DECISION-SUFFICIENT EPISTEMIC EVALUATION: WHEN TO REDESIGN AND WHEN TO ACQUIRE

Anonymous authors
Paper under double-blind review

ABSTRACT

AI evaluation often optimizes additional measurement before asking whether measurement is a valid repair: when a representation erases an action-relevant distinction, more data through that representation cannot recover it. We formalize this prior repair question by distinguishing semantic erasure from statistical insufficiency through two repair gates. Semantic aliasing induces unavoidable intervention regret, while preserved distinctions may remain statistically unrecoverable. Locally, decision recoverability requires $n \in \mathrm{Range}(G)$ and, when estimable, is governed by $\Gamma_D = m/\sqrt{n^\top G^\dagger n}$. After preservation, classical directional design targets acquisition along the downstream decision direction. Across 2,400 held-out controlled interfaces with hidden failure mechanism and encoder matrix, a blind gate reaches 92.6% diagnosis accuracy and lowers post-routing regret to 0.1175 versus 0.2085 for Always Acquire and 0.1710 for Always Redesign. We further train 1,800 representations without assigning semantic/statistical failure labels. Using only independent calibration separability, the gate attains regret 0.1442, below Always Acquire (0.1480) and Always Redesign (0.1648). Post-hoc decision-subspace overlap tracks the repairability crossover. Results support a contract-specific criterion for deciding whether additional measurement is a valid repair under controlled populations, rather than a universal natural-failure classifier.

1 INTRODUCTION

Optimal design asks what to measure next; decision-sufficient evaluation asks the prior question: when is additional measurement a valid repair?

Evaluation pipelines often conflate uncertainty reduction, representation adequacy and decision relevance, collecting more evidence even when the required distinction is lost.

$$\text{semantic failure} \Rightarrow \textsc{Redesign}, \qquad \text{statistical failure} \Rightarrow \textsc{Acquire}.$$

The Preserve–Recover–Acquire pipeline (Fig. 1) diagnoses whether measurement can repair a failure before optimizing what to measure. Therefore, the key question is whether additional measurement remains a valid intervention before optimizing its acquisition strategy.

What is new? Our contribution is not a new acquisition objective. Classical optimal design asks which observation to acquire once additional measurement is assumed useful. We formalize the logically prior *repair-validity* question: whether additional measurement can repair the current failure at all. This separates representation-induced loss of decision-relevant information, which requires redesign, from preserved but statistically insufficient information, which admits acquisition; we then test whether this distinction can be inferred from observable evidence when the generating failure mechanism is hidden.

Contributions. **(1) Repair-validity criterion.** We formalize semantic insufficiency and decision-subspace recoverability to distinguish REDESIGN from ACQUIRE, providing a criterion for when additional measurement is a valid repair rather than merely a possible refinement. **(2) Acquisition after preservation.** We operationalize decision-aware acquisition through classical directional (c-optimal) design after semantic validation and compare Random, A/D/E-optimal, weight-source and misspecification controls. **(3) Blind and learned-representation repair diagnosis.** Known-contract tests establish repair semantics; a blind study then hides the generating regime and encoder

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.

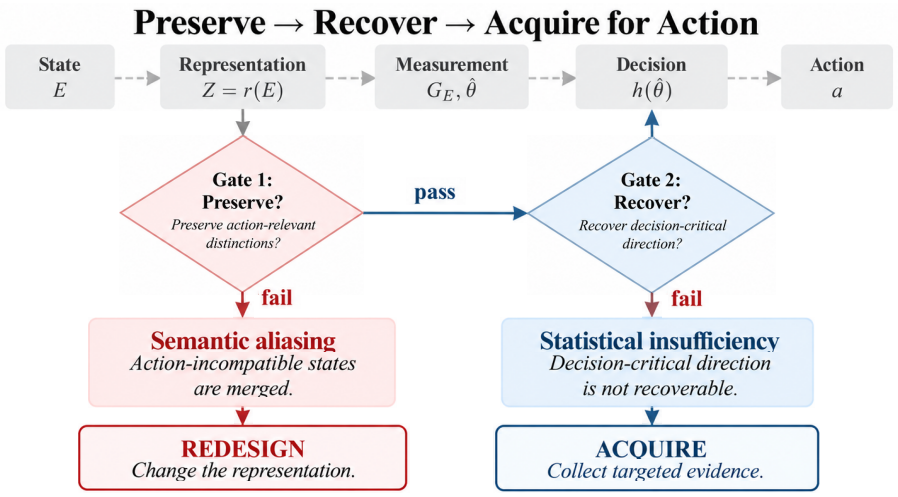


Figure 1: **Preserve–Recover–Acquire for specified downstream decisions under a declared evaluation contract.** Gate 1 tests whether action-relevant distinctions are preserved (REDESIGN if not). Gate 2 tests whether preserved directions are recoverable (ACQUIRE otherwise).

matrix and routes repair from observable calibration diagnostics. A second study removes hand-assigned failure types altogether: 1,800 learned representations arise from varying geometry, noise, and task-relevance pressure, and the same diagnostic principle reduces regret relative to fixed repair policies. Claim-bound and model-grounded tests provide complementary stress tests.

2 Related Work

AI evaluation, calibration, and behavioral diagnosis. Calibration assesses predictive reliability Guo et al. (2017); Geifman & El-Yaniv (2017); Kull et al. (2019); Ovadia et al. (2019); Kadavath et al. (2022). Behavioral evaluation probes beyond accuracy Liang et al. (2023); Ribeiro et al. (2020); Liu et al. (2023); Lin et al. (2022); complementary diagnostics Min et al. (2023); Es et al. (2024); Jacovi & Goldberg (2020) and benchmark analyses Gururangan et al. (2018); Sainz et al. (2023); Zheng et al. (2023); Raji et al. (2021) qualify score interpretation. We instead ask whether representations preserve distinctions required for measurement-based repair.

Measurement, representation, and decision sufficiency. Prior work addresses measurement meaning Cronbach & Meehl (1955); Messick (1995), latent-trait measurement Rasch (1960); Baker & Kim (2004); Reckase (2009), and representation abstraction Tishby et al. (1999); Alemi et al. (2017); Li et al. (2006); Ferns et al. (2004); Abel et al. (2018). Our semantic result relates to classical statistical decision theory Blackwell (1953); Lehmann & Romano (2005); Cover & Thomas (2006). We use representation sufficiency as a repair criterion: erasure requires REDESIGN before measurement.

Optimal design and adaptive evaluation. Classical A/D/E-optimal criteria improve global parameter estimation Pukelsheim (2006); Chaloner & Verdinelli (1995). For a fixed contrast $c^\top \theta$, c-optimal design minimizes $c^\top M^{-1} c$ Elfving (1952); Pukelsheim (2006), and goal-oriented design targets uncertainty in downstream quantities Attia et al. (2018). Our objective $n^\top G^{-1} n$ specializes classical directional design; the singular extension $n^\top G^\dagger n$ requires estimability, $n \in \mathrm{Range}(G)$.

3 Epistemic Measurement Problem

We distinguish claim-level epistemic state E, describing an individual conclusion; system-level competence θ, capturing repeatable epistemic behavior across items; and downstream decision D, mapping measured states to interventions. FASGL provides one empirical instantiation of the coordinate system:

$$E_j^{\mathrm{epi}} = (F_j, A_j, S_j, G_j, L_j),$$

where the coordinates correspond to Formation, Admissibility, Support, Attribution, and Lineage. The system-level epistemic competence is $\theta_m \in \mathbb{R}^K$ ($K = 5$ here). For theoretical binary items,

$$P(Y_{mj} = 1 \mid \theta_m) = \sigma(a_j^\top \theta_m - b_j),$$

while the empirical instrument uses ordinal F/A and the conditional likelihood $P(F)P(A)P(L)P(S)P(G \mid S)$ with proper-prior estimation.

Decision sufficiency as a representation property. For exposed representation $r(E)$ and state-induced decision $D(E)$, decision sufficiency requires

$$r(E) = r(E') \Longrightarrow D(E) = D(E'),$$

equivalently, the representation partition refines the decision partition. This FASGL-agnostic criterion separates representation redesign from acquisition when the relevant distinction is preserved but insufficiently measured.

Three evidence levels. *Formal measurability* concerns preservation and recoverability; *construct validity* concerns response to intended rather than nuisance variation; *external utility* concerns final-task improvement.

4 Epistemic Measurement Theory

Construct validity concerns what variation represents. The formal, logically independent requirements concern preservation of action-relevant distinctions and statistical recovery of retained coordinates.

Semantic sufficiency is a representation property. Let $r : \mathcal{E} \to \mathcal{Z}$ represent E, with actions $\mathcal{A}$ and loss $L(E, a)$. Define

$$A^*(E) = \arg\min_{a \in \mathcal{A}} L(E, a).$$

For a unique optimum with gap $\Delta(E) > 0$, denote it by $a^*(E)$. On this domain, decision sufficiency is

$$r(E) = r(E') \implies a^*(E) = a^*(E'), \qquad \text{equivalently } Q_D \preceq Q_r.$$

With ties, cell compatibility requires

$$\bigcap_{E : r(E) = z} A^*(E) \neq \varnothing,$$

a common optimum over the entire cell. This condition defines compatibility under set-valued optima rather than an equivalence relation between states.

Theorem 1 (Epistemic Aliasing Impossibility)**.** *Suppose E_1, E_2 satisfy $r(E_1) = r(E_2) = z$ but have incompatible unique optimal actions $a_1^* \neq a_2^*$. Let $\mu_i = P(E_i \mid Z = z) > 0$ and define the action gaps*

$$\Delta_i = \min_{a \neq a_i^*} \{L(E_i, a) - L(E_i, a_i^*)\} > 0.$$

Then every policy π measurable with respect to $Z = r(E)$ incurs conditional excess risk

$$\mathbb{E}[L(E, \pi(Z)) - L(E, a^*(E)) \mid Z = z] \geq \min\{\mu_1 \Delta_1, \mu_2 \Delta_2\} > 0.$$

Hence any representation cell that aliases states requiring incompatible interventions induces unavoidable intervention regret.

Regret depends on action-loss gaps as well as aliasing frequency (proof: Appendix B). A noisy analogue makes the neural-representation case explicit.

Corollary 1 (Approximate aliasing)**.** *Under equal priors and minimum incompatible-action gap Δ, if two states induce representation distributions with total-variation distance $\delta < 1$, every representation-measurable policy incurs excess risk at least*

$$\frac{\Delta}{2}(1 - \delta).$$

Thus exact aliasing is the endpoint $\delta = 0$ of a continuous indistinguishability–regret relation; proof is in Appendix B.

Full-vector versus decision-relevant identifiability. For calibrated binary items with $0 < p_j < 1$, let

$$J_{\text{item}} = [a_1, \ldots, a_N]^\top$$

and

$$W = \text{diag}(p_j(1 - p_j)).$$

Fisher information is

$$G_E(\theta) = \sum_j p_j(1 - p_j) a_j a_j^\top = J_{\text{item}}^\top W J_{\text{item}} \succeq 0.$$

Lemma 1 (Full-Vector Local Identifiability)**.** $\text{rank}(J_{\text{item}}) = K$ *iff* $G_E \succ 0$. *Hence full item-discrimination rank identifies the entire K-dimensional competence vector locally. However, this full-vector requirement is stronger than necessary for a specified downstream decision.*

Let $h(\theta) = 0$ be a smooth binary decision boundary, with unit normal

$$n = \nabla h / \|\nabla h\|$$

and local margin

$$m = |h(\theta)| / \|\nabla h\|.$$

The decision boundary h is treated as part of the downstream evaluation specification. A possibly singular instrument can still identify the local decision contrast precisely when

$$n \in \text{Range}(G_E).$$

When this holds, define

$$v_D = n^\top G_E^\dagger n, \qquad \Gamma_D = \frac{m}{\sqrt{n^\top G_E^\dagger n}},$$

where $G_E^\dagger$ is the Moore–Penrose pseudoinverse. If

$$n \notin \text{Range}(G_E),$$

the decision-critical contrast contains an unidentifiable component; for the recoverability test we set

$$v_D = \infty$$

and

$$\Gamma_D = 0.$$

Proposition 1 (Decision-Relevant Recoverability)**.** *Under regular local asymptotics restricted to the identifiable subspace and a local linearization of h,*

$$P(\textit{wrong decision}) \approx \Phi\left(-\frac{m}{\sqrt{n^\top G_E^\dagger n}}\right) = \Phi(-\Gamma_D), \qquad n \in \text{Range}(G_E).$$

Thus full-vector identifiability is sufficient but not necessary for a specified decision. A rank-deficient instrument can be decision-sufficient whenever the boundary normal lies in its identifiable subspace and Γ_D is large enough.

For example,

$$G = \text{diag}(10, 0)$$

is globally singular. For

$$n = (1, 0)^\top,$$

the decision is locally recoverable with

$$v_D = 0.1;$$

for

$$n = (0, 1)^\top,$$

it is not because

$$n \notin \text{Range}(G).$$

Hence

$$\text{global non-identifiability} \not\Rightarrow \text{decision non-identifiability}.$$

For the full-rank controlled instruments,

$$G_E^\dagger = G_E^{-1}.$$

4.1 DECISION-AWARE ACQUISITION

When Gate 1 passes but Gate 2 does not, candidate item j contributes

$$G_j = w_j a_j a_j^\top,$$

with

$$w_j = p_j(1 - p_j).$$

In the singular case, acquisition must first make the decision contrast estimable and then reduce

$$n^\top (G_t + G_j)^\dagger n.$$

In the full-rank/regularized implementation used in our experiments,

$$j_t^* = \arg\min_j n^\top (G_t + G_j)^{-1} n.$$

This is c-optimal design with $c = n$ Elfving (1952); Pukelsheim (2006). Sherman–Morrison gives the exact one-step reduction:

$$S_j^D = \frac{w_j (n^\top G_t^{-1} a_j)^2}{1 + w_j a_j^\top G_t^{-1} a_j}, \qquad j_t^* = \arg\max_j S_j^D. \tag{1}$$

If

$$a_j^\top G_t^{-1} n = 0,$$

then

$$S_j^D = 0 :$$

global information need not inform the decision. The implementation computes

$$w_j$$

from the current MAP estimate

$$\hat{\theta}_t :$$

$$\hat{p}_j = \sigma(a_j^\top \hat{\theta}_t - b_j).$$

True-θ weights are used only as an oracle-information comparator.

Proposition 2 (Greedy decision-normal variance reduction)**.** *Under Proposition 1, for a fixed decision normal n, margin $m > 0$, and $G_t \succ 0$, maximizing S_j^D minimizes the one-step local approximation to wrong-decision probability over the candidate pool.*

This result characterizes one-step local improvement and does not characterize global optimality of adaptive designs.

Adaptive evaluation loop. Given $\hat{\theta}_t$ and decision normal n_t at round t, if

$$n_t \notin \mathrm{Range}(G_t),$$

acquisition first expands the estimable subspace toward n_t. Once

$$n_t \in \mathrm{Range}(G_t),$$

evaluation stops if

$$\Gamma_{D,t}$$

meets the precision threshold. Otherwise the full-rank/regularized implementation selects

$$j_t = \arg\max_{j \in \mathcal{C}_t} \frac{\hat{w}_j (n_t^\top G_t^{-1} a_j)^2}{1 + \hat{w}_j a_j^\top G_t^{-1} a_j}, \qquad \hat{w}_j = \hat{p}_j(1 - \hat{p}_j).$$

Equation (1) is defined only on the estimable subspace. Gate 1 remains necessary because acquisition cannot recover erased distinctions.

5 CONTROLLED EPISTEMIC INTERVENTIONS

FASGL instantiates controlled factual/RAG, observation-limited reasoning and agent/tool-tracing items. Pure interventions target one role, compound interventions multiple roles, and surface controls preserve the declared state. Target responses assess manipulation sensitivity; off-target movement and missingness bound construct validity (construction and binding rules: Appendix F).

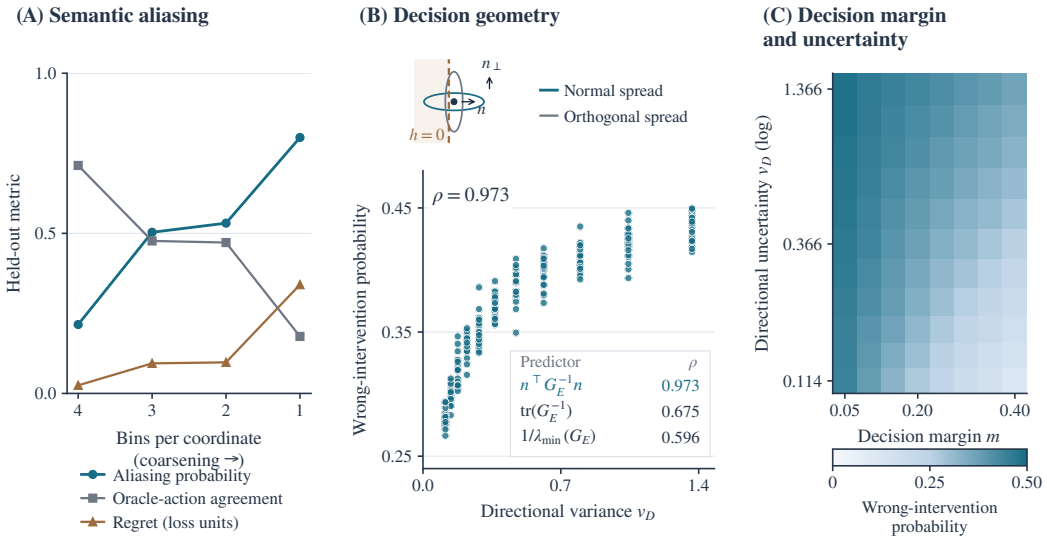


Figure 2: **Two failure mechanisms.** (A) Semantic coarsening increases aliasing and intervention regret. (B) Directional Fisher variance is more strongly associated with wrong-intervention probability than global information summaries; (C) error varies jointly with decision margin and directional uncertainty. B/C use Gaussian-surrogate decisions; contours are schematic and correlations descriptive.

6 Experimental Evaluation

Target studies use Gemma, Qwen and Mistral; calibration uses separate systems. Controlled simulations have known ground truth; model-generated studies assess empirical relevance and validity boundaries. Studies retain separate protocols, with uncertainty procedures and instrument specifications in the appendices.

6.1 RQ1–RQ2: Preservation Failure and Decision Recoverability

RQ1: Does representation loss create irreducible decision regret? The controlled reconstruction uses 20,000 independent-uniform five-coordinate states and an equal-weight deficit intervention oracle, fitting cell policies on 16,000 states and evaluating on 4,000 held-out states. Nested adjacent-bin merges coarsen each coordinate from four to three, two, then one bin. Decision-class aliasing (held-out within-cell action-conflict probability) increases from 0.215 to 0.799, oracle-action agreement decreases from 0.712 to 0.178, and mean intervention regret rises from 0.025 to 0.340. Aliasing and regret increase monotonically across the four coarsenings, testing decision-partition preservation.

RQ2: When is a preserved decision statistically recoverable? Geometry varies information along a fixed decision normal and an orthogonal direction across 220 full-rank five-parameter instruments, each with 60 binary items. Figure 2B applies the true boundary to $\theta + \epsilon$ using 2,000 perturbations $\epsilon \sim \mathcal{N}(0, G_E^{-1})$ per instrument. At margin 0.20, boundary-normal variance $n^\top G_E^{-1} n$ has Spearman $\rho = 0.973$ with empirical wrong-intervention probability, compared with 0.675 for $\mathrm{tr}(G_E^{-1})$ and 0.596 for $1/\lambda_{\min}(G_E)$. Across eight margins from 0.05 to 0.40, the same family's mean empirical error falls from 0.465 to 0.248. At margin 0.40, error ranges from 0.116 to 0.364 across the tested variance levels. Figure 2C averages 20 instruments at each of 11 variance levels. This agrees with $\Gamma_D = m/\sqrt{n^\top G_E^{-1} n}$ under the imposed covariance model; finite-binary estimation is assessed separately.

Aliasing varies representations with actions fixed; geometry assumes preserved distinctions. Both are FASGL-independent; Γ_D is evaluated under the stated covariance model.

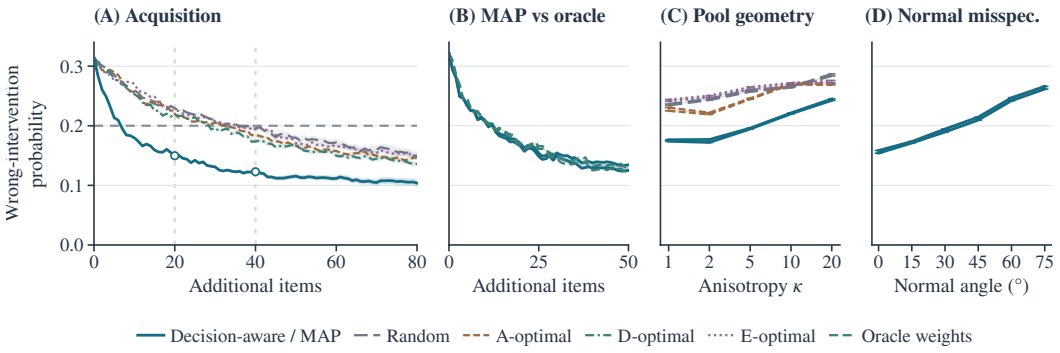


Figure 3: **Decision-aware acquisition and robustness.** (A) Decision-aware acquisition reduces wrong-intervention probability relative to Random and A/D/E-optimal under frozen reconstruction. (B) MAP- and oracle-weight bounds are close, without establishing equivalence. (C) Decision-aware remains favorable to A/E-optimal and Random across pool anisotropy. (D) Error increases with acquisition-normal misspecification; evaluation stays fixed. C/D use 20 additional items. A shows midpoint lines and bound ribbons; B–D show bound edges/ranges. Missing outcomes are retained; intervals are bounds rather than confidence intervals. Dashed $P_{\text{wrong}} = 0.20$: reference, not stopping rule.

6.2 RQ3: When Recoverable, Can Acquisition Help?

After preservation, acquisition policies share a six-item seed instrument and fixed candidate pool across 30 repetitions of 100 paired histories, with budgets up to 80 additional items. We compare the MAP-based rule in Eq. (1) with Random and A/D/E-optimal acquisition. At 20 additional items, decision-aware P_{wrong} has missing-outcome bounds $[0.149, 0.151]$, versus $[0.218, 0.222]$, $[0.217, 0.220]$, and $[0.224, 0.227]$ for A/D/E-optimal, and $[0.228, 0.230]$ for Random. At 40 items, the corresponding bounds are $[0.120, 0.126]$ versus $[0.181, 0.188]$, $[0.170, 0.177]$, $[0.192, 0.196]$, and $[0.192, 0.199]$. Missing-outcome bounds $[W/N, (W + M)/N]$ are not confidence intervals. Decision-aware upper bounds fall below all comparator lower bounds at both checkpoints. Retrospective error-0.20 crossings occur in 28 of 30 repetitions; missing outcomes leave two unresolved.

Panels B–D assess weight-source, pool-geometry and normal robustness. At 20 additional items, MAP- and oracle-weight decision-aware bounds are $[0.165, 0.170]$ and $[0.167, 0.170]$; their overlap does not establish equivalence. For randomized pools with $\kappa \in \{1, 2, 5, 10, 20\}$, decision-aware upper bounds stay below A-optimal, E-optimal and Random lower bounds at budget 20. Misspecifying only the acquisition normal from $0°$ to $75°$ increases error bounds from $[0.154, 0.159]$ to $[0.262, 0.266]$, with the true evaluator fixed. Appendix D details these robustness controls and additional endpoints.

Together, the results define an ordered repair logic: preservation precedes recoverability, and acquisition is admissible only when the decision-relevant distinction remains represented,

$$Q_D \preceq Q_r \;\longrightarrow\; [n \in \text{Range}(G), \Gamma_D] \;\longrightarrow\; S_j^D.$$

6.3 RQ4: Can Repairability Be Diagnosed?

Known-contract sanity check. Does a contract-conditioned gate select the appropriate first repair? The factorial study compares Gate-guided, Always Acquire, Always Redesign and Random on Gaussian observations of five-coordinate states. Equal-rank encoders either erase the decision normal or preserve it with insufficient precision. Acquire adds 16 observations under the unchanged encoder; Redesign re-encodes the four initial noisy observations without new evidence. The gate uses only declared-encoder representation and precision diagnostics. The endpoint is intervention regret

$$R_{\text{int}} = |n^\top x| - a n^\top x$$

for the declared binary action $a \in \{-1, +1\}$. Appendix K.3 specifies evaluation conditions.

Across 30 seeds with paired histories, semantic Acquire regret is 0.372 versus 0.152 for Gate/Redesign; statistical Redesign regret is 0.152 versus 0.037 for Gate/Acquire. The paired

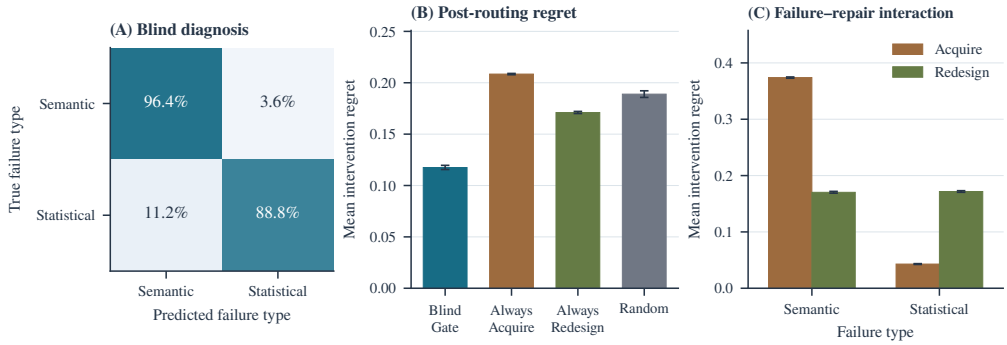


Figure 4: **Blind repair diagnosis under controlled contracts.** (A) The gate infers semantic versus statistical failure from held-out observable calibration probes without access to encoder matrices or regime labels. (B) Blind routing lowers mean intervention regret relative to fixed repair policies. (C) The predicted crossover remains visible: Redesign is effective after semantic erasure, whereas Acquire is effective after statistical insufficiency.

failure-by-repair interaction is 0.335 (95% CI [0.329, 0.340]). Gate yields zero wrong repairs and 100% mechanism-prescribed repair accuracy under declared encoder contracts, evaluating routing rather than learned universal failure discovery. In the equal-regime mixture, Gate regret is 0.095 (95% whole-seed bootstrap interval [0.092, 0.098]), versus 0.205, 0.152 and 0.179 for Always Acquire, Always Redesign and Random. Redesign restores information but may require subsequent acquisition.

The full known-contract crossover curves are reported in Appendix K.3.

Blind diagnosis without regime labels. Can repair be selected without revealing the failure-generating regime? The preceding factorial study establishes repair semantics under a declared encoder contract, but its gate has direct access to encoder structure. We therefore run a separate blind diagnosis study in which the gate receives neither the semantic/statistical regime label nor the encoder matrix, corruption flag, hidden state, oracle repair label, or realized loss. For each unknown interface it observes only independent noisy calibration probes in the exposed representation together with their downstream action labels. A pair-disjoint calibration classifier estimates action separability; its threshold is selected on separate meta-calibration interfaces and then frozen before evaluation. Low separability routes to REDESIGN; otherwise the interface routes to ACQUIRE. Repair budgets match the controlled contract: Acquire adds 16 observations under the unchanged representation, whereas Redesign re-encodes the same four initial noisy observations without adding evidence. Full protocol and accounting are in Appendix K.4.

Across 30 seeds and 2,400 held-out interfaces, blind diagnosis accuracy is 92.6%: semantic recall is 96.4% and statistical recall is 88.8% (Fig. 4A). The errors are consequential rather than hidden. Under semantic erasure, Blind Gate regret is 0.1772 versus 0.1703 for Always Redesign and 0.3740 for Always Acquire; under statistical insufficiency, Blind Gate regret is 0.0578 versus 0.0430 for Always Acquire and 0.1717 for Always Redesign. In the equal-regime mixture, Blind Gate regret is 0.1175, compared with 0.2085 for Always Acquire, 0.1710 for Always Redesign, and 0.1889 for Random routing (Fig. 4B–C). Relative to Always Acquire, the paired seed-level regret reduction is 0.0909 (95% seed-bootstrap CI [0.0886, 0.0931]); relative to Always Redesign it is 0.0535 [0.0519, 0.0551].

Learned-representation repairability. Does repairability emerge when representation failure is learned rather than assigned? We train rank-three linear representations on interfaces that vary state covariance, observation noise, decision direction, and task-relevance pressure, without semantic/statistical failure labels. ACQUIRE freezes the learned representation and increases the evidence budget from four to 20 observations; REDESIGN relearns it with stronger decision relevance at the original evidence budget. The blind gate observes only 24 independent paired calibration probes and a threshold frozen on separate meta-interfaces; critically, it never observes the learned subspace, decision normal, relevance weight, noise level, post-hoc overlap, oracle repair, or realized evaluation loss. Full details are in Appendix K.5.

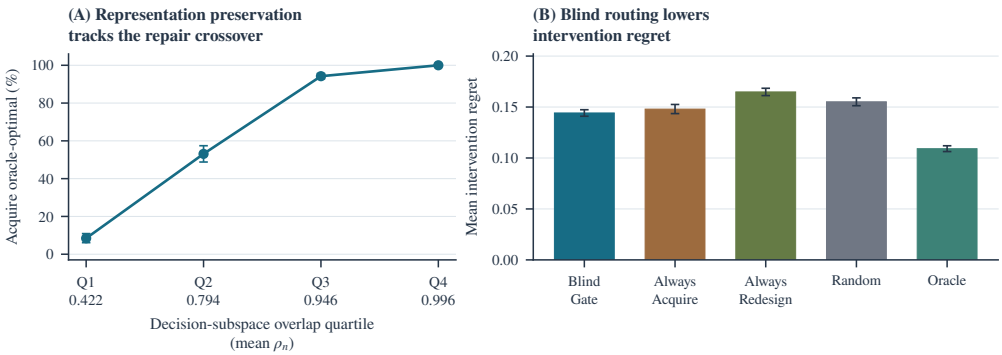


Figure 5: **Learned-representation repairability.** (A) Across post-hoc decision-subspace-overlap quartiles, the fraction of interfaces for which Acquire is oracle-optimal rises from 8.4% to 100.0%; overlap is diagnostic only and is never observed by the gate. (B) Across 1,800 held-out interfaces, Blind Gate has lower mean intervention regret than Always Acquire, Always Redesign, and Random; whiskers are 95% whole-seed bootstrap CIs.

Across 30 seeds and 1,800 held-out interfaces, Blind Gate regret is **0.1442**, below Always Acquire (0.1480), Always Redesign (0.1648), and Random (0.1551); the oracle is 0.1091 (Fig. 5B). Paired seed-level reductions are 0.0038 [0.0013,0.0066] versus Always Acquire and 0.0206 [0.0161,0.0252] versus Always Redesign. Exact repair-label accuracy is only 65.4%, yet routing regret is lower than all fixed policies: not all routing errors are equally costly, so decision quality is the primary endpoint rather than label accuracy.

Post-hoc decision-subspace overlap tracks the continuous crossover: in the lowest-overlap quartile, Acquire is oracle-optimal for only 8.4% of interfaces, whereas in the highest-overlap quartile it is oracle-optimal for 100.0%. This is not presented as the surprising fact that task-aligned training preserves task directions; rather, it tests whether the repair criterion tracks which intervention actually lowers held-out regret after representation learning. Additional evidence becomes a valid repair as decision-relevant structure is increasingly preserved. This remains a controlled learned-representation result, not a claim about naturally occurring deployment failures.

7 Discussion

Decision sufficiency need not require full rank: $n \in \mathrm{Range}(G)$ can suffice. More generally, acquisition is useful only when the action-relevant distinction remains represented and recoverable; Preserve–Recover–Acquire therefore asks whether measurement is an admissible repair before optimizing what to measure next.

The evidence progressively removes oracle knowledge: known-contract tests establish repair semantics, blind routing hides the encoder and regime, and learned representations remove hand-assigned failure types. The latter still uses controlled linear subspaces, so it does not establish diagnosis of naturally occurring deployment failures. Additional claim-bound, model-grounded, and real-model qualification tests are reported in Appendix A and qualify rather than enlarge the central claim.

Scope and limitations. Local Fisher/acquisition approximations condition on supplied rule h and may fail under misspecification, nonstationarity, or moving geometry. Pools are synthetic, FASGL is manually specified, and calibration probes are labeled. External validation requires nonlinear competing representations, blinded natural failures, and held-out repair outcomes (Appendix K.2).

8 Conclusion

Preserve–Recover–Acquire distinguishes representation failures requiring Redesign from preserved but imprecise distinctions admitting Acquire. Blind controlled diagnosis reaches 92.6% accuracy; across 1,800 learned representations without failure labels, calibration-based routing lowers regret relative to fixed repair policies and tracks a continuous repairability crossover. The result is a contract-specific repair criterion, not a universal natural-failure classifier.

AI USE STATEMENT

During manuscript preparation, the authors used AI-assisted tools for technical support, including code debugging, figure formatting, LaTeX editing and language refinement. The authors retain responsibility for all scientific claims, experimental designs, analyses, interpretations and final verification of the manuscript.

REFERENCES

Guo, Chuan and Pleiss, Geoff and Sun, Yu and Weinberger, Kilian Q. "On Calibration of Modern Neural Networks." *Proceedings of the 34th International Conference on Machine Learning*. pp. 1321–1330. 2017.

Geifman, Yonatan and El-Yaniv, Ran. "Selective Classification for Deep Neural Networks." *Advances in Neural Information Processing Systems*. 2017.

Kull, Meelis and Perello-Nieto, Miquel and Kängsepp, Markus and Silva Filho, Telmo and Song, Hao and Flach, Peter. "Beyond Temperature Scaling: Obtaining Well-Calibrated Multi-Class Probabilities with Dirichlet Calibration." *Advances in Neural Information Processing Systems*. vol. 32. 2019.

Ovadia, Yaniv and Fertig, Emily and Ren, Jie and Nado, Zachary and Sculley, D. and Nowozin, Sebastian and Dillon, Joshua and Lakshminarayanan, Balaji and Snoek, Jasper. "Can You Trust Your Model's Uncertainty? Evaluating Predictive Uncertainty under Dataset Shift." *Advances in Neural Information Processing Systems*. vol. 32. 2019.

Kadavath, Saurav and Conerly, Tom and Askell, Amanda and others. "Language Models (Mostly) Know What They Know." *arXiv preprint arXiv:2207.05221*. 2022.

Liang, Percy and Bommasani, Rishi and Lee, Tony and others. "Holistic Evaluation of Language Models." *Transactions on Machine Learning Research*. 2023.

Ribeiro, Marco Tulio and Wu, Tongshuang and Guestrin, Carlos and Singh, Sameer. "Beyond Accuracy: Behavioral Testing of NLP Models with CheckList." *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. pp. 4902–4912. 2020.

Liu, Yang and Iter, Dan and Xu, Yichong and Wang, Shuohang and Xu, Ruochen and Zhu, Chenguang. "G-Eval: NLG Evaluation using GPT-4 with Better Human Alignment." *Proceedings of EMNLP*. pp. 2511–2522. 2023.

Lin, Stephanie and Hilton, Jacob and Evans, Owain. "TruthfulQA: Measuring How Models Mimic Human Falsehoods." *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*. pp. 3214–3252. 2022.

Min, Sewon and Krishna, Kalpesh and Lyu, Xinxi and Lewis, Mike and Yih, Wen-tau and Koh, Pang Wei and Iyyer, Mohit and Zettlemoyer, Luke and Hajishirzi, Hannaneh. "FActScore: Fine-grained Atomic Evaluation of Factual Precision in Long Form Text Generation." *Proceedings of EMNLP*. pp. 12076–12100. 2023.

Es, Shahul and James, Jithin and Espinosa Anke, Luis and Schockaert, Steven. "RAGAs: Automated Evaluation of Retrieval Augmented Generation." *Proceedings of EACL: System Demonstrations*. pp. 150–158. 2024.

Jacovi, Alon and Goldberg, Yoav. "Towards Faithfully Interpretable NLP Systems: How Should We Define and Evaluate Faithfulness?" *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. pp. 4198–4205. 2020.

Gururangan, Suchin and Swayamdipta, Swabha and Levy, Omer and Schwartz, Roy and Bowman, Samuel R. and Smith, Noah A. "Annotation Artifacts in Natural Language Inference Data." *Proceedings of NAACL-HLT*. pp. 107–112. 2018.

Sainz, Oscar and Campos, Jon Ander and García-Ferrero, Iker and Etxaniz, Julen and de Lacalle, Oier Lopez and Agirre, Eneko. "NLP Evaluation in Trouble: On the Need to Measure LLM Data Contamination for Each Benchmark." *arXiv preprint arXiv:2310.18018*. 2023.

Zheng, Lianmin and Chiang, Wei-Lin and Sheng, Ying and others. "Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena." *Advances in Neural Information Processing Systems*. vol. 36. 2023.

Raji, Inioluwa Deborah and Bender, Emily M. and Paullada, Amandalynne and Denton, Emily and Hanna, Alex. "AI and the Everything in the Whole Wide World Benchmark." *Advances in Neural Information Processing Systems Datasets and Benchmarks*. 2021.

Cronbach, Lee J. and Meehl, Paul E. "Construct Validity in Psychological Tests." *Psychological Bulletin*. vol. 52, no. 4, pp. 281–302. 1955.

Messick, Samuel. "Validity of Psychological Assessment: Validation of Inferences from Persons' Responses and Performances as Scientific Inquiry into Score Meaning." *American Psychologist*. vol. 50, no. 9, pp. 741–749. 1995.

Rasch, Georg. "Probabilistic Models for Some Intelligence and Attainment Tests." Danish Institute for Educational Research. 1960.

Baker, Frank B. and Kim, Seock-Ho (eds.). "Item Response Theory: Parameter Estimation Techniques." Second edition. CRC Press. 2004.

Reckase, Mark D. "Multidimensional Item Response Theory." Springer. 2009.

Tishby, Naftali and Pereira, Fernando C. and Bialek, William. "The Information Bottleneck Method." *Proceedings of the 37th Allerton Conference on Communication, Control and Computing*. pp. 368–377. 1999.

Alemi, Alexander A. and Fischer, Ian and Dillon, Joshua V. and Murphy, Kevin. "Deep Variational Information Bottleneck." *International Conference on Learning Representations*. 2017.

Li, Lihong and Walsh, Thomas J. and Littman, Michael L. "Towards a Unified Theory of State Abstraction for MDPs." *International Symposium on Artificial Intelligence and Mathematics*. 2006.

Ferns, Norm and Panangaden, Prakash and Precup, Doina. "Metrics for Finite Markov Decision Processes." *AAAI Conference on Artificial Intelligence*. pp. 950–951. 2004.

Abel, David and Arumugam, Dilip and Lehnert, Lucas and Littman, Michael. "State Abstractions for Lifelong Reinforcement Learning." *Proceedings of the 35th International Conference on Machine Learning*. vol. 80, pp. 10–19. PMLR. 2018.

Blackwell, David. "Equivalent Comparisons of Experiments." *Annals of Mathematical Statistics*. vol. 24, no. 2, pp. 265–272. 1953.

Lehmann, Erich L. and Romano, Joseph P. "Testing Statistical Hypotheses." Springer. 2005.

Cover, Thomas M. and Thomas, Joy A. "Elements of Information Theory." Wiley. 2006.

Pukelsheim, Friedrich. "Optimal Design of Experiments." SIAM. 2006.

Chaloner, Kathryn and Verdinelli, Isabella. "Bayesian Experimental Design: A Review." *Statistical Science*. vol. 10, no. 3, pp. 273–304. 1995.

Elfving, Gustav. "Optimum Allocation in Linear Regression Theory." *The Annals of Mathematical Statistics*. vol. 23, no. 2, pp. 255–262. 1952.

Attia, Ahmed and Alexanderian, Alen and Saibaba, Arvind K. "Goal-Oriented Optimal Design of Experiments for Large-Scale Bayesian Linear Inverse Problems." *Inverse Problems*. vol. 34, no. 9, pp. 095009. 2018.

A Supplementary Empirical Validation

The following analyses are moved from the main text because they qualify scope and external validity rather than carry the central Preserve–Recover–Acquire argument.

A.1 Claim-bound repair semantics validation

Claim-bound validation uses structured E: claim, scope, available/used evidence, observation channel, provenance, resources and reference bindings, not score vectors. Enumerating deterministic transitions a_0, a_A, a_S, a_L defines the oracle through $L(E, a)$: post-state admissibility, warrant and provenance deficits plus intervention cost (zero for no change, 0.1 otherwise), independent of assigned labels.

Table 1 reports mean regret across 108 states and eight seeds. Using known interface and measurement diagnostics, the gate selects Redesign (richer exposure of the same noisy evidence) or Acquire (more measurements through the unchanged interface) before physical repair selection. The gate keeps the interface when measurement-compatible states share a common optimum; this rule includes 54 tied-optimum states. The protocol specifies state grammar, exposure, transitions, loss, sensors and tie handling.

These repair-semantics results use three synthetic templates sharing one grammar; sensor repeats are not independent model samples.

Table 1: **Claim-bound repair validation.** Descriptive mean intervention regret under the specified finite-world contract.

Regime	Gate	Always Acquire	Always Redesign
Semantic compression	0.0105	0.0954	0.0105
Precision insufficiency	0.00112	0.00112	0.0105
Sufficient control	0	0	0

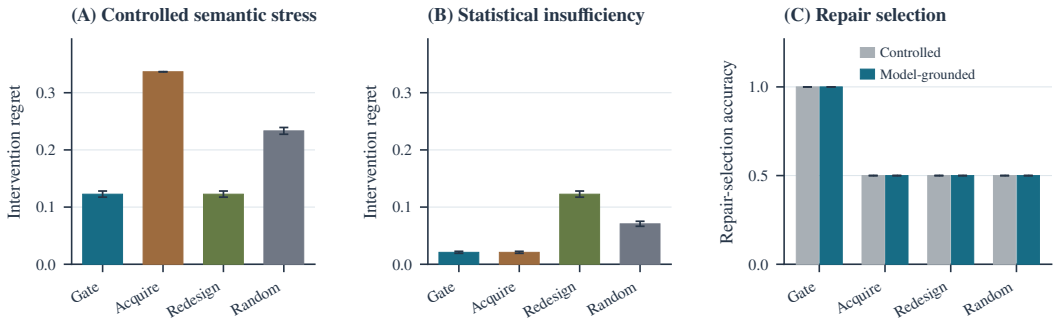


Figure 6: **Model-grounded controlled repair-selection stress test.** (A) Imposed semantic corruption and (B) measurement insufficiency reproduce the repair asymmetry. (C) Repair-selection accuracy alongside the controlled study. Regret uses 14 complete archived states; repair labels use 12 nonzero-margin states. Whiskers resample 30 sensor/policy seeds conditional on this fixed archive. Ten incomplete states remain recorded.

A.2 Model-Grounded Controlled Repair-Selection Stress Test

Does repair asymmetry persist on model-generated substrates? We reuse 24 archived initial Gemma/Qwen responses from 12 families. Suppressing A/S while retaining F/G/L creates action-conflicting cells without changing underlying states. Statistical stress preserves the contrast $(S - A)/\sqrt{2}$ but imposes measurement insufficiency. Both branches use the Gaussian sensor, gate, repair policies and loss of Section 6.3.

Across 30 seeds, regret uses 14 complete states from eight families; repair accuracy uses 12 nonzero-margin states from seven families. Ten incomplete states remain unimputed; two exact ties have undefined repair labels. Repetitions add no independent model states. Semantic Gate/Redesign regret is 0.123 versus 0.337 for Acquire; statistical Gate/Acquire regret is 0.021 versus 0.123 for Redesign (Fig. 6). Gate has zero wrong repairs and 100% repair accuracy on defined cases. Overall Gate regret is 0.072 [0.069, 0.075], versus 0.179, 0.123 and 0.152 for the three comparators. The 95% whole-seed intervals are conditional on the fixed available archive; all-planned missingness bounds overlap (Appendix K.5).

A.3 Real-Model Qualification and Boundaries

Across 360 rows, 132 items and 33 families, conditional-instrument separation holds on system-specific complete observations, but is unresolved on common support. Construct responses are role-sensitive but coupled. Only 157/396 surface pairs jointly observe A/S/G/L; authority assessment is correct in 3/47 scoreable responses, native production in 0/71 (qualification, response and availability: Appendices G–I and K.6).

A.4 Measurement-to-Intervention Boundary

Across three protocols, descriptive FASGL action agreement is 41.7%, 50.0%, and 66.7%, versus Random at 22.2%, 15.4%, and 41.7%. With validated operators, Oracle/FASGL/Random match pre-specified repairs on 24/24, 16/24, and 10/24 cases, supporting intervention relevance beyond random selection.

Distinct endpoints remain: selected repair yields 6/8 known in-domain epistemic recoveries and 2/9 transfer recoveries, versus task recovery of 0/20 and 0/30. The confidence-calibrated gate selects no actions among 540 held-out cases; realized action utility remains unassessed. Appendix J details endpoints, nulls and calibration.

B Formal Theory and Proofs

B.1 Semantic Sufficiency and Epistemic Aliasing

Decision sufficiency is characterized by the partition condition $Q_D \preceq Q_r$. Its consequence is Theorem 1.

Proof of Theorem 1. Condition on $Z = z$. Because E_1 and E_2 share the same representation, any deterministic Z-measurable policy chooses one action $a = \pi(z)$ for both. If $a = a_1^*$, state E_2 incurs excess loss at least Δ_2, contributing at least $\mu_2\Delta_2$. If $a = a_2^*$, the analogous contribution is at least $\mu_1\Delta_1$. Any other action is suboptimal for both and incurs at least $\mu_1\Delta_1 + \mu_2\Delta_2$. Taking the smallest of these cases yields the stated lower bound. A randomized policy is a convex combination of deterministic actions and cannot reduce the minimum below the best deterministic case. □

Noisy representation extension. If the observation distributions generated by two incompatible states are not identical but have total-variation distance $\delta < 1$, no representation-based decision rule can perfectly distinguish them. For equal state priors and action gap at least Δ, the binary testing error is at least $(1-\delta)/2$, giving excess risk at least $\Delta(1-\delta)/2$. Thus exact aliasing is the endpoint of a continuous indistinguishability-regret relation.

B.2 Local Instrument Separability

Proof. For any $v \neq 0$:

$$v^\top G_E v = (J_{\text{item}} v)^\top W (J_{\text{item}} v).$$

Since $W \succ 0$ (non-degenerate items), this is positive iff $J_{\text{item}} v \neq 0$. Thus $G_E \succ 0$ iff J_{item} has trivial nullspace, i.e., $\operatorname{rank}(J_{\text{item}}) = K$. □

B.3 Scalar Insufficiency Corollary

If there exist $\mathsf{E}^{(1)}, \mathsf{E}^{(2)}$ such that $s(\mathsf{E}^{(1)}) = s(\mathsf{E}^{(2)})$ but $\mathcal{D}_{\text{dec}}(\mathsf{E}^{(1)}) \neq \mathcal{D}_{\text{dec}}(\mathsf{E}^{(2)})$, then the scalar representation s is not decision-sufficient for $\mathcal{D}_{\text{dec}}$.

B.4 Measurement Geometry and Precision

For a regular model with calibrated item parameters, $G_E(\theta) = J^\top W J$. Under standard asymptotic efficiency, $\operatorname{Cov}(\hat{\theta}) \approx G_E^{-1}$. Thus:

$$\lambda_{\max}(\operatorname{Cov}(\hat{\theta})) \approx \frac{1}{\lambda_{\min}(G_E)}.$$

$\lambda_{\min}(G_E) = 0$ implies non-identifiability; small positive $\lambda_{\min}$ implies locally identifiable but poorly conditioned measurement; larger $\lambda_{\min}$ indicates improved precision.

B.5 Decision-Relevant Geometry Derivation

Full-rank derivation. When $G_E \succ 0$, regular local asymptotics give

$$\hat{\theta} - \theta \approx N(0, G_E^{-1}).$$

Linearizing the supplied decision rule around θ,

$$h(\hat{\theta}) \approx h(\theta) + \nabla h(\theta)^\top (\hat{\theta} - \theta).$$

With $n = \nabla h / \|\nabla h\|$ and $m = |h(\theta)| / \|\nabla h\|$, the boundary-normal perturbation has variance $n^\top G_E^{-1} n$, yielding

$$P(\text{wrong}) \approx \Phi\left(-\frac{m}{\sqrt{n^\top G_E^{-1} n}} \right).$$

Singular extension. For $G_E \succeq 0$ singular, restrict the local asymptotic model to the identifiable subspace $\mathrm{Range}(G_E)$. If $n \in \mathrm{Range}(G_E)$, the covariance of the estimable contrast is represented by $G_E^\dagger$, so

$$v_D = n^\top G_E^\dagger n, \qquad P(\text{wrong}) \approx \Phi\left(-\frac{m}{\sqrt{n^\top G_E^\dagger n}}\right).$$

If $n \notin \mathrm{Range}(G_E)$, the decision-critical contrast is not estimable; we therefore define its decision variance as $+\infty$ rather than interpreting the pseudoinverse quadratic form as zero. This yields Proposition 1.

B.6 The Three-Part Validity Decomposition

Principle 1 (Decision-Sufficient Epistemic Measurement Conditions)**.** *A measurement is adequate for a specified downstream decision only if three requirements are distinguished:*

Condition 1 — Construct Validity (empirical): *variation attributed to each construct corresponds to reproducible epistemic differences.*

Condition 2 — Semantic Sufficiency (formal): *$Q_{\mathcal{D}_{dec}} \preceq Q_r$, so the representation does not merge states requiring incompatible actions.*

Condition 3 — Decision-Relevant Recoverability (formal/local): *the decision-critical direction is estimable, $n \in \mathrm{Range}(G_E)$, and sufficiently precise. For target local error tolerance $\varepsilon < 1/2$,*

$$\Gamma_D = \frac{m}{\sqrt{n^\top G_E^\dagger n}} \geq \Phi^{-1}(1-\varepsilon).$$

Full-vector local identifiability, $\mathrm{rank}(J_{\mathrm{item}}) = K$, is a stronger global condition and is not required when the decision depends only on an identifiable subspace.

Semantic insufficiency is a Gate 1 failure and requires representation redesign. If Gate 1 passes but $n \notin \mathrm{Range}(G_E)$, acquisition must add information spanning the decision-critical direction. If $n \in \mathrm{Range}(G_E)$ but Γ_D is too small, acquisition should sharpen precision along that direction. More samples can repair statistical insufficiency, but cannot reconstruct a distinction already erased by the representation.

C Singular Information and Decision-Subspace Recoverability

For singular G, the quadratic form $n^\top G^\dagger n$ is interpreted as finite decision variance only after checking estimability. With

$$G = \begin{bmatrix} 10 & 0 \\ 0 & 0 \end{bmatrix}, \qquad G^\dagger = \begin{bmatrix} 0.1 & 0 \\ 0 & 0 \end{bmatrix},$$

$n_1 = (1,0)^\top \in \mathrm{Range}(G)$ has finite local variance 0.1. In contrast, $n_2 = (0,1)^\top \notin \mathrm{Range}(G)$; the pseudoinverse quadratic form alone would misleadingly return zero, so decision variance is defined as $+\infty$ outside the estimable subspace. This is the corresponding estimability condition in singular c-optimal design.

D Active-Acquisition Robustness

The acquisition study includes three robustness controls, each using 20 repetitions of 100 histories per condition and method. They share the population law and true task but use independent study random streams; pairing is maintained within each comparison. The reported ranges are missing-outcome bounds $[W/N, (W+M)/N]$ with $N = 2{,}000$ planned histories rather than confidence intervals.

Estimated versus oracle item weights: the MAP-based acquisition rule computes $\hat{w}_j = \hat{p}_j(1-\hat{p}_j)$ from the current MAP estimate $\hat{\theta}_t$ and updates it after each response; the oracle arm replaces acquisition weights with true-θ likelihood weights while leaving the MAP estimator and evaluator unchanged. MAP/oracle error bounds are $[0.165, 0.170]$ / $[0.167, 0.170]$ at 20 additional items and $[0.129, 0.138]$ / $[0.125, 0.133]$ at 40. Both arms retain their finite-sample advantage over A-optimal, E-optimal, and Random at these endpoints; statistical equivalence is not established.

Random-pool robustness: 20 independently generated base pools are transformed across $\kappa = 1, 2, 5, 10, 20$ with randomized orientation, the specified one-strong/four-equal spectrum, and fixed expected trace, without post-hoc normalization. Here κ is the target population second-moment eigenvalue ratio, not realized Fisher conditioning. At budget 20, decision-aware bounds in that order are $[0.174, 0.177]$, $[0.172, 0.177]$, $[0.195, 0.196]$, $[0.220, 0.222]$, and $[0.243, 0.245]$. Corresponding A-optimal bounds are $[0.225, 0.231]$, $[0.219, 0.222]$, $[0.245, 0.246]$, $[0.269, 0.271]$, and $[0.269, 0.271]$; E-optimal bounds are $[0.241, 0.244]$, $[0.248, 0.251]$, $[0.261, 0.265]$, $[0.269, 0.272]$, and $[0.272, 0.276]$; Random bounds are $[0.234, 0.237]$, $[0.243, 0.246]$, $[0.257, 0.260]$, $[0.264, 0.266]$, and $[0.284, 0.287]$. Robustness is evaluated over the declared pool family.

Normal misspecification: only the acquisition normal is perturbed through $0°$, $15°$, $30°$, $45°$, $60°$, and $75°$; the true boundary, labels, and evaluator remain fixed within the paired comparison. At budget 20, error bounds are respectively $[0.154, 0.159]$, $[0.171, 0.174]$, $[0.190, 0.195]$, $[0.210, 0.215]$, $[0.242, 0.247]$, and $[0.262, 0.266]$. The observed degradation supports dependence on correct acquisition geometry. Across the three studies, 255 MAP failures and 3,580 unavailable later checkpoints are included in the missingness accounting; no observations are imputed.

D.1 Proof of Proposition 2

Let $v_j = n^\top (G_t + G_j)^{-1} n$. Under Proposition 1,

$$P_j(\text{wrong}) \approx \Phi\left(-\frac{m}{\sqrt{v_j}}\right).$$

For fixed $m > 0$, this expression is strictly increasing in v_j. Therefore the item minimizing v_j minimizes the one-step local approximation to wrong-decision probability. The result is conditional on correct local specification of the boundary normal and item information contributions and does not imply global optimality of repeated greedy selection when those quantities evolve.

E FASGL Measurement Specification

E.1 Claim-Level Epistemic State

For each claim j, the epistemic state is:

$$\mathsf{E}_j^{epi} = (F_j, A_j, S_j, G_j, L_j) \in \mathcal{E}$$

where each component is a normalized measurement coordinate $q_k \in [0, 1]$:

- F_j: Formation quality (claim specification, variables, referents, scope)
- A_j: Admissibility quality (can observation establish required distinctions)
- S_j: Support quality (did realized evidence warrant the claim)
- G_j: Attribution quality (who/what warrants it)
- L_j: Lineage quality (can origin be reconstructed)

Table 2: FASGL roles (one empirical instantiation of $K = 5$).

Role	Definition
Formation (θ_F)	Ability to formulate a claim with correct variables, referents, scope, granularity, and semantic resolution
Admissibility (θ_A)	Ability to assess whether the available observation mechanism can distinguish the distinctions required by the claim
Support (θ_S)	Ability to marshal realized evidence whose strength (quantity, diversity, reliability) is sufficient to warrant the claim
Attribution (θ_G)	Ability to correctly identify the source or authority that warrants the claim
Lineage (θ_L)	Ability to maintain and reproduce a complete, auditable provenance chain

For failure profiles we define $r_k = 1 - q_k$ (higher = more failure).

The certification and promotion are derived policy outputs:

$$C_j = \kappa_\phi(\mathsf{E}_j^{epi}), \qquad P_j = \pi_\psi(C_j, \rho_j, \xi_j).$$

Importantly, C and P are not latent competencies; they are policy-driven outputs. Changing the certification policy does not change the system's epistemic competence.

E.2 Latent Epistemic Competence

The measurement target is system-level latent competence:

$$\boldsymbol{\theta}_m^{\text{epi}} = (\theta_1, \ldots, \theta_K) \in \Theta_E \subseteq \mathbb{R}^K.$$

Each θ_k represents the system's capability to reliably satisfy epistemic role k.

We instantiate $K = 5$ using five FASGL roles (Formation, Admissibility, Support, Attribution, Lineage). The FASGL roles are distinct measurement dimensions, not assumed statistically independent, orthogonal, or causally independent. We allow $\boldsymbol{\theta} \sim \mathcal{N}(0, \Sigma_\theta)$ with non-zero off-diagonal terms. Full FASGL definitions are provided in Appendix E.

E.3 Measurement Instrument

The instrument consists of calibrated items:

$$\mathcal{I} = \{(x_j, \mathbf{a}_j, b_j, O_j)\}_{j=1}^N$$

where x_j is the item content, $\mathbf{a}_j \in \mathbb{R}^K$ is the discrimination vector, b_j is the difficulty parameter, and O_j specifies the oracle/annotation protocol.

For binary responses, the theoretical measurement model (MIRT) is:

$$P(Y_{mj} = 1 \mid \boldsymbol{\theta}_m) = \sigma(\mathbf{a}_j^\top \boldsymbol{\theta}_m - b_j)$$

where $\sigma(z) = 1/(1 + e^{-z})$.

Calibration. The general model requires explicit scale-identification conventions, such as a location constraint ($\mathbb{E}[\theta] = 0$) and fixed-sign anchors. The empirical instruments fix item-role assignments and calibrate role-specific difficulties or partial-credit thresholds on their declared calibration populations; evaluation does not learn unrestricted cross-role loadings. During evaluation, calibrated parameters and estimator conventions are fixed; only $\boldsymbol{\theta}_m$ is estimated. All Fisher-information and local-identifiability statements condition on calibrated item parameters.

Operational measurement and observation validity. The final calibrated instrument is a conditional, prior-regularized measurement model. Formation has ordinal scores 0–5, Admissibility 0–3, and Support, Attribution and Lineage are binary when available. Its specified likelihood factors as $P(F)P(A)P(L)P(S)P(G \mid S)$, with

$$\operatorname{logit} P(G = 1 \mid S = s) = \theta_G - b_G + \beta(2s - 1).$$

The five roles remain distinct; $G \mid S$ denotes conditional interpretation of the existing Attribution role, not an additional dimension or unconditional authority competence. The conditional calibrated model uses the specified MAP estimator and proper priors with full joint local-Laplace covariance, including item and β uncertainty. The fixed-item binary model above is the theoretical geometry setting, not a claim that this empirical ordinal/conditional model is identical to it. Previous instrument variants are not assumed measurement-equivalent to the conditional calibrated model.

An LLM response becomes an observation only after parsing, representation validation and consistent execution binding. Valid but scientifically incorrect responses remain observations. A missing substantive claim, unavailable evidence or invalid representation can leave a coordinate null; null is neither a zero score nor a repaired response. Attribution interpretation retains the observed S branch, and missing provenance is not automatically Lineage failure.

Estimates are accompanied by assigned, completed, parsed and available counts; coverage by family, checkpoint and condition; reasons for unavailable observations; and controller, prompt and scorer specifications. Qualification describes the conditions of measurement rather than replacing the score. Joint calibration uncertainty, conditional family stability and response availability are reported separately; missingness remains potentially informative and cross-population comparability unestablished.

F CONTROLLED INTERVENTION CONSTRUCTION

F.1 INTERVENTION TAXONOMY

Pure interventions target one FASGL role, surface controls change presentation while preserving declared item-state properties, and compound interventions target multiple roles. All variants preserve their source-family assignment and condition definitions.

Each base item x_j contains:

$$x_j = (c_j, e_j, O_j, s_j, g_j, l_j, y_j^{GT})$$

where c_j is the claim, e_j the evidence, O_j the observation mechanism, s_j the support structure, g_j the source attribution, l_j the lineage trace, and y_j^{GT} the ground truth.

The role-targeting hypotheses concern observed responses rather than assumptions imposed by construction:

1. **Pure interventions:** $\mathcal{T}_k$ targeting role k should primarily affect r_k (failure indicator).
2. **Surface controls:** Surface changes (wording, style) should not materially alter the measured claim-level epistemic response.
3. **Confounded interventions:** $\mathcal{T}_{i+j}$ should produce expected multi-dimensional response patterns.

F.2 CONSTRUCTION VALIDATION

Construction validation compares clean and transformed states while preserving source-family pairing, task truth, and required non-target epistemic properties. Warrant and provenance validation, including 216 lossless Lineage serialization assessments, establishes consistency of the stated references and representations; publisher authenticity and invariance of generated responses remain outside these checks. Pairing and resampling account for shared source families and physically identical items.

Definition 1 (Construct Validity)**.** *A measurement construct is valid only if variation attributed to that construct corresponds to reproducible epistemic differences rather than superficial, benchmark-specific, or evaluator-specific artifacts.*

Construction validation and construct response are separate. Item-state checks certify the specified finite-state contract but do not replace empirical construct evaluation. Construct validity requires empirical evidence for the specified measurement object, while semantic sufficiency and decision-relevant recoverability are formal necessary conditions for the specified downstream decision; full-vector identifiability is a stronger global condition. Unavailable coordinates remain null, not zero.

Table 3: FASGL roles and intervention types used in controlled construction.

Role	Intervention Type	Operationalization
Formation	Pure ($\mathcal{T}_F$)	Precise entity $\rightarrow$ ambiguous; numeric $\rightarrow$ vague
	Surface control ($\mathcal{T}_F^{surf}$)	Wording/style change, epistemic property unchanged
Admissibility	Pure ($\mathcal{T}_A$)	$\exists x_1, x_2 : c(x_1) \neq c(x_2)$ but $O(x_1) = O(x_2)$
	Surface control ($\mathcal{T}_A^{surf}$)	Observation description style change
Support	Pure ($\mathcal{T}_S$)	Reduce evidence quantity, diversity, or reliability
	Surface control ($\mathcal{T}_S^{surf}$)	Evidence presentation style change
Attribution	Pure ($\mathcal{T}_G$)	Attribute to wrong or unverified source
	Surface control ($\mathcal{T}_G^{surf}$)	Source formatting change
Lineage	Pure ($\mathcal{T}_L$)	Break or truncate provenance chain
	Surface control ($\mathcal{T}_L^{surf}$)	Lineage presentation style change

F.3 Evaluation Scope and Estimands

Controlled tasks span RAG/factual QA, observation-limited reasoning and agent/tool tracing, with Gemma-2 9B IT, Qwen2.5 7B Instruct and Mistral 7B Instruct as the target checkpoints. Calibration uses separate systems. Study populations and administrations are protocol-specific rather than a common twelve-system crossing.

The conditional calibrated measurement model uses 360 rows from 33 declared families; profile and surface analyses use their specified target-item frame; the observed role-response analysis crosses six source families per mechanism with three checkpoints; and the authority study uses 12 source families. Confidence calibration uses 60 disjoint fitting and evaluation families. These populations retain their corresponding measurement specifications and source/estimator assignments. Repeated checkpoint administrations, physical aliases and policy-shared actions are not independent family counts.

Constructed variants are analyzed as generated conditions rather than independent acquired observations.

For pure intervention $\mathcal{T}_k$, standardized magnitudes and descriptive selectivity/leakage quantities are defined as:

$$z_\ell^E = \frac{|\Delta r_\ell|}{\sigma_\ell^E}, \qquad SI_E(\mathcal{T}_k) = \frac{z_k^E}{\sum_{\ell=1}^K z_\ell^E + \varepsilon}, \qquad L_k = \frac{\sum_{\ell \neq k} z_\ell^E}{z_k^E + \varepsilon}.$$

For confounded intervention $\mathcal{T}_\mathcal{K}$:

$$L_\mathcal{K} = \frac{\sum_{\ell \notin \mathcal{K}} z_\ell^E}{\sum_{k \in \mathcal{K}} z_k^E + \varepsilon}.$$

These definitions do not impose independence, orthogonality, or additivity. Standardized response summaries retain their reference-scale limitations. The current native role-response analysis instead reports signed, normalized observed-coordinate contrasts and availability; it does not define a new standardized selectivity measure. Compound response-level summaries are distinguished from $L_\mathcal{K}$. Compound response-level summaries therefore do not validate the current warranting-authority observable.

G Measurement Qualification and Identifiability

Table 4: **Measurement qualification and coverage geometry.** Coverage-dependent profile separation is compared with the local 95% posterior radius; precision-scaling coefficients summarize total and weakest-direction behavior. Conditioning statistics summarize the finite-design diagnostics used for the geometry analysis.

Diagnostic	Estimate	Reference / uncertainty
Common-item profile distance	2.697	local radius: 3.327
Common-family profile distance	3.032	local radius: 3.327
System-specific complete distance	6.011	local radius: 3.327
Equal-family-weighted distance	6.214	local radius: 3.327
Total precision-scaling slope	0.992	95% CI [0.980, 1.004]
Weakest-direction slope	0.981	95% CI [0.970, 0.992]
Finite conditioning cases	71,941 / 72,000	59 non-finite
Bootstrap replicates	399	conditioning analysis

This appendix presents conditional instrument qualification, coverage-conditioned profile comparison, and information-geometry precision analyses.

G.1 Conditional Qualification

Conditional qualification evaluates 360 matched-execution rows, 132 source-specific items and 33 declared families from Phi-3.5 Mini Instruct, SmolLM2 1.7B Instruct and StableLM2 Zephyr 1.6B.

Table 5: Held-out predictive checks for the calibrated instrument.

Domain	Test rows	Domain-held-out NLL	Family-held-out NLL
Agent/tool	120	442.80	388.75
Observation	120	419.59	348.95
RAG	120	439.53	387.46

Each domain fit uses 240 training and 120 held-out rows. The two holdout procedures use different training information; these are prediction diagnostics, not domain-invariance tests.

Under the specified conditional, prior-regularized model, family-held-out total negative log likelihood is 1125.16, including joint S/G NLL 277.65; combined domain-held-out NLL is 1301.91. Table 5 gives the domain-specific held-out prediction diagnostics.

The fitted dependence parameter is $\beta = 4.742$, with local 95% posterior interval [3.522, 5.961]. Observed Attribution success occurs in 0/233 Support-zero rows and 121/127 Support-one rows. With no observed S=0/G=1 success, the direction $\theta_G \mapsto \theta_G - t$, $\beta \mapsto \beta + t$ improves the unpenalized likelihood toward a limit; proper priors keep the MAP finite. Uncertainty propagation uses the full joint 702-parameter local covariance, conditional branch uncertainty and fixed S-state reference. Prediction for unseen items and domains integrates over these fitted models without refitting.

Interpretation is conditional on this calibration population, fitted model and local uncertainty procedure. Neither β nor positive posterior curvature establishes causality, semantic validity, universal calibration or likelihood-only identification. Local intervals do not cover model selection or arbitrary posterior tail behavior, and native diagnosis alone does not establish latent transport.

G.2 Coverage-Conditioned Profile Separation

The fixed Gemma Base–Qwen Retrieval comparison separates under system-specific complete observations: 32 and 41 rows out of 132 assigned per system yield a covariance-standardized distance of 6.011 against the joint 95% local Gaussian radius 3.327. Equal-family weighting gives 6.214. Separation is not established on common items (15/15 rows; distance 2.697) or common families (19/20 rows; 3.032), and full-assignment separation is not estimated (Table 6).

Table 6: Profile separation under different observation-support populations.

Observation population	Rows Gemma/Qwen	Standardized distance	Origin excluded
Full assigned	132/132	Not estimated	Not established
System-specific complete	32/41	6.011	Yes
Common items	15/15	2.697	No
Common families	19/20	3.032	No
Equal-family weights	32/41	6.214	Yes

The inherited joint 95% local Gaussian radius is 3.327. Distances use scenario-specific covariance; they are not raw effect sizes. Common-support non-separation is not evidence of equality.

The comparison propagates the full 702-parameter shared calibration covariance through local-Laplace/joint-delta uncertainty propagation. G is a same-S conditional contrast, with the shared β term cancelling; F is at the observed upper category, so its finite estimate is prior-regularized. Scenario-specific covariance, composition and prior influence change together, so the reported distances are conditional on their respective observation populations.

Observation availability defines the comparison population. Gemma has F available on all 132 rows but A/S/G/L only on 32; Qwen has F on 119 and the remaining coordinates on 41. Gemma has 100 and Qwen 78 no-substantive-claim rows; 13 Qwen representations are invalid. Coverage counts include all of these cases. Common-item analysis contains 15 items from seven families; common-family analysis contains eight families with residual within-family imbalance. Neither defines a population independent of observation availability.

All 2,000 family-resampling replicates and 33 leave-family-out diagnostics retain separation within the observed-mask scope. These diagnostics establish only stability within the observed-mask pat-

tern, not a full-assignment effect or exact posterior/frequentist coverage. The exploratory historical outcome comparison is reported separately and is not used for current qualification. The combined evidence supports coverage-conditioned interpretation rather than universal ranking or population-level outcome equivalence.

G.3 Geometry and Identifiability

For the fixed logistic instrument, $G_E = J^\top W J$ links the item Jacobian to measurement precision. Known-truth simulations use 80 independent blocks of 100 latent vectors, ten angular designs, 200 items and two response replications per item; information includes both replications. Conditional on finite full-rank estimates, total squared error scales approximately with $\mathrm{tr}(G_E^{-1})$ (slope 0.992, 95% whole-block interval [0.980, 1.004]), and weakest-direction squared error with $1/\lambda_{\min}(G_E)$ (0.981, [0.970, 0.992]; Table 7).

Table 7: Conditional error scaling with information geometry.

Squared-error endpoint	Predictor	Slope	95% block interval
Total error	$\mathrm{tr}(G_E^{-1})$	0.992	[0.980, 1.004]
Weakest-direction error	$1/\lambda_{\min}(G_E)$	0.981	[0.970, 0.992]

Both error and predictor means use the same finite-MLE mask; intervals resample independent whole blocks. Approximate scaling does not establish exact finite-sample efficiency.

The archived geometry analysis distinguishes full rank, a weak but identifiable direction, and rank collapse using the same latent vector. A small positive $\lambda_{\min}(G_E)$ limits weakest-direction precision even when full-vector local identifiability holds. At exact rank deficiency, $J_{\mathrm{item}} v = 0$ leaves response probabilities unchanged along $\theta + tv$, so the full vector is not uniquely identified. In the collapsed design, the F–S contrast is lost while an identifiable four-dimensional subspace remains. A posterior prior penalty does not restore likelihood information in the null direction; no finite full-space uncertainty region is assigned.

Full rank supplies local identification under positive response weights, but finite samples can still yield non-finite MLEs. Error and predictor summaries use the same finite-estimate subset; information summaries include all true-parameter cases. Whole-block intervals preserve dependence among angular designs. Separation cases are included in failure counts; the earlier single-response setting did not meet its fixed criterion. These results characterize approximate scaling of conditional synthetic precision; broader finite-sample behavior and diagnosis/repair utility require separate validation.

H Construct Response Evidence

H.1 Response Summaries and Aggregation

Do controlled interventions produce structured responses across epistemic roles? Target-response magnitudes are summarized as complete-pair-weighted means of absolute clean-minus-pure changes, with Formation divided by five and Admissibility by three. Across mechanisms targeting each role, the magnitudes are F: 0.074, A: 0.433, S: 0.690, G: 0.364 and L: 0.313. This descriptive pooling averages within-pair absolute changes rather than taking the absolute value of a signed mean. Different magnitudes characterize response sensitivity under the tested manipulations rather than construct importance, model competence, or epistemic-role ranking.

The mechanism-level matrix includes all 11 pure mechanisms and 396 attempts: 393 generations complete and 391 responses are schema-valid. Among the 198 assigned pairs, 104 have complete five-coordinate measurements (218 individual full vectors for the mechanism-level analysis). Each mechanism crosses six independent source families with three checkpoints; its 18 assigned pairs are not 18 independent families. Signed contrasts $\Delta_k = \overline{q_{k,\mathrm{clean}} - q_{k,\mathrm{pure}}}$ are positive for deterioration and retain negative and off-target movements. Support quantity, diversity and licensing interventions show coupled Support/Attribution responses. These are structured observed-coordinate patterns on mechanism-specific complete-pair populations rather than fitted latent effects or population-level tests; unavailable coordinates remain unavailable.

Target sensitivity coexists with off-target movement. The pooled comparison contrasts mean absolute target change with the within-pair sum of absolute changes in the other four roles, weighted by complete-pair counts. The off-target sum is a descriptive aggregate rather than a per-coordinate dominance test or standardized selectivity measure. Formation's signed effects also change with availability conditioning, and zero signed means can conceal opposing responses. Thus controlled item-state perturbations do not imply isolated measured responses, orthogonal dimensions or an identified independent-factor structure (Appendix H.2).

Paired Support and Attribution responses illustrate this coupling. Among 29 Support pairs with both coordinates available, all 20 Support declines coincide with Attribution declines; the remaining nine show no change in either coordinate. The observed contrast correlation is 1.000 within this selected scope, whose availability condition differs from the complete-five matrix. Co-variation establishes a response pattern, while construct identity, causal mediation and independent latent axes require additional evidence. The evidence supports role-sensitive response structure with coupling, while complete-panel latent selectivity is not established.

H.2 Detailed Pure-Intervention Responses

Clean/pure source-family bindings are preserved. Null and failed responses remain part of the measurement analysis; missing coordinates are neither treated as zero nor imputed from their partners.

The mechanism-level response analysis uses the common complete-five-pair population within each mechanism. The four Formation target means on that population are 0, -0.020, 0 and -0.075; corresponding F-available-pair means are approximately $+0.022$, $+0.038$, $+0.022$ and $+0.012$. Changing availability conditioning changes the observed pattern and must not be described as a uniform absence of Formation response. Support interventions produce coupled S/G declines. Admissibility changes only its targeted coordinate in nine of ten complete pairs, while the Lineage mechanisms show different magnitudes and off-target movements. Descriptive means may conceal opposing within-pair changes.

Historical observable-contract analyses are reported separately and are not used for current construct-response inference.

I Operational Boundaries

I.1 Surface Sensitivity and Availability Boundary

Matched JSON-format perturbations reveal both latent-distance and availability changes. The conditional calibrated measurement model compares original and matched reformatted administrations on the same item frame using the Euclidean latent-distance endpoint. The available-channel median is 1.702 (95% interval [0, 4.349]); the complete-five restriction gives 1.076 [0, 3.182] (Table 8). Practical equivalence requires an independently justified margin on this latent scale; neither calibration uncertainty, score-category resolution nor the historical raw-score margin supplies one.

Table 8: Surface sensitivity on the calibrated latent scale.

Observation panel	Median distance	95% interval	Equivalence
Available-channel	1.702	[0.000, 4.349]	Margin absent
Complete-five pairs	1.076	[0.000, 3.182]	Margin absent

Intervals refer to the Euclidean latent-distance endpoint. The historical raw-score reporting margin cannot be transported to this latent scale.

Among 396 assigned pairs, A/S/G/L are available on both sides in 157, reference-only in 58, surface-only in 46, and neither in 135. These transitions describe measurement support rather than correctness. Checkpoint shifts have opposing signs, which pooled summaries can conceal (Table 9).

I.2 Warrant-Level Attribution Boundary

The authority-aligned pilot binds a warranting-authority oracle separately from evidence-provider and recorder identities. Its 12 source families span two authority mechanisms, three domain framings and two basis templates; direct assessment and native claim production are independent requests. Direct assessment is correct in 3/47 scoreable responses, with 25 null abstentions, including

Table 9: Availability transitions under matched JSON formatting.

A. Availability composition

Checkpoint / channels	Both	Reference-only	Surface-only	Neither
All / F	379	4	13	0
All / A,S,G,L	157	58	46	135
Gemma / A,S,G,L	29	3	33	67
Mistral / A,S,G,L	110	20	0	2
Qwen / A,S,G,L	18	35	13	66

B. Availability shift

Checkpoint / channels	Shift (pp)	95% interval (pp)
All / F	2.273	[-0.070, 4.545]
All / A,S,G,L	-3.030	[-8.665, 2.914]
Gemma / A,S,G,L	22.727	[14.441, 32.492]
Mistral / A,S,G,L	-15.152	[-24.658, -6.619]
Qwen / A,S,G,L	-16.667	[-29.659, -4.112]

Availability means observed, including valid zero; it is not correctness. Intervals are pointwise basic intervals using the archived family resamples. A/S/G/L share the availability pattern, not necessarily the score.

all 24 Gemma assessments. Native production is correct in 0/71, with one Qwen native JSON invalid and unscored. All 144 attempts complete and 143 responses are schema-valid (Table 10). None of 35 scoreable native model-family pairs tracks both authority states: selection is unchanged, with 69 responses selecting evidence provider E and two recorder R. Provenance-role selection therefore characterizes the tested operationalization of authority attribution rather than general attribution capability.

Table 10: Warrant-level authority identification in direct and native responses.

Checkpoint	Direct correct/scoreable	Direct null	Native correct/scoreable	Native invalid
Gemma	0/0	24	0/24	0
Qwen	0/23	1	0/23	1
Mistral	3/24	0	0/24	0
Qwen 32B (separate cohort)	14/22	2	0/24	0

Correct/scoreable denominators exclude null or invalid responses, reported separately. Direct 0/0 is undefined. The separate cohort has only 2/10 fully scoreable families tracking both authority states; full control/switch and provenance accounting is in Appendix I.2.

Separate auxiliary cohort. The separate Qwen2.5-32B-Instruct cohort provides an auxiliary format/checkpoint sensitivity analysis (Table 10). The auxiliary cohort provides sensitivity evidence across formats and checkpoints while preserving the primary authority-boundary interpretation. Observed differences across formats, executions and checkpoints do not identify causal effects, and aggregate direct accuracy does not establish robust authority tracking.

Provenance capture failures are zero in both cohorts. In the primary native cohort, 57 L=0 values reflect missing claim–evidence declarations, 14 have L=1, and one is unscored; these reflect missing claim–evidence declarations rather than capture failures. Opposing auxiliary Lineage changes can cancel in aggregate, so a zero mean does not imply zero leakage. All models, null cases and unfavorable responses are included in the analysis.

J Intervention Relevance and Utility Boundary

This appendix separates descriptive action agreement, recovery under validated input-state operators, and the diagnosis-to-task-utility boundary.

J.1 Action Agreement and Intervention Relevance

Deficit-based, actionability-weighted and explicit repair-operator studies provide descriptive action-selection evidence (Table 11). The initial study uses 18 fresh families (A: 4, S: 4, G: 3, L: 3, clean: 4), six per domain. The weighted study uses 13 fresh families (A: 4, S: 3, L: 3, clean: 3), allocated 5/4/4 to RAG/Observation/Agent. The repair-operator study uses 12 fresh families (A/S/L: four each), four per domain. Every family is evaluated with Gemma and Qwen; checkpoint outputs and action branches are dependent observations within a source family.

Table 11: Action agreement with pre-specified Oracle assignments.

Study	Families	Conditions	FASGL	Random
Initial mapping	18	A/S/G/L, clean	41.7%	22.2%
Weighted mapping	13	A/S/L, clean	50.0%	15.4%
Repair operators	12	A/S/L	66.7%	41.7%

Oracle agreement is 100% by construction.

The initial policy selects the largest available response deficit across A/S/G/L; Random samples those four repairs or no change. The weighted policy applies fixed weights 1/1/0.8 to A/S/L; its Random comparator samples A/S/L or no change. The repair-operator study uses $(1 - q_k)c_k$ only where both the coordinate and the operator resources are available, with the same A/S/L compatibility weights; nulls are not zero deficits. Random samples the available repair operators, without no change as a candidate when resources exist. Deterministic tie handling and an explicit no-change fallback for all-unavailable measurements are applied. Initial defects and Oracle actions are fixed before responses, independently of model outputs. Operators act on copies of the same defective state with the same checkpoint and generation budget; evaluation uses the recorded executed states. Families, operators and request outcomes are fixed independently of observed action outcomes.

FASGL agrees with the Oracle assignment in 41.7%, 50.0% and 66.7% of decisions, compared with 22.2%, 15.4% and 41.7% for the corresponding Random policies. Action agreement has descriptive 95% intervals [0.542, 0.833] for FASGL and [0.208, 0.625] for Random in the repair-operator study. Agreement provides intervention-relevant signal within each study. Different populations, eligible roles and Random action sets preclude a causal improvement claim across protocols; response-score recovery ordering and cross-study gap reduction remain unestablished. These studies provide feasibility evidence under their respective protocols rather than a pooled utility estimate.

J.2 Mechanism Recovery Under Validated Operators

The repair-operator study tests whether the measured deficits correspond to actionable input-state mechanisms. The four families per A/S/L role and two checkpoints share initial states across Oracle, FASGL and Random branches. All 96 administrations complete: 24 initial and 72 post-action; 89 outputs are schema-valid and 7 Qwen outputs remain invalid. Input-state mechanism evaluation is defined for 72/72 branches independently of response validity.

Diagnosis uses the unchanged response-dependent q_A, q_S, q_L. Mechanism validation instead uses independent executed-state checks:

$$
\begin{aligned}
M_A &= \mathbf{1}\{O(x_1) \neq O(x_2)\}, \qquad C(x_1) \neq C(x_2), \\
M_S &= \mathbf{1}\{W(C, E^*) = 1\}, \\
M_L &= \frac{\#\{\text{required reference edges present}\}}{\#\{\text{required reference edges}\}}.
\end{aligned}
$$

The A operator adds the prespecified minimal discriminative observation; its certificate concerns two designated alternatives, not all possible native information. The S operator supplies a fixed evidence subset satisfying authority, relevance and distinct primitive acquisition-unit requirements. This is structural source independence, not an empirically validated independence of source errors. The L operator restores missing reference edges; invalid extra branches are not assigned a successful completeness score. Task correctness is evaluated separately. Mechanism recovery $\Delta M = M_{\text{after}} - M_{\text{before}}$ therefore does not substitute for Δq or ΔTask.

Table 12: Mechanism recovery under validated operators.

Role	Oracle ΔM	FASGL ΔM	Random ΔM
A	1.000 [1.000, 1.000]	0.750 [0.500, 1.000]	0.625 [0.250, 1.000]
S	1.000 [1.000, 1.000]	1.000 [1.000, 1.000]	0.250 [0.000, 0.500]
L	0.500 [0.500, 0.500]	0.438 [0.313, 0.500]	0.500 [0.500, 0.500]

Target mechanism recovery in the repair-operator study: mean ΔM with descriptive 95% whole-family bootstrap intervals.

Oracle has positive mechanism recovery in all three roles. FASGL is closer to Oracle than Random for A/S; in L, FASGL recovers 0.438 compared with 0.5 for both Oracle and Random. The executed Support operator also restores provenance relations, so a different selected action can recover Lineage. Exact action agreement and mechanism recovery consequently measure different properties; the Lineage comparison does not establish either isolated-role failure or repair independence. Oracle recovery validates these constructed operators, not learned repair competence or task recovery. The results meet the descriptive feasibility criterion for the tested operators; broader utility requires separate validation.

Mechanism means first average checkpoints within family and then weight families equally within role. The 2,000 whole-family bootstrap draws jointly resample both checkpoints and every policy branch. The mechanism intervals in Table 12 are not simultaneous superiority tests and have limited resolution at four families per role; construction-induced degenerate intervals are not a general certainty claim. Reported summaries use the original estimates and resampling distribution.

J.3 Diagnosis, Repair, and Task Utility Boundary

Descriptive localization. Retrospective diagnosis uses 108 existing initial responses from 36 additional families, without adding any of these families to fitting. The 90 defect responses and 18 clean responses are balanced across domains and checkpoints. Agent transfer uses twelve of those families and is not an independent cohort. The FASGL diagnostic policy (FULL) uses all five roles, with explicit unavailable categories and Support-conditioned Attribution. Its localization is 35.6%; Table 13 compares it with outcome-only, scalar, XGBoost and SBERT-MLP and reports its descriptive family interval. FULL's overall bounded interval is [0.108, 0.604], wider than its descriptive bootstrap interval. All 18 simultaneous FULL-minus-comparator intervals include zero. Clean recognition is 0/18. Localization therefore provides descriptive signal without establishing learned-baseline superiority, calibrated action accuracy or the safety of acting on a diagnosis.

Table 13: Diagnostic localization and clean-control recognition.

Method	Defect localization	Clean recognition	Agent transfer
FULL	35.6%	0.0%	36.7%
Outcome-only	23.3%	0.0%	26.7%
Scalar	16.7%	27.8%	13.3%
XGBoost	12.2%	0.0%	26.7%
SBERT-MLP	12.2%	33.3%	23.3%

Retrospective diagnosis uses 108 responses from 36 additional families (90 defect, 18 clean responses). Agent transfer uses a subset of these families. FULL localization has a descriptive 95% family interval of 23.3–48.9%. All 18 simultaneous bounded FULL-minus-comparator intervals include zero; learned-baseline superiority remains unresolved. These are not repair outcomes.

Availability is full for 44 and partial for 64 additional-cohort episodes, with no all-unavailable episode in that cohort; unavailable episodes remain included in the accounting. FULL defect localization is 20.0% for Gemma, 26.7% for Qwen and 60.0% for Mistral. These strata differ in model/defect composition and do not establish that completeness causes validity. FULL selects correction on all 18 clean controls in this diagnosis-only assessment; no repair harm is measured by that selection count.

Endpoint distinction. Response-dependent diagnosis and task change remain separate from the independently checked input-state mechanisms in Table 12. Fixed-panel task-change means remain undefined in Table 14.

Selected repair and task outcomes. The separate selected-repair evaluation compares diagnosis-selected repair, four fixed baseline policies and a newly sampled no-change control under independent task and protected-state oracles. Identical episode/action administrations are shared across policies while logical policy identities are retained: 54 initial episodes and nine unique clean controls produce 432 policy/track records, not 432 independent episodes. The 206 fresh attempts yield 197 schema-valid outputs and nine retained Qwen failures, without retries or replacements.

Table 14: Intervention endpoints across recovery, task and action boundaries.

Endpoint	Metric	Observed result	Interpretation boundary
Selected repair (in-domain)	Task / epistemic recovery	**FULL**: Task 0/20; unknown 0. Epistemic 6/8; unknown 1.	Instrument-defined recovery is distinct from task recovery.
In-domain		**Outcome-only**: Task 0/20; unknown 2. Epistemic 0/8; unknown 0.	
In-domain		**Scalar**: Task 0/20; unknown 1. Epistemic 0/8; unknown 0.	
In-domain		**XGBoost**: Task 0/20; unknown 1. Epistemic 5/8; unknown 1.	
In-domain		**SBERT-MLP**: Task 0/20; unknown 0. Epistemic 0/8; unknown 0.	
In-domain		**Sham**: Task 0/20; unknown 0. Epistemic 0/8; unknown 0.	
Selected repair (transfer)	Task / epistemic recovery	**FULL**: Task 0/30; unknown 0. Epistemic 2/9; unknown 0.	Separate studies, not a pooled utility effect.
Transfer		**Outcome-only**: Task 0/30; unknown 0. Epistemic 0/9; unknown 0.	
Transfer		**Scalar**: Task 0/30; unknown 0. Epistemic 0/9; unknown 0.	
Transfer		**XGBoost**: Task 0/30; unknown 0. Epistemic 1/9; unknown 0.	
Transfer		**SBERT-MLP**: Task 0/30; unknown 0. Epistemic 0/9; unknown 0.	
Transfer		**Sham**: Task 0/30; unknown 0. Epistemic 0/9; unknown 0.	
Operator task change	ΔTask	**Oracle**: NA; bounds [-0.125, 0.042]; defined pairs 21/24.	Missingness bounds, not confidence intervals; NA is undefined, never zero.
		FASGL: NA; bounds [-0.125, 0.042]; defined pairs 22/24.	
		Random: NA; bounds [-0.083, 0.042]; defined pairs 22/24.	
Action calibration	NLL, raw $\rightarrow$ scaled	In-domain: 1.58 $\rightarrow$1.54; Transfer: 1.73 $\rightarrow$1.70.	Zero selected actions leaves action-region calibration and repair recovery unassessed.
	Brier, raw $\rightarrow$ scaled	In-domain: 0.75 $\rightarrow$0.74; Transfer: 0.81 $\rightarrow$0.80.	
	Calibrated confidence AUROC	In-domain: 0.739; Transfer: 0.643.	
	Availability + deficit eligible	In-domain: 157/540; Transfer: 148/540.	

Continued on next page

Table 14 (continued)

Endpoint	Metric	Observed result	Interpretation boundary
	Final action coverage	In-domain: 0/540; Transfer: 0/540.	
	Clean-control selections	In-domain: 0/90; Transfer: 0/90.	
	Maximum confidence	In-domain: 0.784; Transfer: 0.469.	

Selected repair: FULL's in-domain point rate is undefined: six known epistemic recoveries among eight at risk with one unresolved post state give identification bounds 75–87.5%, not confidence limits. Tracks share episodes and have only two independent test families each. All known epistemic recoveries are in Mistral. FULL clean corrective selections are 5/6 in-domain and 6/6 in transfer; unchanged public fields do not demonstrate state harm.

Operator task change: NA is an undefined fixed-panel mean, never zero. Bounds are saved missingness bounds, not confidence intervals. These are separate studies, not a pooled utility effect; instrument-defined recovery is distinct from task recovery and general semantic improvement. Nulls and failed executions remain included.

Action calibration: Both tracks evaluate the same 540 held-out responses from 30 families. Temperature was fitted only on 30 disjoint calibration families and fixed before held-out evaluation. The action threshold is > 0.80. Zero selected actions leaves action-region calibration and repair recovery unassessed rather than validating selective safety.

FULL's in-domain epistemic recovery is 6/8 known recoveries with one unresolved post state; its point rate is undefined and its identification bounds are 75–87.5%. The transfer rate is 2/9 with a descriptive family-bootstrap interval 0–66.7%, without a nominal coverage guarantee at two families. Task recovery is 0/20 in-domain and 0/30 in transfer; all known epistemic recoveries are in Mistral. FULL selects corrective actions on 5/6 in-domain and 6/6 transfer clean-control records, with zero changed public fields because their resource states were already restored. False intervention selection is not thereby demonstrated state harm. Native recovery and task success remain distinct from transport of the conditional calibrated measurement model and general semantic improvement.

Fresh post-action availability is 69/69 partial Gemma vectors; Qwen has 5 full, 51 partial and 9 unavailable among 65 attempts; Mistral has 72/72 full vectors. All models remain represented. Missing initial or post states do not become failures, and no complete-case model deletion is used.

Action-calibration boundary. The confidence-calibrated action-selection pilot uses 60 fresh source families, split into 30 calibration-fit and 30 held-out evaluation families, each crossed with three checkpoints and six conditions. All 1,080 requests are attempted; four Qwen failures remain unavailable. In-domain temperature 0.607624 and source-domain-only transfer temperature 0.411439 are fixed before held-out evaluation. Both tracks evaluate the same 540 evaluation records, of which three are unavailable executions; the missing-aware diagnosis policy retains them.

Table 14 shows proper-score improvement but no action coverage. Confidence-ranking AUROC has descriptive whole-family 95% intervals [0.703, 0.773] in-domain and [0.597, 0.692] for the pooled transfer model. Agent-only transfer comprises 180 records from ten held-out families: AUROC 0.650, 50/180 preconfidence-eligible and 0/180 selected actions. This reflects ranking agreement with assigned conditions rather than repair success or causal attribution.

Conditional action-rate bootstrap intervals collapse at zero, which does not prove a zero population rate. The two-stage calibration/evaluation bootstrap gives an in-domain action-rate interval [0, 0.061]; selected accuracy is undefined in 7,714/9,999 draws, so no interval is formed by discarding those draws. No held-out confidence exceeds 0.80, including for full-vector Mistral. The result separates confidence calibration from demonstrated repair utility under this fixed action-selection policy. Neither descriptive localization nor input-state restoration alone establishes repair superiority. Measurement-guided intervention remains distinct from demonstrated task utility.

K Reproducibility and Scope

K.1 Fixed Analysis Specification

Each study uses fixed source-family, item, checkpoint and condition assignments, together with specified controllers, scorers, references and calibration models. Analyses account for family-level

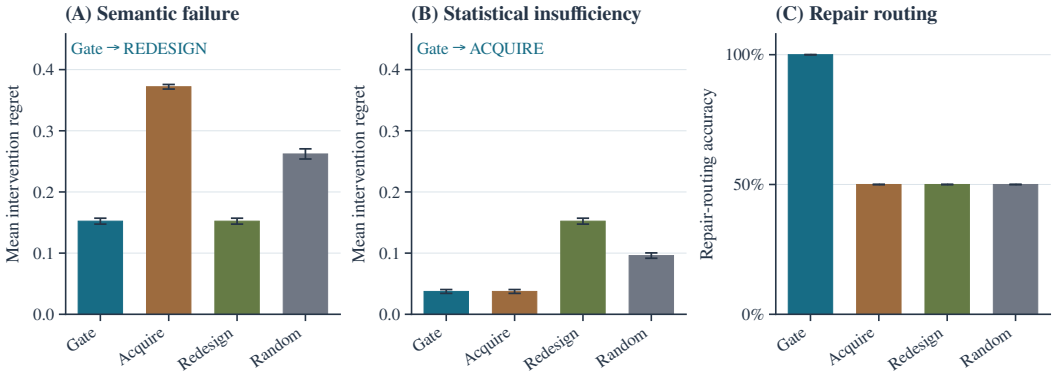


Figure 7: **Known-contract repair-selection sanity check.** (A) Redesign reduces regret after semantic erasure; Acquire does not. (B) Acquire reduces regret under statistical insufficiency; same-evidence Redesign does not. (C) Mechanism-prescribed repair accuracy under declared encoder contracts. Whiskers are pointwise 95% paired whole-seed bootstrap intervals over 30 seeds.

splits and dependence across checkpoint and policy administrations. Invalid outputs, absent claims and unavailable coordinates are reported separately, without imputation or conversion to failure scores. Reported results follow the prespecified estimates, thresholds, eligibility rules and uncertainty procedures; outputs are not altered after analysis. The reproducibility specification describes these assignments and analysis procedures.

K.2 Scope Boundaries

Measurement scope. FASGL is a candidate five-role instantiation, with no independence, orthogonality or additivity assumption. Evidence concerns controlled domains and fixed checkpoints under study-specific allocations. Calibration is conditional on the specified response model and proper priors; current test–retest reliability, model-family invariance and formal DIF inference remain unestablished. Historical outcome definitions and raw-score margins retain their specified interpretation scope.

Interpretation scope. Profiles are coverage-conditioned: common-support separation and the full-assignment effect remain unestablished, rather than supporting a universal ranking or recovery of missing epistemic states. Nulls and availability remain part of measurement. Coupled role responses, provenance versus warrant, and surface equivalence without a justified margin retain the boundaries documented in Appendices G–I. Geometry concerns identification and conditional precision under the specified local model.

Intervention scope. Diagnostic signal, mechanism recovery, confidence calibration and task utility are distinct endpoints. Validated operators recover specified input-state properties under the tested family, checkpoint and action scope. Coupled operators, undefined diagnosis/task means, sparse repair evidence and zero calibrated action coverage limit claims of generalization to reliable action and task benefit. Appendix J provides the full endpoint and uncertainty analyses; no general repair superiority or causal improvement across protocols is established.

K.3 Repair-Selection Protocols and Accounting

Shared diagnostic contract. The controlled and model-grounded studies use orthonormal-row encoders R, a supplied unit decision normal n, observations $X_j = x + \epsilon_j$ with $\epsilon_j \sim \mathcal{N}(0, I_5)$, and action $a \in \{-1, +1\}$. The loss $L(x, a) = |n^\top x| - an^\top x$ gives zero loss to the sign oracle; it is an intervention-decision endpoint, not an LLM task score. The gate first tests $n \in \mathrm{Row}(R)$. If this fails it selects REDESIGN; otherwise it tests directional variance against the target 0.0625 and selects ACQUIRE when precision is insufficient. Its inputs are encoder, normal, declared sensor noise, observation count, variance target and tolerance 10^{-12}. The gate receives no branch labels, corruption indicators, hidden states, oracle repair labels or realized losses. Encoder structure is supplied as part of the controlled diagnostic contract; semantic validity is not learned.

Four noisy initial vectors give variance 0.25 for an estimable unit contrast; acquiring sixteen more reduces it to 0.05. Redesign applies identity encoding to the same four retained noisy vectors, not the noiseless state. The Gaussian sample-mean contrast is the MLE without a prior. A non-estimable contrast has no fitted value or finite variance and uses the declared $+1$ policy fallback; exact action ties also use $+1$. The statistical branch gains no decision information from same-evidence redesign. Resources are observation counts and restored channels, without an economic exchange rate. Semantic redesign may still need acquisition afterward. No repair is asserted to improve every noisy history.

Controlled population. For each orientation, $x = u \pm mn$ with orthogonal Gaussian nuisance u and $m \sim \mathrm{Uniform}[0.15, 0.60]$. Equal-rank four-row encoders respectively erase n or a nuisance direction, ruling out rank alone as a regime identifier. Thirty seeds (2026092301–2026092330) generate 120 orientation blocks, 3,840 paired families, 7,680 sign-specific states and 15,360 two-branch histories. Signs, branches and policies share potential evidence; their 61,440 records are not independent observations. All histories have defined outcomes. Zero numerical failures occurred; 11,533 non-estimable fallback records remain explicit. The paired failure-by-repair interaction is 0.335 [0.329, 0.340], defined as $(R_{\mathrm{Acquire}} - R_{\mathrm{Redesign}})_{\mathrm{semantic}} - (R_{\mathrm{Acquire}} - R_{\mathrm{Redesign}})_{\mathrm{statistical}}$.

K.4 Blind Repair Diagnosis Protocol

Information restriction. The blind study uses the same five-dimensional population law, four-initial-observation budget, 16-observation Acquire action, same-evidence Redesign action, and intervention loss as the controlled repair-selection study. The diagnostic gate does *not* receive the encoder matrix R, semantic/statistical regime label, corruption indicator, hidden state x, oracle repair label, or realized loss. For each unknown interface it receives only an independent labeled calibration set consisting of noisy exposed representations $z = R\bar{X}$ and the corresponding downstream binary action label. Thus the blind diagnostic tests whether the action distinction remains empirically separable in the exposed representation rather than reading regime identity from the declared encoder geometry.

Calibration and frozen threshold. Each unknown interface receives 24 paired calibration probes generated independently of its evaluation states. Opposite-action pairs are kept pair-disjoint between diagnostic training and holdout partitions. A logistic calibration classifier is fitted on one half of the probe pairs and evaluated on the held-out pairs using balanced accuracy. A separate meta-calibration set of 16 semantic and 16 statistical interfaces per seed selects one scalar separability threshold; the threshold is then frozen before blind evaluation. Scores below threshold route to REDESIGN; scores at or above threshold route to ACQUIRE. In both blind studies, separate action decoders use the same calibration states and labels: Acquire fits on observation means from the initial four plus sixteen added observations under the original representation; Redesign fits on only the original four under the redesigned representation. Diagnostics use only initial evidence; evaluation states and labels never enter decoder fitting. No evaluation state or realized intervention loss enters threshold selection.

Evaluation population and accounting. Thirty seeds generate 40 held-out semantic and 40 held-out statistical interfaces per seed, for 2,400 blind interfaces in total. Each interface is evaluated on 80 opposite-sign state pairs (160 states) with shared potential evidence across repair policies. Diagnosis accuracy is averaged over interfaces; intervention regret is averaged over all evaluation states and then summarized at the seed level. Seed bootstrap intervals use 10,000 resamples of the 30 seed summaries.

Blind-diagnosis results. Overall diagnosis accuracy is 92.62%, with 96.42% semantic recall and 88.83% statistical recall. Mean equal-regime-mixture intervention regret is 0.1175 for Blind Gate, 0.2085 for Always Acquire, 0.1710 for Always Redesign, and 0.1889 for Random routing. Blind Gate improves over Always Acquire by 0.0909 (95% seed-bootstrap CI [0.0886, 0.0931]) and over Always Redesign by 0.0535 [0.0519, 0.0551]. By regime, semantic regret is 0.1772/0.3740/0.1703 for Blind Gate/Always Acquire/Always Redesign; statistical regret is 0.0578/0.0430/0.1717. The experiment establishes blind repair diagnosis only within the declared controlled family; it does not establish diagnosis of naturally occurring AI failures.

K.5 Learned-Representation Repairability Protocol

Population and learned representation. Each interface samples a five-dimensional Gaussian state geometry with a random orthonormal covariance basis and log-uniform eigenvalues in $[0.15, 4.0]$, a random unit decision normal n, observation noise $\sigma \in [0.45, 2.4]$ on a log scale, and a task-relevance weight $\alpha \in [0.01, 5]$. Eight hundred training states are drawn independently. The rank-three learned subspace is the leading eigenspace of $C + \alpha nn^\top$, where C is the empirical state covariance. This is the population analogue of a linear encoder trained jointly for reconstruction structure and decision relevance; no semantic/statistical failure label is generated or used. Redesign relearns the same rank-three representation with relevance weight $\max(\alpha, 4)$ and no additional observations. Acquire freezes the original representation and increases the observation count from four to 20.

Blind calibration and evaluation. Each interface receives 24 independent opposite-action calibration pairs with margins sampled uniformly from $[0.12, 0.70]$. A logistic calibration classifier is trained and evaluated on pair-disjoint halves; held-out balanced accuracy is the sole gate statistic. For each of 30 seeds, 32 separate meta-interfaces choose one scalar threshold minimizing meta-interface routing regret; the threshold is frozen before 60 held-out interfaces are generated. Scores below threshold route to Redesign and scores above threshold to Acquire. Scores equal to the threshold are routed to Acquire. Held-out interfaces contain 80 opposite-action evaluation pairs. The gate never observes W, n, α, σ, decision-subspace overlap, oracle repair, or realized evaluation loss. The full evaluation therefore contains 1,800 held-out interfaces. Decoder evidence budgets follow Appendix K.4. Whole-seed 95% percentile intervals use 10,000 bootstrap resamples of the 30 seed summaries.

Results and mechanism diagnostic. Mean regret is 0.1442 [0.1410,0.1474] for Blind Gate, 0.1480 [0.1436,0.1525] for Always Acquire, 0.1648 [0.1612,0.1684] for Always Redesign, 0.1551 [0.1513,0.1590] for Random, and 0.1091 [0.1063,0.1120] for the oracle. Gate repair-label accuracy is 0.6544 [0.6328,0.6756]. The paired gate improvement is 0.0038 [0.0013,0.0066] over Always Acquire, 0.0206 [0.0161,0.0252] over Always Redesign, and 0.0109 [0.0073,0.0146] over Random. Decision-subspace overlap $\rho_n = \|Wn\|_2^2$ is computed only post hoc. Across overlap quartiles, mean ρ_n values are 0.422, 0.794, 0.946, and 0.996; the corresponding fractions for which Acquire is oracle-optimal are 0.084, 0.531, 0.942, and 1.000. This diagnostic supports a continuous repairability interpretation but is not an input to routing.

Scope. Repair type is not assigned by construction after representation learning; instead, the better repair is determined by held-out regret. The encoder family remains linear and the data-generating process controlled. We therefore describe the result as *emergent repairability in learned representations*, not as naturally occurring deployment failure; nonlinear and real-deployment validation remain open.

Model-grounded population and corruption. The archive contributes 24 initial responses from Gemma and Qwen over 12 families, with no fresh generation or rescoring. The semantic encoder retains F/G/L and suppresses A/S; the statistical encoder retains A/S/G/L for $n = (0, -1, 1, 0, 0)/\sqrt{2}$. Nine overlapping archived opposite-action pairs collide under the imposed mask. Twelve private counterfactual checks validate the construction but are not additional model responses or metric observations. A separate rank-matched preserving control verifies that the gate does not route from rank alone.

Thirty fresh seeds (2026092501–2026092530) and 32 sensor repetitions per request yield 46,080 planned branch histories. Of 24 requests, 22 are schema-valid and 14 have complete FASGL states from eight families. Eight have unavailable Attribution and two invalid responses lack all coordinates; none is imputed. Defined regret uses 26,880 branch histories; the 19,200 incomplete-state slots retain unavailable outcomes. The two complete zero-margin states contribute zero regret but undefined repair labels, leaving 23,040 defined repair-label histories from twelve requests in seven families. These are sensor repeats on a small fixed archive, not new independent model samples. The 184,320 policy records comprise 87,271 estimated, 20,249 explicit non-estimable fallback and 76,800 unavailable-state outcomes; numerical failures are zero.

Uncertainty and interpretation. Both studies use 10,000 paired whole-seed percentile bootstrap draws, with bootstrap seeds 2026092399 (controlled) and 2026092599 (model-grounded). Each

draw resamples thirty seed bundles with replacement, preserving all within-seed dependence and using shared indices across metrics and contrasts. Model-grounded estimates divide summed defined outcomes by summed defined counts; zero-denominator draws would remain undefined rather than be discarded. None occurred. Intervals are pointwise Monte Carlo intervals, conditional on the fixed archive and availability in the model-grounded study, not population-generalization or simultaneous guarantees.

The model-grounded paired interaction is 0.316 [0.314, 0.318]. Its all-planned conservative regret missingness bounds are [0.042, 1.220] for Gate, [0.104, 1.283] for Always Acquire, [0.072, 1.250] for Always Redesign and [0.089, 1.267] for Random, using the inherited cap $2\|n\|_1$. These overlapping identification bounds are not confidence intervals; superiority is restricted to defined cases. All-planned repair accuracy is undefined where a unique repair reference is absent. Perfect gate routing is largely algebraic under known contracts. Fixed-policy overall accuracy is $1/2$ in the balanced design; Random also has exactly $1/2$ because its choice is shared across the two opposite-reference branches. These degenerate intervals do not establish deployment certainty. Individual gate repairs increase realized regret in 905 semantic and 232 statistical controlled histories, and in 1,246 semantic and 207 statistical model-grounded histories; these outcomes remain included.

Provenance and endpoint separation. Reproduction of the controlled and model-grounded studies uses the recorded raw states, source and checkpoint identities, seed assignments, encoder and corruption specifications, sensor noise, diagnostic inputs, potential repair outcomes and resource counts. Repeated execution of all thirty seed runs in each study reproduces the results, and consistency checks confirm internal agreement. Model-grounded uncertainty is conditional on the archived claim-level substrate and imposed Gaussian observation model; checkpoint test–retest reliability is not estimated. Task recovery and action calibration are evaluated separately in Appendix J, preserving the distinction between formal repair routing, mechanism recovery, confidence calibration and task utility.

K.6 Additional Real-Model Qualification Detail

The conditional instrument is qualified as a measurement model on 360 matched-execution rows, 132 source-specific items, and 33 families. Support–Attribution dependence is $\beta = 4.742$ [3.522,5.961], and known-truth geometry recovers the predicted weakest-direction scaling (slope 0.981, 95% interval [0.970, 0.992]). System separation is strongly coverage-dependent: Gemma Base–Qwen Retrieval separates on system-specific complete observations (distance 6.011 versus local radius 3.327), whereas common-item and common-family distances of 2.697 and 3.032 leave common-support separation unresolved; equal-family weighting gives a separate coverage-adjusted distance of 6.214. This is an important negative result: apparent system separation can be created or erased by the observation support on which the profile is defined.

Controlled construct interventions produce 396 attempts, 393 generations, 391 schema-valid responses, and 104 complete five-coordinate pairs. Observed target-coordinate changes across F/A/S/G/L are 0.074/0.433/0.690/0.364/0.313. The asymmetry is informative: Support responds strongly to intended evidence manipulations, while Formation is comparatively stable; Support and Attribution often move together, indicating that the empirical coordinates are not orthogonal causal factors. These experiments are interpreted as manipulation sensitivity and coupling evidence rather than external validation of a unique ontology.

Operational tests expose a second boundary: measurement availability and authority attribution remain difficult in practice. Only 157/396 surface pairs jointly observe A/S/G/L, while 58 are reference-only, 46 surface-only, and 135 observe neither. Authority is especially difficult: direct authority assessment is correct in 3/47 scoreable responses and native authority production in 0/71. These findings delimit the observation support and authority interpretation of the measurement. Appendix G–I gives the complete likelihood, covariance, construct-response, and availability analyses.