the distinct pairs. On full permutation families, a target supported on $k < d$ coordinates costs k queries, while full identity costs $d - 1$. If a target-relevant pair has an empty disagreement set the target is not identifiable at any budget; if P_q is empty the cost is zero. The finite validator checks the $d = 3$ instance exhaustively, where the target taken as the image of one source costs one query against the identity's two.

The repair-relevant target has two readings and they price differently. Taking $q(\theta) = S_\epsilon(\theta)$, the set of actions that are ϵ-safe for θ, asks the query set to recover the whole set. Returning one safe action asks less: a query set certifies this decision when every transcript cell has a nonempty safe-action intersection. The exact cost of this set-valued decision is governed by the minimal higher-order conflicts characterized in Appendix E.9; pairwise separation gives the exact cost when the safe-action family is 2-Helly. Since recovering the full safe-action set also certifies a shared action, and identifying the interface recovers that set,

$$Q^\star_{\mathrm{dec}} \;\leq\; Q^\star_q \;\leq\; Q^\star_{\mathrm{identity}},$$

The middle inequality can be strict. In a two-interface catalog with safe-action sets $\{a, b\}$ and $\{a, c\}$ and a probe that distinguishes the interfaces, action a is certifiable without a query, while one query recovers the full safe-action set. Thus $Q^\star_{\mathrm{dec}} = 0$ and $Q^\star_q = Q^\star_{\mathrm{identity}} = 1$ in this catalog. The ϵ-repair quotient groups interfaces with identical safe-action sets and preserves every safe-action membership query. The measured three-level cost hierarchy is reported in Appendix J.

E.7 THE REPAIR QUOTIENT IS THE COARSEST SAFE ABSTRACTION

The body says the ϵ-repair quotient is the coarsest abstraction that preserves every safe-action membership query. Both halves of that claim are proved here, since a coarsest element has to be shown to exist as well as to preserve.

Write $S_\epsilon(\theta)$ for the ϵ-safe actions of interface θ, and define the repair relation by

$$\theta \sim_\epsilon \theta' \quad \Longleftrightarrow \quad S_\epsilon(\theta) = S_\epsilon(\theta'),$$

with $\Theta/\!\sim_\epsilon$ the repair quotient.

Proposition. $\sim_\epsilon$ is an equivalence relation, and it is the coarsest relation on Θ with the property that two states related by it answer every safe-action membership query alike.

Proof. Equality of sets is reflexive, symmetric and transitive, so $\sim_\epsilon$ is an equivalence relation and the quotient is well defined.

For the property, let $\theta \sim_\epsilon \theta'$ and let a be any action. Since the safe sets are equal, $a \in S_\epsilon(\theta)$ if and only if $a \in S_\epsilon(\theta')$, so the two states answer the membership query for a alike, and this holds for every a.

For coarseness, let R be any equivalence relation with the property, and suppose $\theta R \theta'$. If the property failed to imply $S_\epsilon(\theta) = S_\epsilon(\theta')$ there would be an action a in exactly one of the two sets, and the two states would answer the membership query for a differently, contradicting the property. So $\theta R \theta'$ implies $\theta \sim_\epsilon \theta'$, hence $R \subseteq \sim_\epsilon$, and $\sim_\epsilon$ is the coarsest such relation. □

Sharpness. The quotient is strictly coarser than behavioral equivalence whenever two behaviorally distinct interfaces share their safe-action sets. The permutation families the validator checks are the smallest instance: a full permutation family is behaviorally separated at every coordinate, while its members' safe sets can coincide on the target's support, which is why the identity cost exceeds the target cost on those instances and the two coincide on families where the target is injective.

E.8 THE LOCAL GATE INHERITS THE CERTIFICATE

The body states the inheritance in Section 4; it is repeated here with its proof. Write $S_\epsilon(\theta, x)$ for the actions that are ϵ-safe for interface θ at receiver context x, and $\Theta(D)$ for the interfaces consistent with the public transcript D.

Proposition 4. Fix a state with receiver context x and sender message m, and let $a_1, \dots, a_k$ be the target actions recorded at the k nearest anchors carrying message m, with contexts $x_1, \dots, x_k$. Suppose

- (A1) *local stability:* there is a radius $\delta > 0$ such that for every $\theta \in \Theta$ and every pair x, x' with $\|x - x'\| \le \delta$, the safe sets agree, $S_\epsilon(\theta, x) = S_\epsilon(\theta, x')$;
- (A2) *the neighbours are inside the radius:* $\|x - x_i\| \le \delta$ for every $i \le k$;
- (A3) *the gate holds:* the recorded actions agree, $a_1 = \cdots = a_k =: a$, and the recorded messages agree;
- (A4) *the recorded action is certified at the anchors:* $a \in \bigcap_{\theta \in \Theta(D)} S_\epsilon(\theta, x_i)$ for every $i \le k$.

Then $a \in \bigcap_{\theta \in \Theta(D)} S_\epsilon(\theta, x)$, so the action the local gate installs at x is certified safe for every interface the transcript leaves open.

Proof. By (A2) every neighbour lies within δ of x, so (A1) gives $S_\epsilon(\theta, x_i) = S_\epsilon(\theta, x)$ for every θ and every $i \le k$. Hence

$$\bigcap_{\theta \in \Theta(D)} S_\epsilon(\theta, x_i) = \bigcap_{\theta \in \Theta(D)} S_\epsilon(\theta, x) \qquad \text{for every } i \le k.$$

By (A4) the action a lies in the left-hand intersection, so it lies in the right-hand one, which is the claim. □

What each premise costs. The four premises are not of the same kind, and separating them is the point of stating the proposition.

(A3) is enforced by the rule: the deployed procedure abstains unless the neighbours agree, so a state on which it acts satisfies the premise by construction, and the abstention rate is the share of states where it does not.

(A2) is measured. The distance to the k-th neighbour is recorded for every state, so the share of accepted states whose neighbour radius stays under any candidate δ is measured directly, and the body reports the neighbour distances alongside the accepted rate.

(A1) and (A4) are the conditions Proposition 4 carries, and neither follows from the other two. (A4) is the substantive one: it says the action recorded at an anchor is safe for every interface the full transcript admits, which is the certificate's condition applied at the anchor. (A1) is the geometric one: it says the safe-action set does not change within the radius, so the neighbour's evidence transfers to the state.

The evaluation therefore bounds (A2), enforces (A3), and reports the fidelity of the rule it audits, which the audit of Appendix B.4 reads against the candidate set. A state whose neighbours disagree is left to abstention, and the reference-action mismatch rate is the share of issued states on which the installed action departs from the receiver's own policy, which is a fidelity quantity.

E.9 The exact cost of a shared-action certificate

The exact set-valued decision object is a conflict hypergraph. Let X be a finite probe set, $\rho_\theta : X \to \mathcal{Y}$ the deterministic response of candidate interface θ, and $S_\epsilon(\theta)$ its nonempty safe-action set. For a nonadaptive query set Q and true candidate $\theta_\star$, the surviving cell is

$$C_Q(\theta_\star) = \{\theta \in \Theta : \rho_\theta|_Q = \rho_{\theta_\star}|_Q\}.$$

Let $\mathcal{M}_\epsilon$ be the inclusion-minimal empty-intersection families,

$$\mathcal{M}_\epsilon = \left\{ A \subseteq \Theta : \bigcap_{\theta\in A} S_\epsilon(\theta) = \emptyset, \quad \bigcap_{\theta\in B} S_\epsilon(\theta) \neq \emptyset \ \forall B \subseteq A,\ B \neq A \right\}.$$

For each such family define its separating probe edge

$$E(A) = \{x \in X : \exists \theta, \theta' \in A, \rho_\theta(x) \neq \rho_{\theta'}(x)\}.$$

Theorem (minimal-conflict query cost). The exact nonadaptive cost for certifying a shared safe action is

$$Q^\star_{\mathrm{dec}} = \min_{Q\subseteq X}\{|Q| : Q \cap E(A) \neq \emptyset \text{ for every } A \in \mathcal{M}_\epsilon\}.$$

If some $E(A)$ is empty, no finite query set certifies the decision for that family. If $\mathcal{M}_\epsilon$ is empty, the cost is zero.

Proof. Suppose Q hits every $E(A)$ and consider any surviving cell $C_Q(\theta_\star)$. If its safe-action intersection were empty, it would contain an inclusion-minimal empty-intersection family A. Every member of A agrees with $\theta_\star$ on Q, so $Q \cap E(A)$ would be empty, which contradicts the hitting property. Thus every surviving cell has a common safe action. Conversely, if Q misses $E(A)$ for some A, all members of A agree on Q. For any $\theta_\star \in A$, the surviving cell contains A and has empty safe-action intersection, so Q cannot certify a shared action. The two implications give the equality. □

Pairwise boundary and the non-Helly witness. Let $\mathcal{M}^{(2)}_\epsilon$ contain the minimal conflicts of size two. Its probe edges give

$$\tau(\mathcal{H}^{(2)}_\epsilon) \leq Q^\star_{\mathrm{dec}},$$

and equality holds when the safe-action family is 2-Helly, so every minimal empty intersection is a pair. The sets $\{a, b\}$, $\{b, c\}$, and $\{a, c\}$ have pairwise nonempty intersections and an empty triple intersection. They produce no pairwise edge and one higher-order edge, which is the exact finite witness used by the noisy certificate experiment. Its enumerated costs are one query for a selected action, one for a target symbol, and $d - 1$ for full permutation identity.

E.10 WHERE THE CERTIFICATE IS NON-VACUOUS

What a certificate can say at a state is bounded by whether the transcript reaches it. Take the family of receiver maps that agree with the recorded anchors, and let x be a state whose (context, message) pairs the anchors do not probe. Two members of that family can differ at x in any way, so for any message m there is a consistent map under which installing m at x reproduces the reference action and another under which it does not. The safe sets of those two maps at x intersect emptily, and $\mathcal{S}_\epsilon(D)$ at x is empty with them. The certificate is therefore non-vacuous at the states the anchors reach, which is the family the finite statements are computed on, and the message-level object of Section 3 is the reading the audit builds at every state.

E.11 A TIME-UNIFORM CONFIDENCE SET

The noisy experiment uses a row-wise confidence set for a finite permutation family. Let m queried source rows have true symbols j_s, and let each response pass through a known d-ary replacement channel with correct probability p and wrong-symbol probability $q = (1-p)/(d-1)$, where $p > q > 0$. For row s, let $N_{s,i}(n)$ count response symbol i after n observations. The implemented candidate set is

$$\widehat{C}_{s,n} = \{j : \Delta_{s,j}(n) < b_{m,d,\delta}(n)\},$$
$$\Delta_{s,j}(n) = \max_i \big(N_{s,i}(n) - N_{s,j}(n)\big) \log \frac{p}{q},$$
$$b_{m,d,\delta}(n) = \log\left(\max\left\{2, \frac{m(d-1)n(n+1)}{\delta}\right\}\right).$$

Proposition. If row selection is predictable from the preceding transcript, then

$$\Pr\Big(j_s \in \widehat{C}_{s,n} \text{ for every audited row } s \text{ and every } n \geq 1\Big) \geq 1 - \delta.$$

For a fixed competitor $i \neq j_s$, the likelihood ratio $L_{s,i}(n)/L_{s,j_s}(n)$ is a nonnegative martingale under the true row. Markov's inequality at each n, followed by the summable allocation $\delta/[m(d-1)n(n+1)]$ and a union bound over rows and competitors, proves the result because $\sum_{n\geq 1} 1/[n(n+1)] = 1$. Predictable selection preserves the conditional likelihood-ratio calculation. The deterministic $p = 1$ branch uses the exact observed-symbol set.

At time t, let $N_s(t)$ be the number of observations collected from source row s. The row sets induce the interface confidence family

$$\widehat{\Theta}_t = \Big\{\theta \in \Theta : \rho_\theta(x_s) \in \widehat{C}_{s,N_s(t)} \text{ for every audited row } s\Big\},$$

where an unqueried row retains its full symbol set. Thus the proposition gives $\Pr(\theta_\star \in \widehat{\Theta}_t \text{ for every } t) \geq 1 - \delta$.

Proof of the stopping claim. On this event, $\bigcap_{\theta\in\widehat{\Theta}_t} S_\epsilon(\theta) \subseteq S_\epsilon(\theta_\star)$ for every t. The event that a returned action at any stopping time falls outside the true safe set is therefore contained in the complement of the confidence event, which has probability at most δ.

For $p < 1$, an explicit finite-sample reading follows from the same construction. Let $\gamma = p - q$, $L_\alpha = \log(m(d-1)/\alpha)$, and $h_n = b_{m,d,\delta}(n)/\log(p/q)$. If

$$n\gamma - \sqrt{2nL_\alpha} \;\geq\; h_n,$$

then, with probability at least $1-\alpha$, every wrong symbol is excluded from every audited row at sample size n. Indeed, for each row and competitor, the variables $\mathbf{1}\{Y = j_s\} - \mathbf{1}\{Y = i\}$ have mean γ and range length two, so Hoeffding's inequality gives $N_{s,j_s}(n) - N_{s,i}(n) \geq h_n$ simultaneously after a union bound; the defining likelihood score then removes i from $\widehat{C}_{s,n}$. This is a sufficient identity-resolution scale. Decision certification can stop as soon as the shared-action intersection becomes nonempty.

At sample size n, an action in $\bigcap_{\theta\in\widehat{\Theta}_n} S_\epsilon(\theta)$ is certified; an empty intersection returns abstention or requests a probe that separates a minimal conflict family.

Experiment grids. The confirmation grid spans widths $d \in \{3, 4, 8, 16\}$ and reliabilities $r \in \{0.55, 0.70, 0.85, 0.95, 1.00\}$, with 10,000 trials per cell, $\delta = 0.05$, and a 128-query cap per source row. The boundary grid spans widths 3, 4, and 8 and six normalized signal levels, including $\eta = 0$ and $\eta = 0.25$, with 10,000 paired trials per cell and the same per-source cap. An independent raw audit recomputes candidate sets, stopping times, pathwise coverage, unsafe-return events, and paired savings; the savings intervals use a normal approximation.

F DECISION IDENTIFIABILITY AND QUERY COST

The target-factorization criterion extends the public-interface quotient to a deployment objective. If a loss or repair action is constant on each behavioral class, all attainable risk values and optimal decisions depend only on the quotient. For a finite family, a nonadaptive query set identifies a target exactly when it intersects every disagreement set between behaviors whose target values differ.

The first statement follows because behaviorally equivalent states induce the same likelihood for every transcript and every policy. A target that differs within an equivalence class therefore cannot be identified; if the target is constant on each class, identifying the quotient determines it. For the query statement, a pair with different targets that agrees on every queried coordinate produces the same observation with different target values and is impossible to separate. If every such pair disagrees on at least one queried coordinate, the target is constant on every observationally remaining equivalence class, so the target is identified.

Width	Full interface	First source	Parity	Label $K = d$	Label $K = d + 2$
$d = 3$	2	1	2	1	impossible
$d = 4$	3	1	3	1	impossible
$d = 5$	4	1	4	1	impossible

Table 21: Exact noiseless nonadaptive query costs on full permutation families and cyclic version histories. A decision target can be cheaper than the full quotient.

Let $q_x(\pi) = \pi(x)$ and consider the symmetric replacement channel with $r > 1/d$. Fano's inequality on the d uniformly weighted target values gives

$$T \geq \frac{(1-\delta)\log d - \log 2}{C(r,d)},$$

where

$$C(r,d) = \max\{0, \log d - H_b(r) - (1-r)\log(d-1)\}.$$

The capacity grows with $\log d$, and the pairwise discrimination exponent relevant to a likelihood certificate is a different quantity. For two distinct labels, the Chernoff information is

$$\kappa(r,d) = -\log\left(2\sqrt{ra} + (d-2)a\right), \qquad a = \frac{1-r}{d-1},$$

which approaches $-\log(1-r)$ at fixed reliability. A coordinate-wise union bound therefore depends on the number of competing labels, while the total cost also depends on how many target coordinates must be certified. These terms are separated in the finite frontier audit.

For $q_S(\pi) = (\pi(x))_{x \in S}$ with $|S| = k$, the exact noiseless cost is k when $k < d$ and $d-1$ for the full bijection interface. Under noise, the target and full-interface procedures use the same replacement channel and likelihood decoder but allocate probes to different disagreement sets. The decision-safe audit covers 40 target versus full frontiers across widths, support sizes, reliabilities, and repeat budgets; every observed target frontier is cheaper. The high-width full-interface audit covers 160 cells from $d = 64$ to 512 with zero recorded bound failures. These are finite-family results, and their exact costs are reported as such.

Decision-coordinate packing. The support result is one instance of a geometry-dependent cost. Suppose the target decomposes into r independent binary coordinates $Z_1, \ldots, Z_r$, and the available queries are partitioned into disjoint blocks, where every query in block i observes a binary channel for Z_i with Chernoff information at most κ_i. If the full target error is at most δ, then the marginal error of each coordinate is also at most δ. The standard binary Chernoff exponent gives

$$T_i \geq (1 - o(1))\frac{\log(1/\delta)}{\kappa_i},$$

so the leading coefficient is at least $\sum_i 1/\kappa_i$. Conversely, when each block contains a query with information at least κ_i, repeating it $\lceil \log(r/(2\delta))/\kappa_i \rceil$ times and applying a likelihood-ratio test makes each coordinate error at most δ/r. A union bound bounds the total query count by

$$T \le \sum_{i=1}^{r} \left\lceil \frac{\log(r/(2\delta))}{\kappa_i} \right\rceil .$$

For fixed r, the lower and upper bounds match in the $\log(1/\delta)$ factor. An exact audit over 72 cells, coordinate counts $r \in \{1, 2, 3, 4, 6, 8\}$, channel noises from 0.10 to 0.40, and errors from 0.01 to 0.10 finds no upper-budget or finite-exponent violation. The result shows that target geometry is what controls query cost, beyond what entropy alone determines; overlapping queries and non-product targets remain outside this theorem.

Decision versus label on the same transcripts. We evaluate a version-label predictor and a behavior-class predictor on identical noisy transcripts. The comparison spans 160 finite cells, with 140 permanent label-alias cells. Across those cells, label accuracy ranges from 0.129 to 0.703, while behavior-class accuracy ranges from 0.760 to 1.000; the largest decision-minus-label gap is 0.871. A matched-budget noise audit concentrates observations on a single relevant source and improves accuracy by up to 0.999 over full-interface recovery at the same total budget. The exact decision-aware costs and noisy rates are given in Appendix F.

For the symmetric replacement channel, let $a = (1-r)/(d-1)$. The pairwise Chernoff information between two distinct labels is

$$\kappa(r, d) = -\log\left(2\sqrt{ra} + (d-2)a\right).$$

At fixed reliability, $\kappa(r, d)$ approaches $-\log(1-r)$ as d grows. The channel therefore does not provide a logarithmic width gain for pairwise label discrimination. We report noisy results as finite certification frontiers under this explicit channel. The target auditor concentrates probes on the coordinates that can change the repair, while the full-interface auditor certifies every coordinate. This keeps the comparison tied to the decision target and to the observed finite budget.

The exact noiseless cost for $q_S(\pi) = (\pi(x))_{x \in S}$ is k when $k < d$ and $d - 1$ for the full bijection interface. The finite noisy audit measures how the corresponding certification frontier changes with d, k, reliability, and the error target. A lower-entropy target can be cheaper than its support mapping, while a parity target can retain the full-interface cost.

G ADVICE-ASSISTED AUTOMATA COMPARISON

Fica and Otop's advice-enhanced active automata learning framework Fica & Otop (2025) uses rewriting advice to infer membership-query answers and reduce the cost of learning a language-equivalent automaton. Its learning target remains a complete automaton object. Our certificate changes the stopping object: the auditor stops when all interfaces consistent with the transcript share an installable repair action. The two ideas are compatible, since advice can reduce the cost of collecting evidence while the target-specific stopping rule determines when the evidence is sufficient for the decision.

We use advice-enhanced learning as the identity-level reference point for query reduction. The matched comparison below uses its query oracle and finite catalog, then adds the target-specific stopping rule. Complete language equivalence and decision certification remain distinct identification objectives.

The comparison uses that matched finite-catalog instance. We lock the public implementation at commit 197f51bccf2f92c979566e98b89b9c6b74fa3ce3 and enumerate the 40 commuting two-state DFA targets induced by the rule $ab \to ba$. The public arms execute the released implementations; the two decision-stop arms use the same teacher and add the target-specific stopping rule. Across all 640 rows, the catalog is complete, the advice and certificate outputs satisfy their stated conditions, and the query accounting agrees with the oracle.

The complete-objective arms finish language-equivalent hypotheses with mean teacher-query costs of 9.2 without advice and 7.9 with advice. Target-aware stopping reaches the selected decision

Decision target	Public L*	Advice L*	Stop	Advice + stop	Cert.
Language acceptance	9.2	7.9	3.1	3.1	100%
Initial transition	9.2	7.9	6.5	5.8	30%
Accepting-state count	9.2	7.9	6.3	5.6	30%
Transition signature	9.2	7.9	7.1	6.3	30%

Table 22: Mean teacher queries over the 40-target catalog. Certification rates are for the combined advice and target-stop configuration.

with 5.75 queries without advice and 5.20 with advice, averaged over the four target families. The target-stop arms retain multiple candidates at certification, which is the empirical separation between complete automaton learning and decision certification. The official BITWISE_ADDITION run gives the corresponding larger-instance comparison: 202 teacher queries without advice and 116 with the authors' advice configuration, with language-equivalent three-state hypotheses in both runs.

Advice reduces query cost under its rewriting scheme. Target-specific stopping ends once the selected decision is identified, while the automaton's full language behavior can remain unresolved.

The released-version MCP evidence includes seven filesystem versions, nine sequential-thinking versions, ten memory versions, and 28 everything versions. The filesystem history has three behavior classes and a label ceiling of 0.429. The sequential-thinking history has three behavior classes, two repair classes, and a label ceiling of 0.333; class-conditioned repair changes held-out workflow success from 0.222 to 1.000. The memory history has three behavior classes, one repair class, and a label ceiling of 0.300; its class-conditioned compatibility layer changes held-out workflow success from 0.300 to 1.000. The everything history has ten behavior classes, two repair classes, and a label ceiling of 0.357; its tool-name and schema repair changes held-out sum success from 0.821 to 1.000. These released package histories support the label-class separation at protocol and task levels. They provide versioned protocol histories for the decision analysis.

The frozen workflow suite uses one canonical plan-sum-memory topology and two held-out alternatives, reverse sum-plan-memory and plan-two-sums-memory. It contains 270 class-workflow units and four policies per unit. Class repair and class-blind fallback both recover every unit; class repair uses 1170 total interactions including probes, whereas fallback uses 1369. The held-out topologies account for a reduction of 141 calls. The paired class-combination bootstrap interval for the interaction-cost difference remains below zero.

Setting	Versions	Behav.	Repair	Label cap	No rep.	Wrong	Class
Seq.	9	3	2	0.333	0.222	0.222	1.000
Mem.	10	3	1	0.300	0.300	0.300	1.000
Everything	28	10	2	0.357	0.821	0.000	1.000

Table 23: Released MCP task histories; Seq., Mem., and Everything denote sequential-thinking, memory, and everything. One public probe identifies the behavior class in each history, and class-conditioned repair restores the held-out workflow.

The decision audit evaluates label and behavior-class predictors on the same transcripts. It contains 160 finite cells, 140 with permanent label aliasing. Label accuracy is 0.129 to 0.703, behavior-class accuracy is 0.760 to 1.000, and the largest class-minus-label gap is 0.871. The matched-budget table reports representative accuracy cells. The separate cost comparison reaches the three certificate levels at matched decision quality, as reported in Section 5.

The decision comparison uses 12 trained seeds per width and 200 evaluation episodes. The version-label rule performs maximum-likelihood version identification and abstains when aliases remain. The behavior-class rule identifies the public behavior class and maps it to a representative version adapter. Every comparison in Table 2 passes the specified recovery, no-repair, and label contrasts.

H Interface budget and finite-sample law

Reference-action mismatch is governed by how many distinct messages the context bucket leaves, and Figure 8 shows it on the training and held-out splits with 95 percent binomial intervals, together with the held-out error of the five candidate statistics against the declared tolerance. The anchor

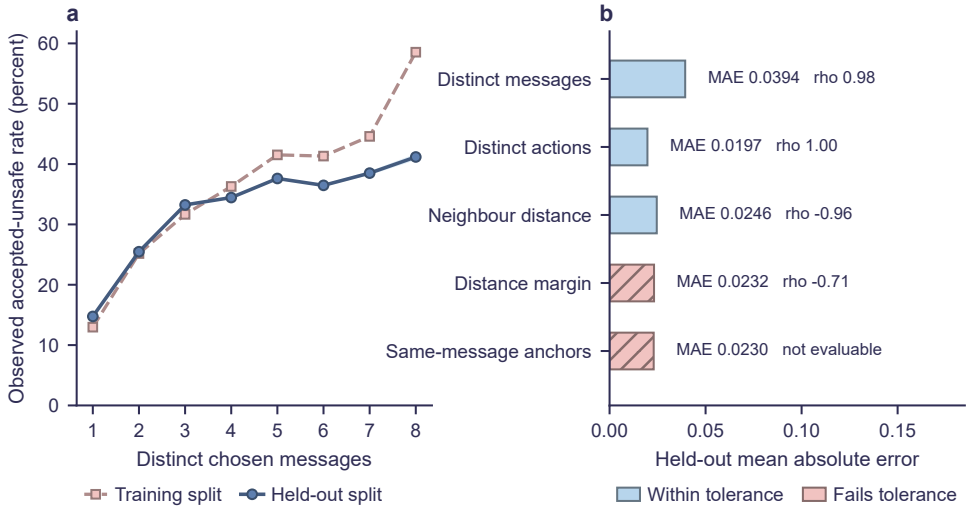

Figure 8: Reference-action mismatch is predictable before the decision. Observed mismatch against the number of distinct chosen messages in the context bucket (left), and the held-out error of five candidate statistics with the declared tolerance (right).

ladder that places the turning point of the budget sweep, and the frontier comparison that separates a decision from an identity, are stated and tabulated in Appendix D.

Supporting interface evidence. On the referential task, a public anchor reveals the receiver's selected wire symbol $w_T(x) = \arg\max_y (Te_x)_y$. For a bijective interface, m anchors leave $(d-m)!$ consistent maps, which yields a $d-1$ identification budget in the classical permutation-identification setting Ko & Teng (1986). Across four admissible transform families, widths through $d = 32$, and trained finite-task checkpoints, the same budget identifies the behavior and a receiver-side adapter restores the matched wire endpoint from 0.05 to 1.00; wrong and random controls remain below 0.07. The detailed family and finite-sample results are in Appendix H. What that behavioral object has to distinguish is settled by the decision it has to support, which is what the experiments measure.

I IMPLICATIONS

Cost follows the target. The decision-level view also predicts the shape of the cost curve, and the prediction holds where the target restricts the goal set: certifying a selected action costs 1.022 queries per state, a repair signature 1.160, and a full identity 1.480, with the cheaper level reaching the same decision quality. The signature level returns the same certified decisions as the selected-action level at strictly higher cost, so its evidence buys the membership guarantee.

Action safety is a state property. Requiring the class target action to belong to the state-wise safe set cuts issue coverage from 0.557 to 0.267 while reference-action mismatch stays at 0.5246 against 0.5281. Safety is a property of the state, so a constraint written on classes cannot deliver it. This is the observation that motivates carrying utility uncertainty in the candidate state.

The procedure. Define the repair target before collecting probes, maintain the compatible interface set, and allocate probes to the disagreements that can change the repair. Read the error rate off the disagreement statistics and choose a risk level. When the compatible states share no safe action, abstain or acquire another anchor.

J HIERARCHY AND STATISTIC DETAIL

J.1 WHAT THE PUBLIC STATISTIC MEASURES

The statistic that predicts reference-action mismatch is a property of the pool: the recorded binning rule is exact-integer with nine bins, which is more than the three neighbours the gate reads, so it counts the distinct chosen messages in the whole same-message bucket. The states that pool is applied to carry a property of their own, the number of candidate messages that reproduce the native action, which bounds the gate's reference-action mismatch however the pool is drawn. The two are nearly independent, so the question is whether the statistic still orders mismatch once the states' over-determination is held fixed.

Both readings are real and they are distinct readings. Pooled into the statistic's own bins and weighted by gate-issued states, reference-action mismatch rises across the nine bins at thirty-two anchors, from 0.0948 at one distinct chosen message to 0.5349 at nine, and across the seven bins the smaller budgets reach. Read across pools, the statistic's rank correlation with the pool's mismatch is $+0.3070$ at thirty-two anchors and $+0.1744$ at sixteen, and at eight it is -0.0079 while the bin means still rise from 0.3079 to 0.4091. The states' mean over-determination correlates with the pool's mismatch by -0.4248, -0.5154 and -0.6800 at those budgets, and the two statistics correlate with each other by only -0.2094, -0.1147 and $+0.0294$. Holding the states' over-determination fixed in thirds at thirty-two anchors leaves the statistic's correlation at $+0.3361$ in the lowest third, $+0.0885$ in the middle and $+0.2853$ in the highest.

The statistic therefore prices the pool and the states' over-determination prices the state, and gate mismatch reflects both. That is also why the bin table orders as sharply as it does while a single pool is a weak predictor: the bins average over states, and the states' own term is what the average absorbs. Reading the statistic per pool is the more conservative of the two for a single pair.

Three geometric dispersion statistics, the k-th neighbour maximum distance, the k-th to k+1-th distance margin, and the same-message anchor count, do not calibrate: their rank correlation with the bin index is negative or undefined, although their mean absolute error can still fall inside the tolerance. The split identifies what carries the information. The signal is the disagreement structure of the bucket, a discrete count.

The hierarchy's levels answer separate questions. Requiring the class target action to belong to the state-wise safe set cuts issue coverage from 0.557 to 0.267, while reference-action mismatch stays at 0.5246 against 0.5281. Safety is a property of the state, so utility uncertainty has to be carried in the candidate state.

The structure of the 24 inconclusive contrasts is itself informative. They occur where the signature and full-identity anchor sets coincide by construction, which is the case when the target scope is not a strict subset of the goal set, so the hierarchy is informative exactly where the target restricts the goal set.

K EVIDENCE BASE

The controlled matrix fixes four cases by construction and reproduces all eighty of its rows with no mismatch; the natural stratification partitions the same evaluation states into $1{,}440{,}000$ grouped states whose certificate counts equal the recorded cells item by item: $485{,}342$ certified decisions reproduce the declared target action, $314{,}911$ do not, and $150{,}046$ force abstention because the compatible candidates require incompatible actions. Separately, the risk ladder is monotone in its discrete statistic across the whole range: reference-action mismatch rises from 0.1396 at one distinct chosen message to 0.5012 at nine, a factor of 3.6. The three geometric dispersion statistics do not track it, with a rank correlation of -0.9636 against the bin index, although their mean absolute error can still fall inside the declared tolerance. The discrete count carries the signal and the metric geometry of the neighbourhood does not.

L Related work, extended

The repair literature, work by work. The main text groups these by the object they change; the per-work contrast is given here. Safe-by-Repair assumes a two-level lattice controller and splits the problem into a local feasibility problem at the counterexample and a global activation problem, both convex Cruz et al. (2022). SpecRepair turns a formal specification into an objective that is negative on violating inputs, searches counterexamples with a global optimiser, penalises them in a resumed training loss, and certifies the result once Bauer-Marquart et al. (2021). Cohen and Strichman search over a small subset of weights with an SMT solver and add soft constraints to keep the repaired decision boundaries close to the original Cohen & Strichman (2022). Yang et al. compute the unsafe input region by reachability analysis and retrain on examples drawn from it, which makes the adversarial examples representative of the unsafe region Yang et al. (2021). Boetius et al. read counterexample-guided repair as a robust optimisation algorithm, prove termination for restrained models, disprove it in general, and derive a quadratic-programming repair for linear regression Boetius et al. (2023). BIRDNN makes behaviour imitation the repair objective and repairs domain-wise, which is the closest published objective to the fidelity reading this paper reports Liang et al. (2024b). RL4F trains a critique generator to steer a fixed model, so the message crosses a frozen interface, but the generator is learned against the end task Akyürek et al. (2023).

Two consequences follow for this paper. The specifications in that literature are given to the operator, which is what makes model modification the natural repair; here the target is the receiver's own behaviour and the specification is not available, so what can be established is a property of the evidence. And the fidelity objective BIRDNN shares with this paper is measured there against a reference the repair may modify, whereas here the reference is frozen and the modification is confined to the wire.

Two refinements of the selective-prediction line bear directly on this setting. One-sided selective prediction bounds the error on the accepted region Gangrade et al. (2021), and the benchmark literature fixes the coverage and conditional-error conventions this paper reports Pugnana et al. (2024), noting that a selector's value can come from the ranking it induces Pugnana & Ruggieri (2022). Abstention also allocates risk unevenly across the states it rejects, and the group-conditional reading is reported where the calibration curve is Lenders et al. (2025). The comparison throughout is against a coverage-matched placebo, a control on the factors the mechanism should not read, which has a lineage in probe evaluation, where control tasks pair each input with a random output Hewitt & Liang (2019).

A certificate built from an agreement region is studied in learning theory, and the two results there are the nearest theoretical neighbours of this paper. A subset of a labelled sample proves a test label when every hypothesis in the class that makes at most b mistakes on it agrees there; the shortest such subset is characterized by the robust hollow star number of the class, and the sample size that produces one is bounded by a per-point certificate coefficient Blum et al. (2025). Certified computation reads the same budget from the other side, bounding the verifications a corrupt dataset needs before its output is ϵ-accurate and giving an instance-optimal scheme for the minimum Gouleakis et al. (2018). Both bound the size of a witness drawn from labelled records in a fixed hypothesis class. The witness here is the candidate set the public evidence leaves open, which no record labels, and the cost is counted in queries, which is what makes the cost a property of the target: the same evidence buys a decision at one price and an identity at another, and the finite family is where the two are computed exactly.

Selective classification under distribution shift takes its guarantee from a shift-corrected draw Tibshirani et al. (2019); Liang et al. (2024a), and a safety filter in control takes its from a model of the plant Ames et al. (2019), so each is only as good as that assumption. The certificate takes its from the interfaces the public evidence leaves open, which is why it holds for a release the auditor has never sampled.

L.1 Probing frozen representations, and identifiability up to a transformation

A probe of a frozen system is a standard instrument, and what it reports is what a representation carries, which can differ from what its objective asked for. Continuous physical measurement is

recovered from frozen features Shkolnikov (2026); outcome value is recovered from frozen policies and used to select among their own sampled actions Zhang et al. (2026); discrete symbols are induced from a frozen latent space without supervision Liu (2026); and prompting is shown to underestimate what a frozen model's hidden states already separate Schumacher et al. (2026). Two of these carry the warning that applies to any such reading: a frozen representation can hold information that is not accessible through a given readout Datta et al. (2026), and a probe's apparent success can be a property of the probe Leask & Moubayed (2025).

The formal neighbours are identifiability results. A learned representation is identifiable in function space up to a linear indeterminacy under stated conditions Roeder et al. (2021), and the observational underdetermination of a representation-head factorization is characterized by a fiber criterion Sevetlidis (2026). The object here is neither: the internal state is never recovered and lies outside what is estimated, and what the audit asks is whether one action is safe for every interface the transcript leaves open, which is why the quotient of Section 3 is behavioural and why the cost is priced over the pairs a target separates.

A third line is an audit of identification: a protocol-level audit prices what a controlled evaluation can identify about its own object Luo et al. (2026), and where a frozen system is observable at all is settled by its architecture Carmichael (2026). Verification spends a probe set against a target under constrained access Wilding et al. (2026), which is the budget this paper prices from the other side.

The direction of the intervention separates a readout from a repair. A probe reads a property out of a frozen object and is judged by whether that property is decodable; a repair writes a message into a frozen receiver and is judged by whether the evidence identifies which message it installed. What stands in for decodability here is the set of messages the anchors leave consistent, which is the quantity the rest of this appendix reads.

L.2 Provenance records

The provenance lock re-audits the frozen capture independently. The fifteen checkpoint hashes, the two hundred and twenty-five ordered pairs and the three disjoint episode schedules reproduce ten million eight hundred thousand evaluation rows, and the declaration hash is recorded alongside the results. The second endpoint carries its own provenance record: eight checkpoint hashes with each training job's recorded configuration, the thirty-six ordered pairs its six checkpoints contribute, its own declaration hash, and its own independent audit, which re-derives every sampled decision from the recorded anchors and reports no replay and no decision-trail mismatch. The source tree contains the scripts, configurations and tests used to regenerate every reported quantity, covering the calibration-budget sweep, the risk reference-action calibration over gate-issued states, the target-query hierarchy, the falsification cases, the finite-sample width sweep, the transform families, the trained-receiver comparison, the MCP fixtures, and the advice-assisted automata comparison.

L.3 A task the message decides

The endpoint is episodic and its message carries the goal: the sender emits one symbol per episode, the receiver decodes it and acts on it, and the horizon is twenty steps. The repair is an adapter fitted from public probes, evaluated against a receiver that carries a mismatched adapter, against the reference pair, and against a policy that ignores the message. The chance returns are 5.0 at width four and 2.5 at width eight.

Width	Seeds	Mismatched	Repaired	Reference
4	10	1.4892	16.2497	16.2497
8	5	0.7925	11.9900	11.9900

Table 24: Task return of the repaired interface against the mismatched and reference arms, of a twenty-step horizon, at the promoted probe budget. The repaired arm reproduces the reference arm row for row.

The paired contrast of the repaired arm over the mismatched one is $+14.7605$ with interval $[13.0007, 16.3368]$ resampling seeds and $[14.5042, 15.0152]$ resampling episodes at width four, and $+11.1975$ with $[10.7446, 11.6370]$ and $[10.8065, 11.6050]$ at width eight. Every seed of both widths moves in the same direction.

The interface is non-degenerate by two independent readings. Every in-contract seed separates all of its messages: message-to-action distinctness is 4 of 4 at width four and 8 of 8 at width eight, the behavioural symmetry set has size one on every seed, and the receiver's greedy reply takes two to three distinct values at each position of a three-value action space. The probe-consistency set, enumerated over all bijections of the alphabet, has size $(d - b)!$ for b probes and is a singleton exactly at $b \geq d - 1$.

Probe access is limited on this endpoint. The estimator declines to act below $d - 1$ probes, and the repaired arm is then identical to the unrepaired one; at and above that budget its return does not vary with the number of probes, so the endpoint measures what the evidence identifies (Appendix L.3). The row counts are $160{,}000$ at width four and $80{,}000$ at width eight, and the recorded audits replay $24{,}000$ rows per population with no violation.

M SUPPLEMENTARY AUDITS

Each audit below extends the main contrast to a case the paper reports separately.

M.1 A TASK-EXTERNAL TARGET

The local gate was also evaluated against a task-external target: whether the selected action solved the task, pooled over the 225 ordered pairs and $15{,}840{,}000$ state rows at the declared configuration of target scope 3 and $k = 3$. The gate's issue-coverage difference from the coverage-matched permutation is exactly zero across every pair, and the paired difference in task utility is not distinguishable from zero: the sign test gives $p = 0.8873$ against the permutation and $p = 0.5543$ against message-only repair. The two targets therefore agree, and they agree in the direction that matters here. The target is a one-step landmark value external to the frozen policy, and it is not an episodic return, so this is a task-external target, on one environment family.

M.2 THE PLACEBO CONSTRUCT

The permutation arm reuses the local gate's issue events. Over 25 files and $3{,}200$ states the permutation reproduces the gate's issue decision on every state, with zero mismatches, and the two coverage shares are identical at 0.618125 for that sample. The first version of this check keyed states on the episode and goal, which repeat across target scope and neighbour count, so it compared rows drawn from different states; the redone check uses the full key.

M.3 SPLIT AND PAIR PARTITION

The gate's performance is reported across four evidence partitions formed by the evaluation and held-out splits and by same-seed and cross-seed pairs, the paired coverage difference from the permutation is exactly zero in all four, and the paired task-utility difference is not distinguishable from zero in any of them ($p = 1.00$, 0.3777, 0.7905, 1.00). Fidelity differences stay within ± 0.014. The same-seed cells carry 15 pairs and the cross-seed cells 210, so the pooled result is not carried by the same-seed controls.

M.4 PER-STATE CALIBRATION

The mismatch calibration is checked on held-out states. On $2{,}669{,}625$ held-out states issued by the local gate, the expected calibration error is 0.0249 and the maximum bin error is 0.1736, concentrated in the largest bucket, where the statistic takes the value 8 on $35{,}073$ states and the predicted rate 0.5854 accompanies an observed rate of 0.4118.

The rank statistics are reported for what they are. The correlation between the bin index and the held-out rate is 0.9762, and the correlation between the bin index and the prediction is the same 0.9762, so on eight points the rank statement is a description of the binning; the quantity the operator uses is the predicted rate, which tracks the observed rate by construction of the protocol.

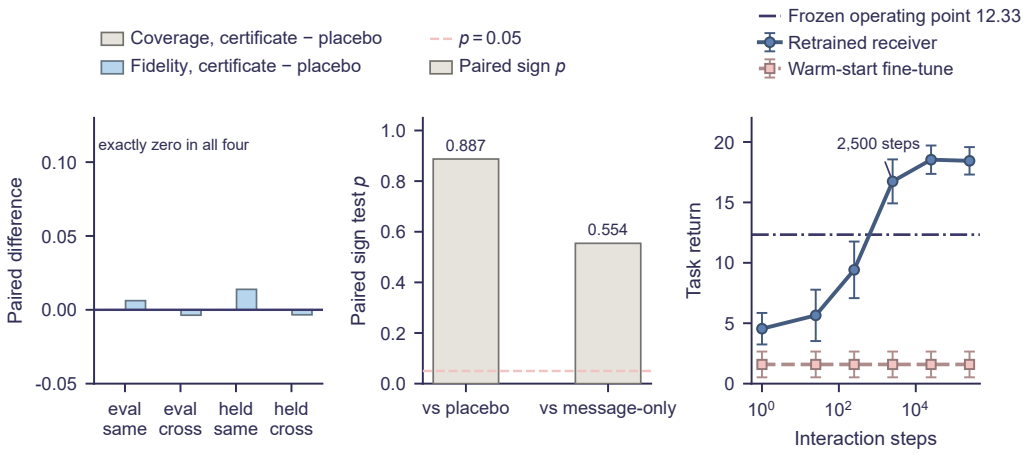

Figure 9: Robustness of the main contrast and the cost of the alternative. The paired coverage difference from the permutation is exactly zero in all four cells of the pair-type partition (a), the task-external target stays inside the noise (b), and retraining from scratch is measured against the frozen operating point (c).

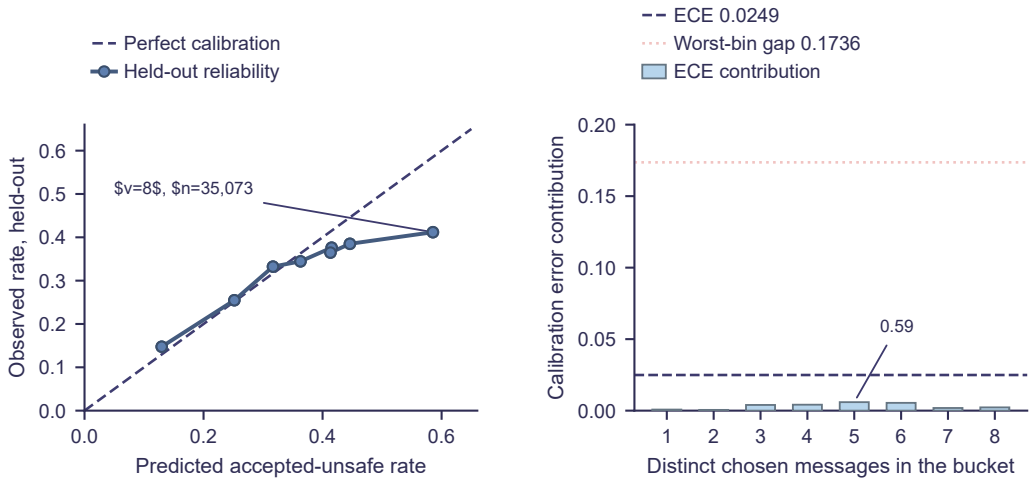

Figure 10: Held-out reliability over the bucket range with the perfect-calibration diagonal (a), and per-bucket contribution to the expected calibration error (b).

M.5 Reference-action mismatch inside each stratum

The rate reported throughout is pooled over gate-issued states, and a pooled rate can conceal where mismatch occurs. Recomputing it inside each stratum needs no new run, because every held-out state carries its goal, sender message and context.

Across the three goals the rate runs from 0.3863 to 0.4073, a spread of two points. Across the fifteen sender messages it runs from 0.3521 to 0.4484, a spread of ten, so the message a state arrived on carries more of the variation than the goal it belongs to. The largest spread is across how crowded a context is, measured by the distance to the nearest other gate-issued context carrying the same message. Contexts packed within 0.25 of another carry a rate of 0.4765 over $224{,}629$ states, those between 0.25 and 0.5 carry 0.3007 over $148{,}172$, and those between 0.5 and 1.0 carry 0.2141 over $27{,}668$. The risk is therefore concentrated where the evidence is dense, which is the same

mechanism the disagreement statistic reads: a crowded neighbourhood is where anchors disagree, and an isolated context is one the interface does not distinguish.

M.6 THE CROSSED STRUCTURE OF THE CHECKPOINTS

The 225 ordered pairs are a complete grid of fifteen sender checkpoints and fifteen receiver checkpoints, so they are not independent draws: pairs sharing a sender share that sender's policy and its message behaviour. A bootstrap that resamples pairs treats them as independent and returns an interval narrower than the claim it supports. Resampling the sender identities and the receiver identities, each with replacement, and reading the statistic over the pairs they induce is the standard treatment for a crossed design.

On the pooled coverage-matched difference it widens the interval by a factor of 2.2 to 2.4 at every budget. The two readings the paper leads with are unchanged. At four anchors the contrast is -0.0022 with interval $[-0.0286, +0.0227]$, which contains zero under either treatment, so the null holds. At sixteen it is $+0.0356$ with interval $[+0.0151, +0.0561]$, which excludes zero under either treatment, so the gain holds. The reading at eight does not transfer: $+0.0082$ has interval $[+0.0028, +0.0136]$ under the pair bootstrap but $[-0.0044, +0.0217]$ under the crossed one, so the separation is established at sixteen anchors. The same treatment applies to the certified-set contrast, where the interval widens by 2.26 at the declared budget and the null there holds as well.

M.7 THE THRESHOLD RULE ON ONE POPULATION

A calibration guarantee invites a direct comparison with the simplest rule that uses it: accept a state when the statistic is small enough. That comparison is only meaningful on one population, so both rules are computed here on the same states, the 225 ordered pairs of the held-out split with $720{,}000$ in-scope states per arm, so both rules are judged on identical states.

The local gate's operating point is issue coverage 0.5568 with reference-action mismatch 0.3930, and message-only repair gives 0.4919 and 0.3812. The threshold rule on the same states traces the frontier: accepting only states whose bucket statistic is 1 gives issue coverage 0.0266 at mismatch 0.0701, a threshold of 5 gives 0.4791 at 0.3824, a threshold of 8 gives 0.5556 at 0.3930, and at 9 the rule issues on exactly the states the local gate issues on and reproduces its operating point to the last digit. The strict ordering of the two levels is therefore not a property of the agreement rule: the calibrated statistic traces the whole mismatch-coverage frontier and the gate's rule sits at its end, which makes the statistic the decision instrument and the agreement rule the mechanism that exposes it.

The conformal protocol behaves the same way on this population. Choosing the threshold on the evaluation split and judging it on the held-out split, every declared level from 0.10 to 0.30 selects the threshold 1, whose held-out risk is 0.0701 at a coverage of 0.0266; no threshold meets the 0.05 level on the training split. The guarantee therefore holds with a wide margin, and the coverage it permits at that margin is small, which is the honest shape of a distribution-free bound on this evidence. The construction returns the gate's own rule, because the nonconformity score takes only three values on these anchors, so the split-conformal threshold and the agreement rule select the same states.

M.8 THE MISMATCH-COVERAGE FRONTIER ACROSS BUDGETS AND TARGET SCOPES

The frontier in the main text is drawn on one population, and an operator choosing a calibration budget needs it at each budget and each target scope, with the width of the interval the resampling gives. That grid was built by rebuilding the gate's issue decision per state per budget from the recorded calibration anchors and reading the in-scope set and the outcome from the recorded rows. The rebuild reproduces the recorded decisions at every cell, zero mismatches over $22{,}500$ state-budget comparisons on each of the two splits, and the statistic it thresholds reproduces the recorded per-state statistic exactly, zero mismatches over $8{,}640{,}000$ states. At the reference cell it reproduces the main-text frontier point by point: a threshold of 1 gives issue coverage 0.0266 with interval $[0.0141, 0.0408]$ at a reference-action mismatch of 0.0701 with interval $[0.0144, 0.1272]$, a threshold of 5 gives 0.4791 at 0.3824, and the local gate's operating point is 0.5568 with interval $[0.5232, 0.5939]$ at 0.3930 with interval $[0.3754, 0.4097]$.

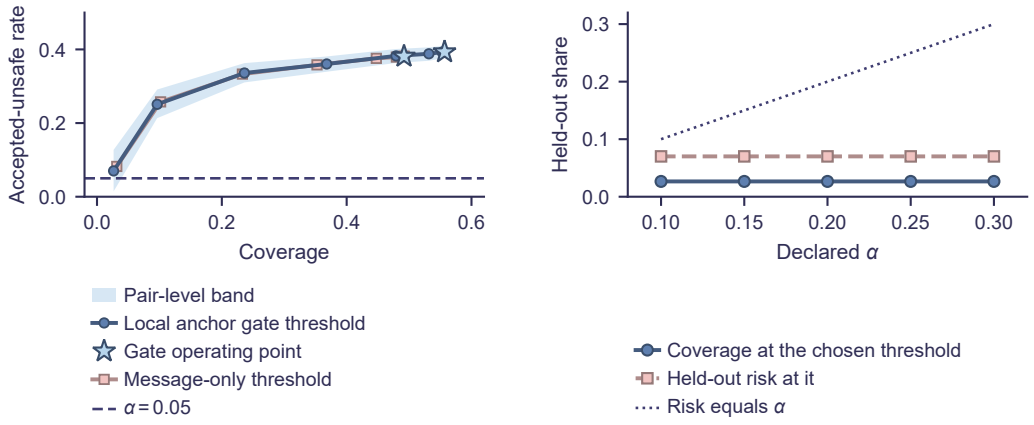

Figure 11: Reference-action mismatch against issue coverage on one population. The frontier traces thresholds of the calibrated statistic for the context gate and message-only repair, with the context gate's operating point marked (a), and the conformal protocol over the declared levels (b).

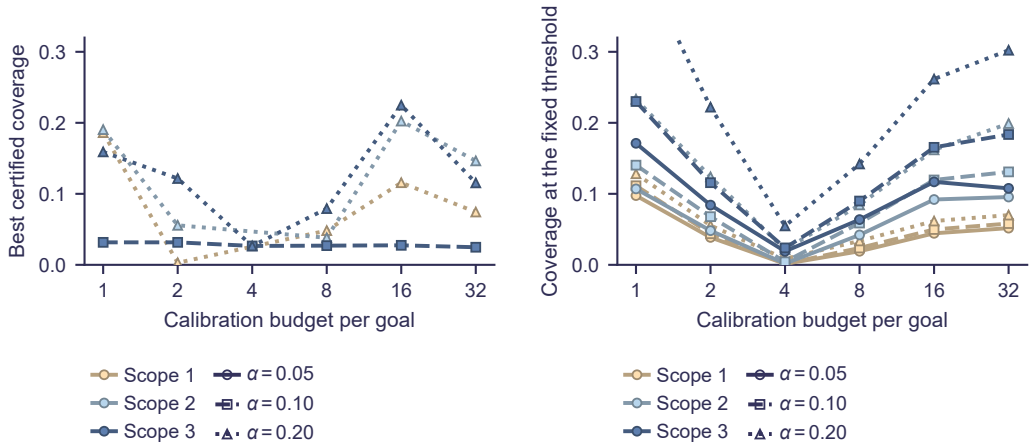

Figure 12: The best issue coverage at each declared mismatch level against the calibration budget, one series per target scope, read on the held-out frontier (a) and with the threshold fixed on the calibration split (b). The low-mismatch end of the frontier is flat in the budget, while the fixed-threshold reading rises from four anchors per goal upward.

The low-mismatch end of the frontier is a property of the statistic. At 0.10 the best held-out issue coverage lies between 0.0248 and 0.0317 at every budget at the full target scope, and the smallest rate any threshold attains anywhere on the grid is 0.0589, which is where the statistic's discreteness puts the floor: a threshold of one admits only the states whose anchors already agree unanimously, and no budget moves that set. What the budget buys is issue coverage at a fixed mismatch level. Fixing the threshold on the calibration split and judging it on the held-out split, issue coverage at a reference-action mismatch of 0.05 rises from 0.0190 at four anchors per goal to 0.1168 at sixteen, and issue coverage at 0.10 rises from 0.0238 to 0.1836 at thirty-two. The two smallest budgets sit below the point where the agreement test binds: a message then carries fewer than three anchors on most goals, so the test admits nearly every state and the coverage is the population itself. The test starts to constrain the decision at four anchors, which is where the rise is measured from. The budget therefore buys issue coverage at a fixed mismatch level, while the frontier's own low-mismatch end, which chooses the threshold on the population it scores and is an upper reading for that reason, stays where it is.

	Scope 1		Scope 2		Scope 3	
Budget	0.10	0.20	0.10	0.20	0.10	0.20
1	none	0.1861	none	0.1905	0.0317	0.1591
2	none	0.0029	none	0.0557	0.0317	0.1216
4	none	none	none	none	0.0266	0.0266
8	none	0.0482	none	0.0381	0.0270	0.0792
16	none	0.1157	none	0.2024	0.0275	0.2248
32	none	0.0744	none	0.1464	0.0248	0.1154

Table 25: The best held-out issue coverage at each reference-action mismatch level, per calibration budget and target scope. An entry reads none where no threshold reaches that level at any coverage, and no cell reaches the 0.05 level. The scope-3 column at 0.10 is the flat end of the frontier.

M.9 A REPAIR MODEL FITTED ON THE SAME ANCHORS

The comparison the main text leads with is between two constructions of the same rule family, so it is checked here against a learned one. For each pair a classifier is fitted on that pair's calibration anchors alone, with the receiver context, the sender message, the candidate message and the goal as inputs. The labels come from the recorded anchor protocol: the first message that reproduced the anchor's action is positive, and the messages scanned before it are negative. At a state the rule installs the highest-scoring candidate, and it abstains below a threshold calibrated on the anchors alone. The pipeline is checked twice before it reports anything. The gate reconstructed from the anchors reproduces the recorded decisions on all 22,500 sampled states, and driving the receiver with the sender's own message reproduces the recorded native action on all 22,500 of them.

Rule	Output rate	Error on issued	Calibration reads
a model fitted on the same anchors	1.0000	0.3291	$\lvert B\rvert\, d$
the model, thresholded to the gate's coverage	0.5374	0.3204	$\lvert B\rvert\, d$
the set the evidence leaves	1.0000	0.0188	$\lvert B\rvert\, d$
the gate	0.5374	0.3057	$\lvert B\rvert$

Table 26: Four rules on the same 22,500 sampled held-out states of all 225 pairs, with the same anchors available to each. Error is the share of issued states where the action departs from the frozen receiver's. The fitted model is the strongest learned alternative at this evidence: it reads the receiver's action under every candidate message, as the set rule does, and its abstention threshold is tuned on the anchors.

M.10 SELECTIVE-PREDICTION BASELINES ON THE SAME ANCHORS

The question a learned rule raises is what it would add when it reads the same evidence. Three baselines were built on the same population, the same anchors and the same features, under the same protocol: the threshold is chosen on the evaluation split and judged on the held-out split. The local gate's issued set on the held-out split is 400,926 states at a reference-action mismatch of 0.3930, and the fitting split carries 399,327 states at a base rate of 0.3940.

The first baseline scores a state by the distance to its k-th nearest same-message anchor. It does not reach the low-mismatch region at all: no threshold on the held-out mismatch-coverage curve attains reference-action mismatch of 0.20, and fixing the threshold on the fitting split gives issue coverage 0.0115 at a mismatch of 0.1244 for the 0.05 level.

The second fits a logistic regression on the anchor distances, the statistic and the pool size, against labelled mismatch outcomes on the fitting split, reaching a training accuracy of 0.6076 against the base rate of 0.3940. On the held-out mismatch-coverage curve it reaches issue coverage 0.0095 at mismatch 0.1000, against 0.0478 at 0.0701 for the declared statistic. The statistic leads at the reported mismatch levels, reaching issue coverage 0.0478 at mismatch 0.10 against 0.0095 for the fitted predictor, while the predictor leads at 0.20, where it reaches 0.1014 against 0.0478. Fixing the threshold on the fitting split sharpens the difference in what the two spend to reach a level: the predictor takes issue coverage 0.0292 at a realized mismatch of 0.0992 for the 0.10 level, while the statistic takes 0.0244 at 0.0302.

The comparison is settled by what each rule consumes. The statistic reads calibration anchors, which are public probes. The fitted predictor additionally reads labelled mismatch outcomes on the fitting

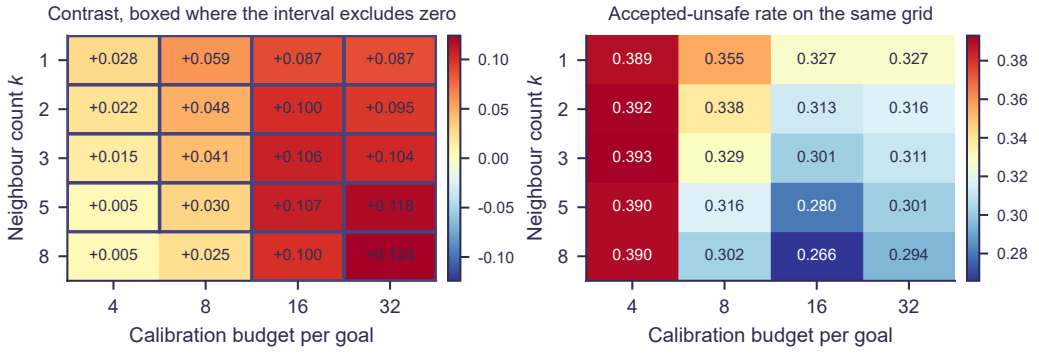

Figure 13: The contrast over the neighbour count and the calibration budget, boxed where the paired interval excludes zero (a), and reference-action mismatch over the same grid (b). The contrast is positive in every cell and the declared setting is at the low end of the range.

split, which means the operator already knows, for those states, whether the receiver's action was the native one. At the low-risk operating points the cheaper input is also the better instrument, and the distance score, which reads the same probes as the statistic, does not reach those points at all.

M.11 The hyperparameter surface, and what each setting means

The reported contrast is measured at one neighbour count and one calibration budget, so the question a reader is entitled to ask is whether it survives a move on either axis. A one-dimensional sensitivity cannot answer it, because the two axes interact: the neighbour count sets how local the bucket is, and the budget sets how many anchors the bucket can draw on. Both were swept together, over five neighbour counts and the four reference budgets, with the local gate keeping its issue events and the installed message shifted by one on the alphabet, so coverage is matched by construction at every cell.

The contrast is positive at all twenty cells and its paired interval excludes zero at eighteen of them. It spans $+0.0046$ to $+0.1241$, and the declared setting, three neighbours at four anchors per goal, sits at the low end of that range with $+0.0148$. The two cells whose interval crosses zero are eight neighbours at four and at eight anchors, where the gate-issued population falls to $239{,}575$ and $73{,}529$ states and the interval widens, while the point estimate keeps its sign. The reported setting is therefore not a tuned maximum: it is the cheapest setting on the grid, and the cells that beat it spend between four and eight times the calibration queries to do so.

The two axes have mechanical meanings. The neighbour count is the size of the unanimity test. The local gate returns a message only when the k nearest same-message anchors agree on it and on the target action, so raising k tightens the test and shrinks the issued population, which is why gate-issued states fall from $400{,}926$ at three neighbours to $246{,}753$ at five on the reference budget. The budget is the size of the pool those neighbours are drawn from, and that pool cannot be moved between goals: each goal's recorded anchors carry a single sender message, so the gate issues exactly on states whose message matches that goal, once it holds at least k anchors. An allocation that moves budget away from a goal removes its issue events. Under a rule that reallocates from the recorded per-goal evidence alone, the mean coverage difference from the recorded schedule is 0.0058 with interval $[-0.0112, +0.0214]$, so the fixed schedule is not leaving coverage on the table.

The threshold is not a setting in this sense at all. It is read off the calibration split by the conformal protocol, and the local gate's issue rule is the endpoint of the frontier that threshold traces.

M.12 The width of each point of the reliability curve

The reliability curve is reported as a rate per value of the statistic, with an expected calibration error and a worst gap, and none of those numbers carries a width. The interval was built by rebuilding the per-pair, per-bin counts from the recorded rows and resampling the ordered pairs, which is the unit the recorded contrasts use. The rebuild is checked before it is used: its issue decision matches

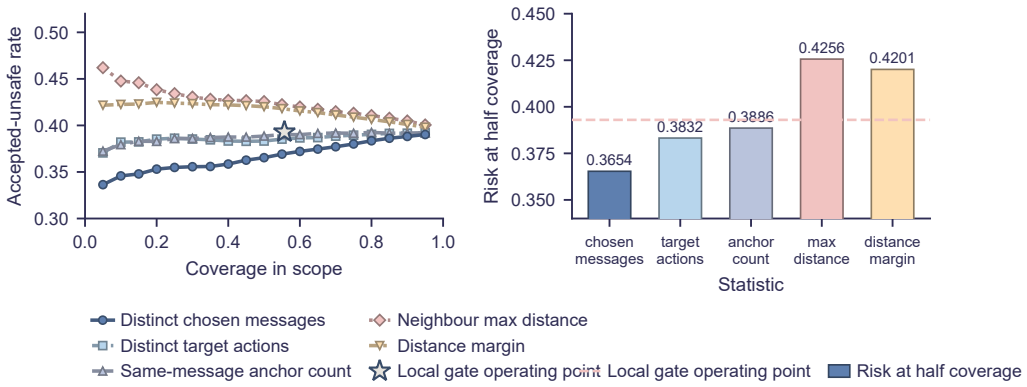

Figure 14: Which number to threshold. Reference-action mismatch against issue coverage for each recorded statistic on the local gate's issued states (a), and the same comparison at half the in-scope coverage (b). The reported statistic is the best decision instrument.

the recorded per-state flag, and its per-cell gate-issued counts match the recorded cells exactly at every one of the six cells the population covers, 128,567, 254,185 and 399,327 on the fitting split and 133,733, 250,055 and 400,926 on the held-out split, for the three target scopes in that order.

The intervals are narrow enough that the ordering of the curve is a reading. Value 1 carries reference-action mismatch of 0.1253 with interval $[0.0580, 0.1914]$, value 4 carries 0.4070 with $[0.3762, 0.4362]$, and value 7 carries 0.5024 with $[0.4471, 0.5570]$. The intervals widen at the top of the range, where the states thin out: value 8 holds 2,811 states and its interval is $[0.3305, 0.6506]$, and value 9 holds 1,796 states in a single pair, so its interval collapses onto its point.

The curve also differs by pair type, which the declaration treats as a unit. Splitting the ordered pairs into the fifteen whose sender and receiver were trained from the same seed and the two hundred and ten whose seeds differ, the expected calibration error is 0.0242 on the same-seed pairs and 0.0105 on the cross-seed ones. The two curves differ in shape. The cross-seed curve rises throughout, while the same-seed curve steps from 0.396 at value 2 to 0.243 at value 3 and then rises, so the pooled monotonicity is carried by the cross-seed pairs.

M.13 Which statistic to threshold

The paper reports that the two discrete counting statistics calibrate and the three geometric ones do not, and uses the primary one to trace the mismatch-coverage frontier. That choice implies a question: does the statistic that calibrates best also order the mismatch best when an operator thresholds it? Every recorded statistic was swept as a threshold rule on the local gate's issued states, so all rows act on like-for-like states at a given issue coverage, and the answer separates two properties that are easy to conflate.

The statistic the paper uses is also the best decision instrument, and by a clear margin: at half of the in-scope coverage it holds reference-action mismatch of 0.3654, against 0.3832 for the statistic with the smaller held-out calibration error, 0.3886 for the same-message anchor count, and 0.4201 and 0.4256 for the two geometric statistics. The ordering is therefore worth stating explicitly, because the two readings do not agree: a smaller calibration error and a better ordering of the decision are two different properties, and on this evidence the statistic to threshold is the one with the coarser fit.

A rule that reads the consistent set uses the statistic better. Issuing wherever the local gate admits gives a reference-action mismatch of 0.3930 at issue coverage 0.5568, while thresholding the primary statistic reaches a lower mismatch in the same coverage regime. The frontier comparison identifies the statistic as the decision instrument and the agreement rule as the mechanism that exposes it.

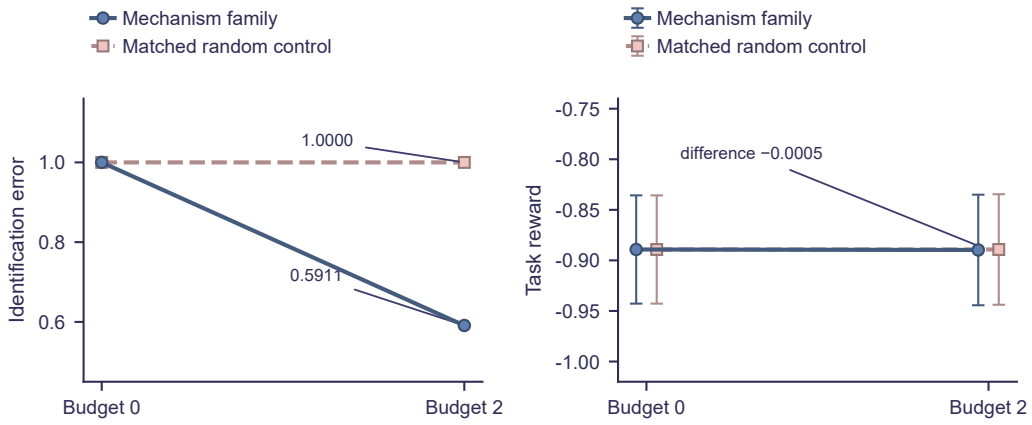

Figure 15: The separation on a second environment. Identification error against the probe budget for the mechanism family and its matched random control (a), and task reward over the same budgets (b). The two coincide at zero probes and separate at two.

M.14 How much of the candidate set a probe must scan, and what the geometric statistics measure

The oracle probe in this paper is an evaluation-only upper bound, and its cost is one query per candidate message. Two questions follow: how much of the candidate set a probe must scan before the realizability verdict is settled, and whether the statistics that fail to calibrate are nonetheless carrying information the discrete ones miss. Both are answered from the recorded per-state action tables over 225 pairs and 720,000 realizable states.

A scan of the first m candidates in the recorded order settles the verdict on 59.4 percent of states at $m = 1$, 74.1 percent at $m = 2$, 87.9 percent at $m = 4$, 96.4 percent at $m = 8$, and 100.0 percent at $m = 16$. Complete recovery therefore needs the whole set, while a scan of half of it leaves the verdict open on under four states in a hundred, which is the cost axis of the evaluation-only bound.

The geometric statistics are not noisier copies of the discrete ones. Over the same states the correlation between the two discrete statistics is 0.356, while the geometric statistics correlate with the primary discrete statistic at -0.025, -0.156 and 0.332, and among themselves at no more than 0.19 in absolute value. They measure something else. What they do not do is refine the decision: inside a fixed value of the primary statistic, the neighbour maximum distance and the same-message anchor count do not split the risk at all, and the distance margin splits off a small subpopulation with a different risk in an inconsistent direction, 0.1283 against 0.3169 at one distinct message and 0.6054 against 0.4266 at three. A statistic that measures something else without refining the decision is consistent with the ordering measured elsewhere in this appendix, where the geometric statistics are the worst instruments for the risk.

M.15 A second environment

The contrast was recomputed on MPE/simple_reference, where 160 of 160 evaluation cells pass their independent audits. With no probes the two families coincide exactly (identification error 1.0000 for both), which checks that the procedure separates nothing on its own. With two probes the mechanism family's identification error falls to 0.5911 while the matched random control stays at 1.0000, and the task reward does not move: -0.8897 against -0.8892, a difference of -0.0005, with the coordination error also unchanged at 0.5089 against 0.5090. A 41-point move in the identification quantity buys nothing on the task on this environment, which is the same separation the main result reports on the primary one.

This document carries the audits the main paper's claims do not rest on, together with the implementation. It holds the secondary audits moved out of the main appendix, the tables and figures

those audits carry, and the runners, audit scripts and configuration files named in the reproducibility statement.

M.16 THE ADAPTER BUDGET FAMILY

The adapter budget sweep behind the finite-sample statements covers two campaigns, 150 summary files and 1,000 recorded arms, each conforming to the declared schema, over probe budgets 0, 1, 2, 4, 8, 14, 15 and 16, and ten seeds. It is the source of the per-source saturation reported in Appendix B of the main paper, and it is listed here so that the number of independent runs behind the interface statements is visible.

Server	Versions	What one probe yields	Label cap	Repair gain		What the repair carries
filesystem	7	3	0.429	exact		label accuracy cap
sequential-thinking	9	class identified in one probe	n/a	0.222 1.000	→	type normalization
memory	10	envelope classes in one read probe	n/a	0.300 1.000	→	graph adapter, compatibility layer
everything	28	ten classes from one probe	n/a	0.821 1.000	→	advertised action mapping

Table 27: The released protocol versions audited, one row per server. Version labels are capped by the public aliases while the behavior-class decision stays exact. The composition suite crosses three frozen topologies with the observed classes.

Target scope	Budget	Output rate	Unsafe	Native	Matched gain	Message-only
1	4	0.5546	0.4069	0.5852	-0.0049	0.5440
1	16	0.4561	0.2928	0.6065	+0.0139	0.0515
1	32	0.5412	0.2999	0.6089	+0.0157	0.0210
2	4	0.5306	0.3968	0.5962	+0.0009	0.4620
2	16	0.4491	0.3062	0.6246	+0.0332	0.0361
2	32	0.5359	0.3145	0.6262	+0.0335	0.0103
3	4	0.5568	0.3930	0.6034	-0.0022	0.4919
3	16	0.4514	0.3013	0.6470	+0.0356	0.0591
3	32	0.5381	0.3106	0.6478	+0.0421	0.0317

Table 28: The three declared target scopes at three budgets, held-out split. Coverage shrinks as the scope widens, and the coverage-matched gain over the permutation grows with the budget within every scope.

Reliability	Capacity	Chernoff	Fano id.	Fano tgt	Measured
0.60	1.0164	0.4679	28.47	2.40	7
0.80	1.7306	0.9334	16.72	1.41	4
0.90	2.1767	1.3933	13.30	1.12	3
0.95	2.4387	1.8375	11.87	1.00	2
1.00	2.7726	inf	10.44	0.88	1

Table 29: The analytic reference beside the measured per-source cost at $d = 16$, in nats per query. The measured cost sits near the target bound and far below the identity bound.

Statistic	Held-out fit	Risk at half coverage
distinct chosen messages	the reported primary	0.3654
distinct target actions	smaller held-out error	0.3832
same-message anchor count	reported	0.3886
neighbour maximum distance	does not calibrate	0.4201
distance margin	does not calibrate	0.4256

Table 30: Every recorded statistic used as a threshold rule on the local gate's issued states, over 225 pairs. Reference-action mismatch is measured among issued states at half the in-scope coverage, so every row acts on a like-for-like set. The reported statistic is also the best decision instrument.

Value	Predicted	Observed	Gap	States	ECE share
1	0.1297	0.1475	+0.0178	111,997	0.00075
2	0.2517	0.2548	+0.0030	456,493	0.00052
3	0.3167	0.3323	+0.0157	676,924	0.00398
4	0.3629	0.3445	-0.0184	603,178	0.00416
5	0.4152	0.3761	-0.0391	405,046	0.00593
6	0.4132	0.3647	-0.0485	299,377	0.00544
7	0.4459	0.3850	-0.0609	81,537	0.00186
8	0.5854	0.4118	-0.1736	35,073	0.00228

Table 31: The reliability table behind the figure, on $2{,}669{,}625$ held-out states issued by the local gate. The expected calibration error is 0.0249 and the largest single gap is 0.1736, at the value 8. The last column sums to the expected error.

Arm	Pairs	Coverage	Utility	Fidelity
Context gate	220	0.5568	0.1681	0.6034
Message-only	188	0.4919	0.1675	0.6026
Coverage-matched permutation	220	0.5568	0.1689	0.6056

Table 32: The task-external target, pooled over 225 ordered pairs. The paired utility difference from the permutation is -0.00079 with interval $[-0.00686, +0.00528]$ and a sign test at $p = 0.8873$.

Split	Pair type	Pairs	Coverage	Coverage	Fidelity	Utility diff.	Sign p
evaluation	same seed	15	0.4869	0.0000	+0.0062	-0.00027	1.000
evaluation	cross seed	210	0.5595	0.0000	-0.0037	+0.00135	0.378
held out	same seed	15	0.5224	0.0000	+0.0139	-0.00438	0.791
held out	cross seed	210	0.5593	0.0000	-0.0034	-0.00052	1.000

Table 33: The split by pair-type partition. The paired coverage difference from the permutation is exactly zero in all four cells, and the fidelity difference stays within ± 0.014.

Budget	Issue rate	Message-only	Matched gain	Matched mismatch	Mismatch
1	0.7943	0.9438	-0.0019	+0.0024	0.3959
2	0.6123	0.7661	-0.0035	+0.0057	0.3998
4	0.5568	0.4919	-0.0022	+0.0040	0.3930
8	0.3134	0.1624	+0.0082	-0.0263	0.3286
12	0.3343	0.0751	+0.0216	-0.0645	0.2971
16	0.4514	0.0591	+0.0356	-0.0788	0.3013
32	0.5381	0.0317	+0.0421	-0.0782	0.3106

Table 34: The budget sweep on the held-out split at target scope 3. Issue coverage is matched between the context gate and its permutation by construction in every row. The matched gain rises monotonically and the reference-action mismatch difference flattens after sixteen anchors.

CODE AND DATA

The supplementary material includes the complete implementation, configuration files, evaluation scripts, archived outputs, and a README describing how to reproduce the reported results. The code and data are provided in an anonymized form for review. All experiments reported in the paper can be reproduced from the anonymized code, configurations, and evaluation artifacts included in the supplementary material. The data are the archived raw outputs of the frozen endpoints with their checkpoint hashes; the compute is a single workstation and no pretrained model beyond the released endpoints is used; the seeds are fixed and listed per run; and every reported quantity states its unit of resampling, its interval and its multiplicity correction. The tree is organised as follows.

- scripts/: the runners that produce each reported quantity, one family per file, each writing raw JSON beside its summary. The calibration-budget sweep, the reference-action calibration over gate-issued states, the target-query hierarchy, the falsification cases, the finite-sample width sweep, the transform families, the trained-receiver comparison and the protocol fixtures are all covered here. The second endpoint is produced by the same runner

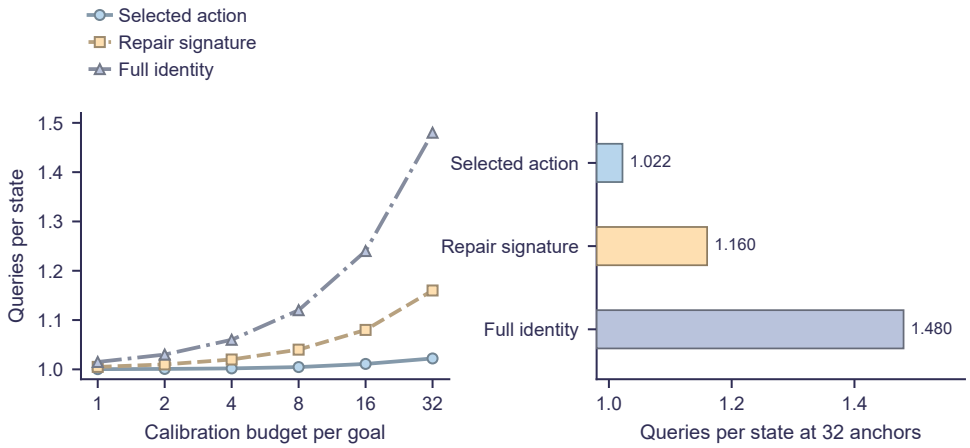

Figure 16: Queries per state for each target level, against the calibration budget (left) and at thirty-two anchors (right), on the held-out split at target size one. Selecting the target action costs 1.022 queries per state against 1.160 for a repair signature and 1.480 for a full identity, and reaches the same decision quality as the signature level.

family, which takes the endpoint as a parameter whose default reproduces the primary endpoint, so its reduction and its audit are the same scripts that reduce and audit the primary one.

- configs/: the declared calibration budgets, target scopes, neighbour counts, seed lists and held-out split definitions that those runners read.
- tests/: the invariant and replay checks. These include the cost identity $C_{\text{cal}} + C_{\text{online}} = C_{\text{total}}$, which holds exactly on every audited cell, and the replay checks that recompute each cell's metrics from the per-state rows.
- artifacts/: the archived outputs, each with its fixed seeds and the log of the independent checker that audited it.
- requirements-lock.txt: the pinned Python environment used for every reported run.

The per-pair rows behind the paired statistics of Figures 2 and 5 of the main paper are included, so the coverage, reference-action mismatch and native-match differences can be recomputed without re-running the environment. Every runner is deterministic given its seed.

M.17 Content-sensitive metrics

A fidelity metric cannot see whether a repair preserved the relation between the message and the action, so that relation was measured directly. Over 40 of the 225 pairs, $2{,}816{,}000$ rows, the mutual information between the receiver message and the receiver action is 0.0247 bits lower for the local gate than for its permutation, and message-cluster purity is 0.0385 lower. Two properties of the measurement bound what it can show: the permutation is a deterministic function of the pair, the target scope and its own seed, so its tighter message-action relation is a property of that construction, and forty pairs are a weak instrument for a difference of this size. The measurement is reported as the direction of the effect under the declared construction.

M.18 The calibration budget does not transfer between sources

The anchors are paid for per source, so the natural question is whether a budget spent unevenly across goals beats the recorded equal split. It cannot, and the reason is structural. Each goal's captured anchors all carry one sender message: across 240 goals over 80 pairs, every goal anchors exactly one message and every one of its anchors carries it. The local gate needs k agreeing anchors inside one sender message, so a goal's issue rate switches on at k anchors and does not improve afterwards, and taking budget away from a goal does not buy coverage elsewhere because no other goal can supply the missing message.

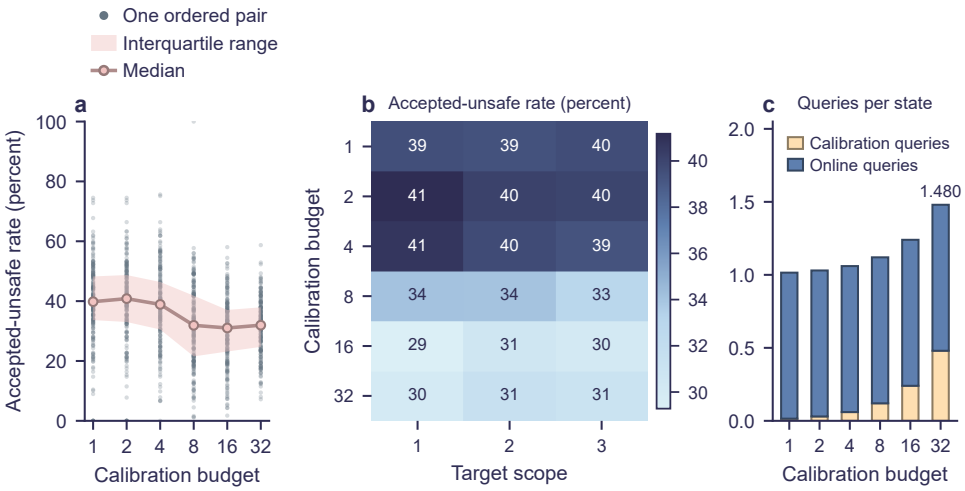

Figure 17: Accepted-unsafe rate over ordered pairs as the calibration budget grows (left), the rate across target scopes and budgets (middle), and calibration and online queries per state (right).

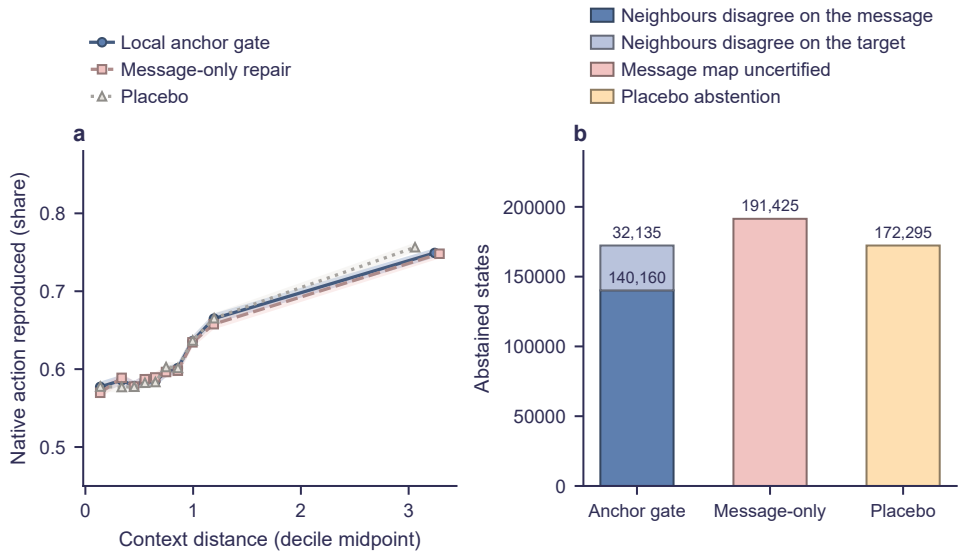

Figure 18: Native action reproduction against context-distance decile for the three rules with 95 percent intervals (left), and the abstention counts by reason (right).

The measured response follows. Giving every goal n anchors and sweeping n over $\{3, 4, 6, 8, 12\}$ moves coverage from 0.5849 to 0.5791, 0.4508, 0.3208 and 0.3444 at mean query costs of 13.2, 18.4, 30.0, 43.3 and 71.3. More evidence does not buy more coverage; past k it costs coverage. The paired coverage difference between three anchors per goal and the recorded four is $+0.0058$ with interval $[-0.0112, +0.0214]$, so the recorded allocation sits on a plateau, and the decline after it has a single cause. Decomposing the abstentions over $256{,}000$ in-scope states per level, the share abstaining because the neighbours disagree on the receiver message rises from 0.3312 at three anchors per goal to 0.4183, 0.5082 and 0.5298, while the share abstaining on the reference action stays between 0.1018 and 0.1581. The extra anchors place disagreeing contexts into the neighbourhood, and the local gate abstains on them as its rule requires. What the anchors carry explains why: reading the capture order, the modal receiver message of a goal's anchors holds 0.8833 of them at the first anchor and 0.5292 by the twelfth, a drifting consensus, while the reference action is close to uniform over the five movement actions at every index, with a modal share near 0.26. The resource

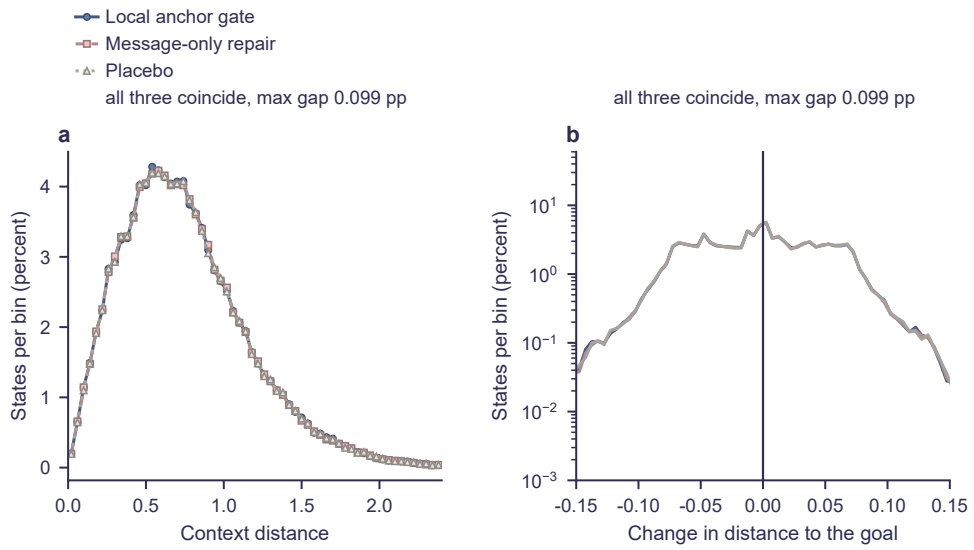

Figure 19: Distribution of states over context distance (left) and over the change in distance to the goal (right). All three rules place their mass in the same places.

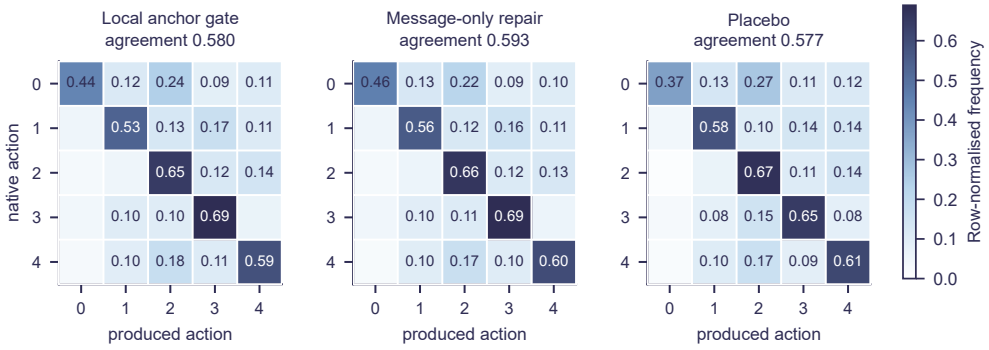

Figure 20: Row-normalised agreement between the frozen native action and the action produced, over about two hundred thousand gate-issued states per rule. The three rules leave the same fingerprint.

that limits the local gate is the agreement structure of the anchor pool, which no allocation of a query budget changes.

M.19 A SOURCE IS WORTH MORE THAN ANCHORS

The budget sweep varies the number of calibrated goals, and the local gate indexes its pool by the sender message, so the two are separate axes. Across the 225 pairs of the primary endpoint every goal carries exactly one sender message, and in 90 pairs two of the three goals carry the same one, so three calibrated goals open two sources. The census depends on the calibration pools alone. At a target scope of two goals 90 pairs open one source and 135 open two; at a scope of three the same split is 90 pairs at two sources against 135 at three.

Opening a source is worth more than deepening one. Moving the target scope from one goal to two adds four anchors in every pair, and whether those anchors open a new source or join an existing one is fixed by the sender. Where they open a source, coverage at three neighbours moves from 0.1743 to 0.3785, a paired gain of $+0.2042$ with interval $[+0.1770, +0.2305]$ over 135 pairs; where they join an existing source, coverage moves from 0.1849 to 0.3149, a gain of $+0.1299$ with interval $[+0.0904, +0.1689]$ over 90 pairs. The intervals do not overlap, and the two groups start from the same place, 0.1743 against 0.1849, so the difference lies in what the added anchors reach.

The mechanism is the one the previous subsection measures. A source the pool has not sampled leaves a third of the states without a bucket, and no amount of evidence on a source the pool has sampled reaches them, because the decision is taken inside the bucket the sender message selects. Deepening a bucket buys agreement among neighbours the pool already holds, and the recorded response counts bound how much agreement is left to buy.

M.20 THE COVERAGE-MATCHED CONTRAST ON THE SECOND ENDPOINT

The main comparison is the local gate against its coverage-matched placebo, and it was recomputed on MPE/simple_reference through the same runner, the same reduction and the same independent audit, over the six checkpoints that environment contributes and all 36 of their ordered pairs, at the declared target scope and neighbour count on the held-out split. All 36 pairs complete, the independent audit re-derives every sampled decision from the recorded anchors alone and finds no replay mismatch and no decision-trail mismatch, and the campaign lands on the same pre-declared branch as the primary endpoint. Local realizability is 0.9864, so the interface reproduces the reference action on almost every state, and the gate issues on 0.2782 of them.

The second endpoint places the content effect further along the budget curve the primary endpoint traces. Coverage is equal in both arms by construction. On the second endpoint the gate matches the target action on 0.7541 of its issued states against 0.5057 for the placebo, a paired difference of -0.2934 on the placebo-minus-gate scale with interval $[-0.4091, -0.1827]$ over 36 clusters, where the primary endpoint reads -0.0043 with interval $[-0.0253, +0.0170]$ over 220 at four anchors and reaches $+0.0356$ once the budget rises. The same instrument, the same reduction and the same control therefore detect a content effect where one is present, which makes the four-anchor reading on the primary endpoint a measured absence. The unconditional count agrees, where the placebo loses 0.0691 of native decision match over all in-scope states, interval $[-0.1039, -0.0356]$.

The two axes the comparison separates are the ones the paper reports throughout. The endpoints differ in architecture as well as in environment, since the second shares one actor between the roles and carries the message in a second action head. The message-only contrast on this endpoint is not coverage-matched, since that arm issues on 0.2456 of in-scope states against 0.2782 for the context gate; the matched construction above aligns the issue coverage. The placebo leads on the task-relevant action by 0.0286 with interval $[+0.0131, +0.0465]$, the same separation between decision fidelity and task utility that the primary endpoint reports.

M.21 A THIRD ENVIRONMENT

MPE/simple_spread was run as a further family: $1{,}248$ evaluation runs of which $1{,}246$ were reduced, over $6{,}230{,}000$ trace rows, crossing eight controller families with capability, intent, capacity and probe budget. The controller-family axis does not vary the outcome on this run family: seven families resolve to the same per-seed checkpoints and five of them produce numbers identical to four decimal places, the matched random control among them. What varies is the capability axis, worth 2.9 reward units, and the capacity axis, worth 0.29. The controller family is inert on this run family, and the capability and capacity axes are what move the outcome.

M.22 THE INTERACTION COST OF THE ALTERNATIVE

The local gate is available without environment interaction. Training a receiver from a fresh initialisation reaches the frozen operating point of 12.3325 at 100 episodes, which is $2{,}500$ interaction steps at the declared episode length of 25, and its return there is 16.7404 ± 1.8196 over twelve seeds. Warm-start fine-tuning of the frozen pair stays at 1.5894 ± 1.0658 at every interaction budget up to $250{,}000$ steps, where training from scratch reaches the frozen operating point after roughly one hundred episodes.

M.23 THE NEIGHBOUR COUNT

The declaration fixes three neighbours and names one and five as its sensitivity, and that axis has not been reported until now. The recorded permutation exists only at the primary setting, so the control here is a declared one: the local gate keeps its issue events at each neighbour count, and the

message it installs is shifted by one on the alphabet, which matches coverage by construction exactly as the recorded permutation does. The action each message induces is read from the per-state action table, whose indexing is checked by the no-adapter identity on every row of the sweep, $720{,}000$ of $720{,}000$.

The covered contrast is the same at every declared neighbour count. On the held-out split at target scope 3, the paired difference against the shifted control is $+0.0279$ with interval $[+0.0114, +0.0439]$ at one neighbour over $720{,}000$ states, $+0.0262$ with interval $[+0.0053, +0.0471]$ at three over $400{,}926$, and $+0.0287$ with interval $[+0.0034, +0.0565]$ at five over $246{,}753$. The sign, the size and the clearance of the interval are unchanged across the range, so the result is robust to the neighbour count.

What the neighbour count moves is issue coverage, and the state counts show it: a single neighbour issues on every state in the panel, three on $400{,}926$, and five on $246{,}753$, because a stricter agreement requirement abstains more often. That is the same division of labour the rest of the paper reports, with the agreement rule setting issue coverage and the evidence setting the quality of what is covered.

The other half of the neighbourhood definition is the metric, and the result survives replacing it. Two alternatives were run over all 225 pairs under the same shifted control: dividing each coordinate by that pair's own anchor standard deviation before taking the Euclidean norm, so that coordinates in different units contribute comparably, and cosine distance, which discards magnitude entirely. Counted by state, the paired difference against the control is $+0.0240$ with interval $[+0.0034, +0.0443]$ under the recorded unweighted Euclidean metric over $416{,}918$ gate-issued states, $+0.0218$ with interval $[-0.0001, +0.0428]$ under standardised Euclidean over $405{,}756$, and $+0.0246$ with interval $[+0.0013, +0.0470]$ under cosine over $388{,}039$. The sign and the magnitude are stable to within four thousandths; the standardised weighting is the one variant whose interval reaches zero at its lower end. The reason the axis is weak is visible in how much of the gate-issued set moves: the two metrics agree with the recorded one on about nine gate-issued states in ten, with Jaccard overlap 0.8918 for standardised against Euclidean and 0.8931 for cosine, so the neighbourhood is largely the same neighbourhood and the contrast is not an artefact of the coordinate system it was measured in.

M.24 The statistic across budgets

The recording declares the disagreement statistics budget independent, which is a statement about their definition: the number of distinct chosen messages in a bucket never consults the budget. Whether the *reliability* of that statistic also transfers across budgets is a separate question, and it is the one an operator asks when choosing a budget. It was therefore measured per budget by rebuilding the bucket, the neighbours and the issue decision from the recorded calibration anchors alone, and validating the reconstruction against the recorded decisions: zero mismatches over $90{,}000$ state-budget comparisons, which also re-derives the local gate rule independently.

The prediction transfers across the whole declared range, and its residual error grows where the statistic's tail becomes reachable: the largest value the statistic attains grows with the budget, from 4 at four anchors per goal to 9 at thirty-two, so the tail enters the comparison only at the higher budgets and only with the fewest states behind it. The expected calibration error is 0.0091 at four anchors per goal, 0.0405 at eight, 0.0199 at sixteen and 0.0473 at thirty-two, against the pooled 0.0249 reported in the main text. The large gaps sit at the largest values of the statistic, which only the larger budgets reach and which hold the fewest states: at thirty-two anchors the worst bucket gap is 0.5452 at the value 8.

Budget per goal	ECE	Worst bucket gap	At value	Reachable to	Scored states
4	0.0091	0.1909	4	4	5,993
8	0.0405	0.5312	6	7	3,553
16	0.0199	0.2973	7	8	5,340
32	0.0473	0.5452	8	8	6,137

Table 35: Calibration per budget on the declared pair split, with the largest value of the statistic each budget can reach. The residual error concentrates in the high values, which hold the fewest states.

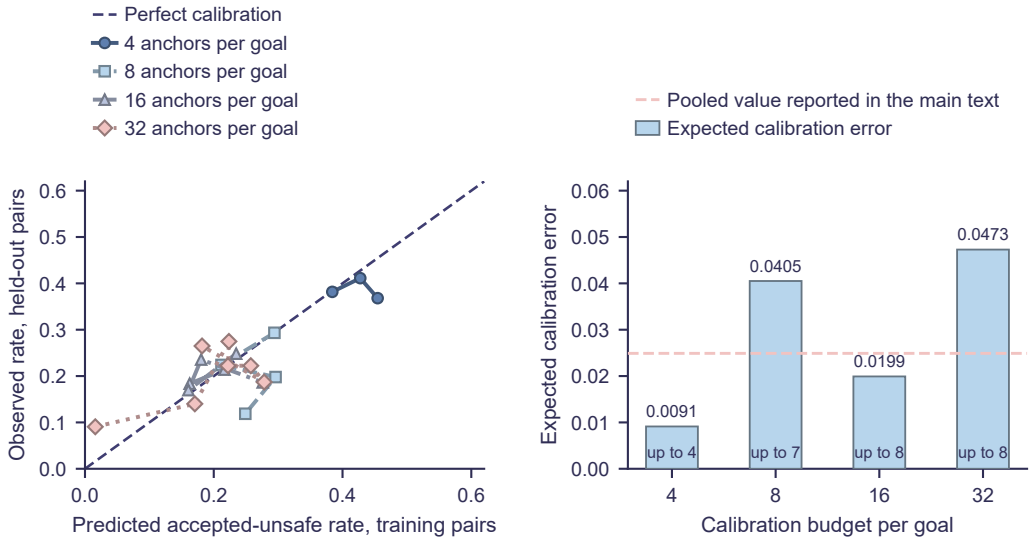

Figure 21: The statistic across budgets. Predicted against observed reference-action mismatch on the declared pair split, one series per budget (a), and expected calibration error per budget against the pooled value (b). The reachable range grows with the budget.

M.25 Which end performs the repair

A repair can be installed at either end of the channel: the sender can transmit the message the receiver needs, or the receiver can overwrite the message it received. The two are the same intervention whenever the interface is invertible, because in that case the receiver's effective input is the installed message either way. That premise is measurable, and it was measured: grouping the held-out rows of every arm by state and asking whether one installed message ever produced two different receiver actions gives zero violations over 303,929 (message, action) pairs across 192,000 states in 60 pairs. The receiver action is therefore a function of the installed message alone, so no experiment on these endpoints can distinguish a sender-side repair from a receiver-side one. The declared families are cyclic, dihedral, invertible row-stochastic, near-permutation and block permutations, all invertible on the alphabet, so on the recorded runs every message is installable from either end. Where the ends separate is therefore not invertibility but image containment, and that is measurable. Declaring a many-to-one interface whose image is the first k symbols and applying it to the same gate-issued messages gives, over 400,926 gate-issued states in 225 pairs, an installability that falls from 1.0000 at $k = 16$ to 0.9702 at $k = 1$. The divergence between the two ends, meaning states the receiver end repairs and the sender end cannot, is exactly zero for every image of eight symbols or more and rises to 0.0014, 0.0029 and 0.0074 at four, two and one. Two of the zero-divergence images are already non-invertible, which locates the condition precisely: a sender-side repair is available whenever the message selected by the gate lies in the image of the interface, and the gate-issued messages need only about half the alphabet, so a substantial collapse of the interface costs almost nothing. The receiver end carries no such condition, and the difference between the ends is bounded by 0.0074 on this population.

A separate interface-level confidence-set audit considers the endpoint transforms. Every recorded interface maps the alphabet onto itself with all outputs distinct, and the certificate audit checks that the true transform is a unique member of its candidate family. An injected non-invertible condition needs endpoints trained under one, and that experiment was run. Five endpoints were trained at width 8 under a many-to-one interface whose image holds four of the eight symbols, so the interface is not invertible, and the same certificate audit was then applied to them over 270 cells. The guarantee holds under it: the minimum realized coverage is 0.9844 against the declared $1 - \delta$ of 0.95, the maximum unsafe-return rate is 0.0156 against a δ of 0.05, and no cell reports an error. The record therefore does not depend on the interface being invertible, and what a collapsing interface costs is paid on the sender-side condition measured above.

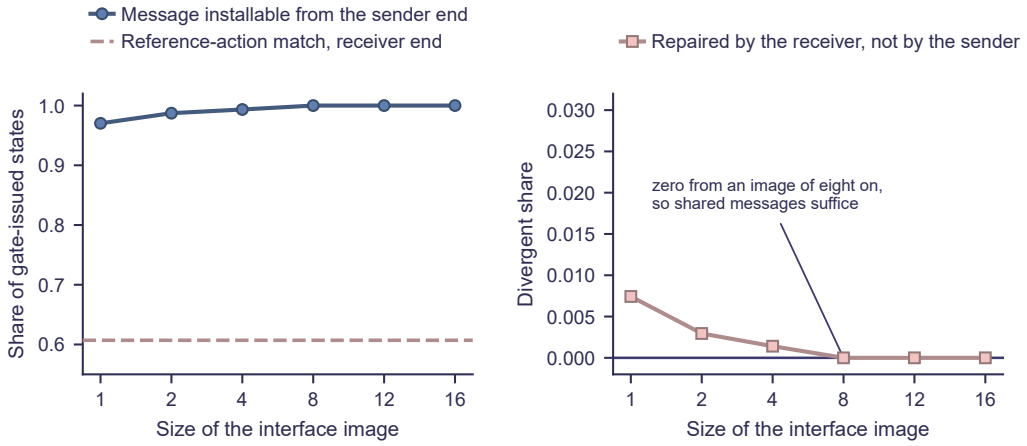

Figure 22: Where the ends separate. The share of gate-issued states whose selected message a sender-end repair can install, against the size of the interface image (a), and the divergence between the ends (b). Installability stays above 0.97 even when the image has one symbol.

M.26 The message-only contrast is a coverage contrast

The permutation control is matched by construction, but the message-only arm is matched by budget, and its coverage falls as its own budget grows, so a matched-budget comparison between the two arms mixes decision quality with coverage. The mix is now resolved by asking what the two arms do on the states they share. Over 225 pairs and the same held-out states, $29{,}553$ states receive an output from both gates, and on every one of them the two install the same message ($29{,}553$ of $29{,}553$) and reach the same decision. The arms do not differ where they overlap: they differ in which states they reach, with $5{,}298$ states issued only by the context gate and $5{,}172$ only by the message-only map.

The matched-budget gain of the context gate over message-only repair is therefore an issue-coverage gain. That is the same conclusion the permutation control reaches from the other side, and it is the mechanism the budget sweep shows: as the budget grows the message-only map becomes stricter and issues on fewer states, while the context gate keeps its issue coverage, so the advantage accumulates on the coverage axis. This is why the contrast against the permutation, whose coverage difference from the context gate is exactly zero, carries the paper's claim and the contrast against message-only repair does not.

AI Use Statement

We used generative AI tools to assist with the conceptual ideation of the mechanism schematic, translation and polishing of the manuscript language, parts of the code, and literature retrieval and discovery, including locating potentially relevant related work. The authors independently verified the authenticity and bibliographic details of every literature item used in the manuscript, including its authors, publication year, title, venue, and persistent identifier when available. We rigorously and carefully reviewed all AI-assisted content and take responsibility for the final manuscript, figures, code, and results.