LATENT-SPACE COUNTERFACTUAL VERIFICATION FOR ROBUST REASONING AGENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

A reasoning trajectory can produce the correct answer while relying on a brittle shortcut. We show that nominal accuracy substantially overestimates strict robust accuracy under semantics-preserving edits, by 21.9% on average and up to 65.4% across 25 model–task configurations and can even reverse model rankings on MATH-500. To detect such failures, we introduce GAUGE, a counterfactual verifier that tests whether model representations respond consistently to controlled input transformations with known answer-level invariances and equivariances. GAUGE learns a contrastive robustness metric and uses conformal calibration to provide statistically valid score thresholds, enabling reference-free scoring at test time. Across 23 configurations, GAUGE predicts which nominally correct outputs fail under held-out edits—a property, we show, predominantly of the prompt rather than the trajectory—with a conditional AUROC of 0.770, outperforming 14 published baselines. Its probe-free distillation reads the candidate's own hidden state, adding +0.021 AUROC over confidence and length and matching pool self-consistency at zero extra generation. Using GAUGE for self-training selection improves strict robust accuracy while maintaining nominal accuracy, recovering 60–67% of available oracle gains.

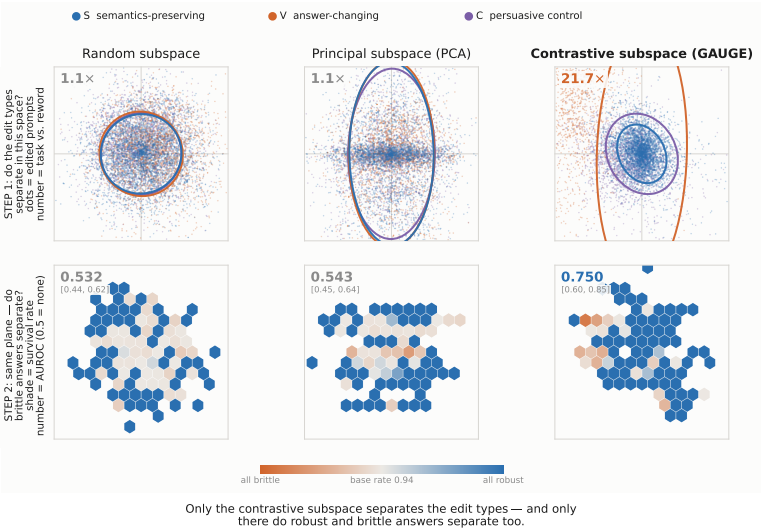


Figure 1: Brittle reasoning becomes visible under the right latent metric. One model's reasoning trajectories (Qwen3-4B on GSM-IC), with every prompt edited in three *typed* ways—typed meaning we know in advance what each edit should do to the answer: *semantics-preserving* ($\mathcal{S}$) rewordings leave it unchanged, *answer-equivariant* ($\mathcal{V}$) edits change the task so it must change in a known way, and *diagnostic controls* ($\mathcal{C}$) insert a misleading hint that a sound derivation must ignore. *Top:* where each edit moves the model's internal state relative to the unedited rollout (origin; 1σ contours), in the leading 2D plane of a random basis, of PCA, and of GAUGE's learned metric; a useful plane keeps $\mathcal{S}$ small while making $\mathcal{V}$ and $\mathcal{C}$ large. *Bottom:* the same candidates, shaded by how often they stay correct under held-out rewordings. Only GAUGE's plane separates the edit families, and only there do brittle and robust candidates occupy different regions (Table 12).

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.

1 INTRODUCTION

A language model can produce a correct answer for reasons that collapse under a minor rewording: symbolic re-instantiation degrades open models (Mirzadeh et al., 2025), irrelevant context derails multi-step arithmetic (Shi et al., 2023), and stated chains of thought often diverge from the computation that produced them (Turpin et al., 2023; Chen et al., 2025). Across every policy family and reasoning suite we evaluate, nominal correctness overstates *strict robust accuracy*—survival under semantics-preserving edits—and on MATH-500 the gap inverts a ranking outright, making the nominally stronger policy the more brittle one (§4.3).

This undermines any pipeline that *selects* rollouts: best-of-J reranking, rejection fine-tuning (RFT), preference optimization. The verifiers such pipelines rely on score only the realized trajectory, never what the model would do under a controlled change of the prompt, so shortcuts and genuine derivations look alike and self-training reinforces the shortcut (Shao et al., 2025; Helff et al., 2026).

We ask a counterfactual question instead: does a trajectory's *latent* response to controlled perturbations predict its survival under held-out edits? GAUGE[1] answers it with **typed edits**—semantics-preserving, answer-equivariant and diagnostic controls (Figure 1)—each carrying a **total answer map** $\tau_g : \mathcal{A} \to \mathcal{A}$ defined for *any* candidate answer, not only the reference one. Totality is what makes scoring reference-free: a probe is judged against the candidate's own mapped answer $\tau_g(\widehat{a}_0)$, which provably agrees with an oracle on correct rollouts (Proposition 12). GAUGE scores the drift these edits induce under a trace-optimal metric that whitens nuisance variation while amplifying task and shortcut sensitivity, and split-conformal calibration turns that score into a threshold with a distribution-free bound on how often it is exceeded—one construction spanning a supervised score, a coupling that needs no robustness labels, and a distilled head that adds no inference rollouts (§3.4).

Contributions.

- **The robustness gap, and what it is.** Nominal correctness overstates strict robust accuracy and can invert model rankings (§4.3); among correct answers that gap is predominantly a property of the *prompt*, even under a label tied to each trajectory's own reasoning, so we cast verification as estimating a policy's reliability on a prompt (§4.2, Appendix H.3).
- **Typed transformation calculus.** Perturbation families carrying total answer maps make counterfactual scoring reference-free, provably matching an oracle on correct rollouts (§3.2, §3.4).
- **Whitened metric, conformal thresholds, distilled head.** A trace-optimal whitened metric scores typed drift with split-conformal score thresholds, and GAUGE-D distils it into a probe-free head costing one inner product, whose hidden-state input adds to confidence and length (§3.3–§4.4).
- **Evidence.** With frozen hyperparameters and disjoint probe and label mechanisms, GAUGE ranks prompt-level reliability above every published verifier baseline we test, and curating self-training data with it raises strict robust accuracy at matched nominal accuracy, in single-turn and agentic settings (§4.4–§4.6).

2 RELATED WORK

Outcome and process verifiers. Verifiers differ in which segment they evaluate: *outcome* verifiers score final answers (Cobbe et al., 2021), *process* verifiers score intermediate steps from human or synthetic labels (Lightman et al., 2024; Wang et al., 2024; Zheng et al., 2025; Zhang et al., 2025; Kamoi et al., 2026), *progress* verifiers estimate success probabilities (Setlur et al., 2025), and *generative* verifiers reason to a verdict (Zhang et al., 2025; Khalifa et al., 2026; Zheng et al., 2023). All inspect only the realized trace, in which a shortcut and a genuine derivation are indistinguishable.

Representation-based verification. A second line reads internal activations: verbalized confidence and semantic entropy calibrate uncertainty (Kadavath et al., 2022; Kuhn et al., 2023; Farquhar et al., 2024), and linear probes recover truthfulness (Zou et al., 2023; Burns et al., 2023; Azaria and Mitchell, 2023; Orgad et al., 2025). Closest to us, Liang et al. (2026) verify traces from activation deltas along an unperturbed rollout—passive drift, which performs poorly here (Table 6). GAUGE instead reads *induced* drift under typed perturbations, without correctness labels.

[1] After Hermann Weyl's *Eichinvarianz* (1918, 1929): laws characterised by the transformations that leave them invariant.

Contrastive activation engineering. Representation engineering and contrastive activation addition fit probes or steering vectors to activation differences of positive and negative prompts (Zou et al., 2023; Panickssery et al., 2024). GAUGE instead scores displacements under three typed families with total answer maps, not binary pairs (§3.2); whitens them against rewording variance via Equation (6), not raw Euclidean shifts; and distils them into a head needing no counterfactual rollouts (§3.4).

Reasoning fragility and reward gaming. Accuracy masks fragility to irrelevant context (Shi et al., 2023), symbolic re-instantiation (Mirzadeh et al., 2025) and option order (Pezeshkpour and Hruschka, 2024; Zheng et al., 2024), and optimizing checkable rewards invites Goodhart effects (Amodei et al., 2016): RLVR gains persist under corrupted rewards (Shao et al., 2025), and models game verifiers with heuristics exposed only by isomorphic tests (Helff et al., 2026). Prior work audits fragility post hoc; GAUGE predicts it at verification time.

Self-training and agentic reliability. Self-training (STaR (Zelikman et al., 2022), RFT (Yuan et al., 2023), DPO (Rafailov et al., 2023), GRPO (Shao et al., 2024)) filters rollouts by terminal correctness, and in agentic environments (τ-bench (Yao et al., 2025), AppWorld (Trivedi et al., 2024), ALFWorld (Shridhar et al., 2021)) unreliable execution, not unsolvable tasks, dominates failure. GAUGE tests action-signature stability, with split-conformal thresholds (Vovk et al., 2005; Lei et al., 2018; Angelopoulos and Bates, 2023; Quach et al., 2024; Mohri and Hashimoto, 2024).

3 METHOD

3.1 PROBLEM SETTING

Instances, rollouts, and action signatures. Let $\mathcal{A}$ be the *answer space*, the set of outputs the answer parser can return (a number, an option, or an action signature), and $\mathcal{D}$ the *instance distribution* of a task. An instance $\xi = (x, a^\star, \mathfrak{E}, m) \sim \mathcal{D}$ is a prompt, a reference answer $a^\star \in \mathcal{A}$, a tool environment, and an answer-equivalence predicate $m : \mathcal{A} \times \mathcal{A} \to \{0, 1\}$. A policy π_θ interacting with $\mathfrak{E}$ emits a rollout $\zeta = (x, r_1, u_1, o_1, \ldots, r_L, u_L, o_L, r_{L+1}, \widehat{a})$ over L tool-calling turns capped at $L_{\max}$, with r_ℓ generated reasoning, u_ℓ tool actions and o_ℓ observations; a deterministic parser reads the final answer $\widehat{a}(\zeta) \in \mathcal{A}$ (interactively, its *action signature*: the ordered environment-modifying tool calls with their arguments, m being signature equality), giving nominal correctness $O(\xi, \zeta) = m(\widehat{a}(\zeta), a^\star)$, and a frozen π_{θ_0} draws J rollouts per instance as the candidate pool $\mathcal{Z}_i = \{\zeta_{ij}\}_{j=1}^{J} \sim \pi_{\theta_0}(\cdot \mid \xi_i)$. At $L = 0$—the single-turn suites, hence most of our results—the rollout is $\zeta = (x, r_1, \widehat{a})$ with $\mathfrak{E}$ inert and signature $\widehat{a}$; nothing downstream changes, since everything below reads only generated hidden states and the parsed answer.

The verification problem and operational pipeline.

Definition 1 (Robustness, and the verification problem)**.** Let $\mathcal{S}_i^{\text{held}}$ be *semantics-preserving edits* (paraphrase, number words, a distractor sentence) held out from the verifier. A candidate ζ_{ij} is *robust* if it is correct on ξ_i and the policy stays correct under every held-out edit:

$$\text{Robust}_{\mathcal{S}}(\xi_i, \zeta_{ij}) = O(\xi_i, \zeta_{ij}) \prod_{s \in \mathcal{S}_i^{\text{held}}} O\big(s(\xi_i), \zeta_{ij,s}\big) \in \{0, 1\}, \tag{1}$$

with $\zeta_{ij,s} \sim \pi_{\theta_0}(\cdot \mid s(\xi_i))$ drawn under the edited prompt. We seek a score $R(\xi_i, \zeta_{ij}) \in \mathbb{R}$ ranking candidates by $\text{Robust}_{\mathcal{S}}$ from π_{θ_0} alone—using at test time neither $a^\star$, nor $\mathcal{S}_i^{\text{held}}$, nor robustness labels—while admitting a distribution-free coverage guarantee.

Test-time instances have no $a^\star$ and cannot be tried against every unseen edit, so GAUGE proceeds in four stages: counterfactual probe trajectories from known answer maps (§3.2); a contrastive subspace separating robustness signal from phrasing noise (§3.3); a spectrum of probe-based and probe-free scores (§3.4); and split-conformal acceptance thresholds for fine-tuning (§3.5).

3.2 PROBE TRAJECTORY GENERATION

Transformation taxonomy with total answer maps.

Definition 2 (Valid typed transformation)**.** Let $\mathcal{A}/\!\sim_m$ be the parseable answer space modulo the equivalence m, with classes $[a]_m$. An edit $g \in \mathcal{T} = \mathcal{S} \dot\cup \mathcal{V} \dot\cup \mathcal{C}$ (semantics-preserving, answer-equivariant, diagnostic-control) maps $\xi = (x, a^\star, \mathfrak{E}, m)$ to $g(\xi) = (x_g, a_g^\star, \mathfrak{E}_g, m_g)$; it is *valid* if a

known, total map $\tau_g : \mathcal{A}/\sim_m \to \mathcal{A}/\sim_{m_g}$ preserves candidate correctness,

$$m_g\big(\tau_g([a]_m), a_g^\star\big) = m(a, a^\star) \qquad \text{for every } a \in \mathcal{A}, \tag{2}$$

so $a_g^\star \in \tau_g([a^\star]_m)$. An unparseable answer maps to $\perp$, which matches no probe (weight $w_g = 1$ in Equation (8)).

Totality—τ_g is defined for *every* candidate answer, not only $a^\star$—is what removes the reference answer from test-time verification. **Semantics-preserving** edits ($\mathcal{S}$, $\tau_s = \mathrm{id}$) reword or rename; **answer-equivariant** edits ($\mathcal{V}$, $\tau_v \neq \mathrm{id}$) change the task, e.g. scale the quantity asked for or redirect an agent subgoal, so the answer must become $\tau_v(a^\star)$; **diagnostic controls** ($\mathcal{C}$, $\tau_c = \mathrm{id}$) add non-causal cues, such as a confidently stated wrong suggestion. $\mathcal{S}$ and $\mathcal{C}$ share $\tau = \mathrm{id}$ but test different failures, so each keeps its own normalizer and weight; a candidate swayed by a cue can still return $\tau_c(\widehat{a}_0) = \widehat{a}_0$, leaving *no answer-level signature*, which is why GAUGE reads hidden activations. Every transformation is programmatic or a number-preserving paraphrase (audit: Appendix H.3); the one answer-dependent cue, rebuilt from the candidate's own answer, reproduces 99.8% of evaluated cues (Appendix R).

Probe bundles and statistical design constraints.

Definition 3 (Probe bundle)**.** For candidate ζ_{ij} on ξ_i, the probe bundle $\mathcal{B}_{ij}$ contains ζ_{ij}, one rollout per probe in $\mathcal{S}_i \cup \mathcal{V}_i \cup \mathcal{C}_i$, and an independent matched noise replicate $\zeta_{ij,N} \sim \pi_{\theta_0}(\cdot \mid \xi_i)$, costing $1 + |\mathcal{S}_i| + |\mathcal{V}_i| + |\mathcal{C}_i| + \mathbf{1}\{\text{noise}\}$ rollouts.

Two constraints keep this honest: the held-out edits $\mathcal{S}_i^{\text{held}}$ of Eq. (1) share no transformation with the probe families $\mathcal{T}$ (the two paraphrases are independent samples), so a verifier cannot memorize perturbation artifacts; and because the bundles for one instance share a prompt and are conditionally dependent, every split, bootstrap interval and conformal calibration is clustered at the prompt-group level.

3.3 SUBSPACE LEARNING FOR ROBUSTNESS SIGNAL REPRESENTATIONS

Hidden-trajectory features. At a fixed layer, let $H_\theta(\zeta) = (h_0, \ldots, h_T)$ be the hidden states of rollout ζ at its model-generated token positions ($h_t \in \mathbb{R}^{d_h}$; observations and prompts excluded).

Definition 4 (GAUGE pooled feature)**.** The trajectory representation pools initial, mid-point, terminal, and average hidden states:

$$\phi(H) = \big[\, h_0;\ h_{\lfloor T/2 \rfloor};\ h_T;\ (T+1)^{-1} \textstyle\sum_t h_t \,\big] \in \mathbb{R}^{4d_h}. \tag{3}$$

Three temporal anchors and a running mean capture both local state and the trajectory's overall direction (defined identically at $T = 0$); Appendix A.2 bounds its losses.

Contrastive metric formulation. Raw Euclidean distance is dominated by generic high-variance directions, so we learn a task-sensitive metric on an offline fit split from transformation types alone, *without robustness labels*. Let ϕ_i^0 and $\phi_{i,g}$ be the original and perturbed pooled rollouts, reduced to low-dimensional representations $z(\phi) = P^\top(\phi - \mu) \in \mathbb{R}^K$ by projection onto the top-K principal components $P \in \mathbb{R}^{4d_h \times K}$, where μ is the mean pooled feature on the fit split. For displacement $\delta_{i,g} = z(\phi_{i,g}) - z(\phi_i^0)$ across family $G \in \{\mathcal{S}, \mathcal{V}, \mathcal{C}\}$, with $N_G = \sum_i |G_i|$ the number of displacements available in that family, we compute the uncentered scatter C_G and its regularized shrinkage covariance $\widehat{\Sigma}_G$:

$$C_G = \frac{1}{N_G} \sum_i \sum_{g \in G_i} \delta_{i,g} \delta_{i,g}^\top, \tag{4}$$

$$\widehat{\Sigma}_G = (1 - \rho) C_G + \rho \frac{\mathrm{tr}(C_G)}{K} I_K + \eta I_K. \tag{5}$$

Conditioning the covariances. Equation (5) is Ledoit–Wolf shrinkage (Ledoit and Wolf, 2004): $\rho \in [0, 1]$ mixes the scatter with the equal-trace target $\mathrm{tr}(C_G) I_K / K$, keeping $\widehat{\Sigma}_G$ well conditioned when N_G is small relative to K, and the ridge $\eta > 0$ makes $B = \widehat{\Sigma}_\mathcal{S}$ invertible in Equation (6); both are fixed by the staged search ($\rho = 0.1$, $\eta = 10^{-4}$; Appendix T), and Theorem 9 bounds how far estimation noise can move the fitted subspace.

The contrastive basis. To isolate essential reasoning shifts ($\mathcal{V}, \mathcal{C}$) from phrasing variance ($\mathcal{S}$), we solve $Au_j = \gamma_j Bu_j$ with

$$A = \widehat{\Sigma}_{\mathcal{V}} + \lambda_C \widehat{\Sigma}_{\mathcal{C}}, \qquad B = \widehat{\Sigma}_{\mathcal{S}} \qquad (\lambda_C \geq 0). \tag{6}$$

The numerator holds the drift a verifier must detect—$\mathcal{V}$ edits, which a sound derivation must follow, and $\mathcal{C}$ cues, which expose shortcuts—and the denominator the rewording drift it must ignore, so $\gamma_j = u_j^\top A u_j / u_j^\top B u_j$ is a signal-to-nuisance ratio; λ_C trades the two signals, from answer-changing edits alone ($\lambda_C = 0$) towards cue sensitivity (frozen at 0.5). The basis $U = [u_1, \ldots, u_r] \in \mathbb{R}^{K \times r}$ ($U^\top B U = I_r$) collects the eigenvectors of the r largest $\gamma_1 \geq \cdots \geq \gamma_r$. This choice is not heuristic: those eigenvectors are exactly the maximizers of $\mathrm{tr}(U^\top A U)$ subject to $U^\top B U = I_r$, whose optimum is $\sum_{j \leq r} \gamma_j$ (Proposition 7, Appendix A).

Whitening geometry and family roles. Because $U^\top B U = I_r$, projecting onto U *whitens* against rewording: the cloud of $\mathcal{S}$ displacements becomes a unit sphere, so a drift is measured in units of ordinary phrasing variation rather than raw activation scale. At full rank ($r = K$) this is purely a change of metric; within it the $\mathcal{V}$ and $\mathcal{C}$ drifts are largest along the leading u_j (Figure 1), and $\gamma_j < 1$ marks directions where rewording dominates. $\mathcal{S}$ normalizes the metric and $\mathcal{C}$ sharpens it; at test time drift under both is penalized.

3.4 The Gauge scoring spectrum

Task coordinates, RMS scales, and reference-free coupling. For pooled feature ϕ, define projected coordinates $q(\phi) = U^\top P^\top (\phi - \mu) \in \mathbb{R}^r$, writing q_0, q_g, and q_N for original, probe, and noise replicate rollouts. Offline root-mean-square scales are computed on the fit split:

$$\sigma_G^2 = \frac{1}{N_G} \sum_i \sum_{g \in G_i} \left\| q_{i,g} - q_i^0 \right\|_2^2, \qquad G \in \{\mathcal{S}, \mathcal{V}, \mathcal{C}\}. \tag{7}$$

Reference-free answer coupling. To score candidates without ground-truth answers, we introduce the reference-free answer coupling w_g^{free} alongside the oracle coupling w_g^{oracle} (which assumes the ground-truth $a_g^\star$ is available):

$$w_g^{\text{free}} = \begin{cases} \lambda_{\text{corr}}, & m_g(\widehat{a}_g, \tau_g(\widehat{a}_0)) = 1 \\ 1, & \text{otherwise} \end{cases} \qquad w_g^{\text{oracle}} = \begin{cases} \lambda_{\text{corr}}, & m_g(\widehat{a}_g, a_g^\star) = 1 \\ 1, & \text{otherwise,} \end{cases} \tag{8}$$

where $\lambda_{\text{corr}} \in [0, 1]$ is the weight given to a probe that agrees with the map-pushed candidate answer.

For valid edits the two couplings coincide on every nominally correct candidate, so dropping the reference answer costs nothing where verification matters (Proposition 12, Appendix A; proof in Appendix D). The identity needs correctness to equal answer equivalence; τ-bench rewards accept several valid action signatures, so both couplings are evaluated there (§4.4). $\lambda_{\text{corr}} = 1$ gives a purely geometric verifier ($w_g^{\text{free}} = 1$); the frozen $\lambda_{\text{corr}} = 0$ counts only disagreeing probes.

Typed residuals and the Gauge score. Normalized latent drifts penalize excess displacement under invariant edits ($\mathcal{S}, \mathcal{C}$), dispersion under task modifications ($\mathcal{V}$), and stochastic execution instability (N):

$$D_{\mathcal{S}} = \frac{1}{|\mathcal{S}_i|} \sum_{s \in \mathcal{S}_i} w_s \frac{\|q_s - q_0\|_2^2}{\sigma_{\mathcal{S}}^2 + \varepsilon}, \qquad D_{\mathcal{C}} = \frac{1}{|\mathcal{C}_i|} \sum_{c \in \mathcal{C}_i} w_c \frac{\|q_c - q_0\|_2^2}{\sigma_{\mathcal{C}}^2 + \varepsilon}, \tag{9}$$

$$D_{\mathcal{V}} = \frac{1}{|\mathcal{V}_i|} \sum_{v \in \mathcal{V}_i} w_v \frac{\left\|d_v - \bar{d}_{\mathcal{V}}\right\|_2^2}{\sigma_{\mathcal{V}}^2 + \varepsilon}, \qquad D_N = \frac{\|q_N - q_0\|_2^2}{\sigma_{\mathcal{S}}^2 + \varepsilon}, \tag{10}$$

where $\varepsilon > 0$ guards against division by a vanishing scale, $d_v = q_v - q_0$ is the equivariant displacement of probe v, and $\bar{d}_{\mathcal{V}} = |\mathcal{V}_i|^{-1} \sum_{v \in \mathcal{V}_i} d_v$ is its family mean.

Subtracting $\bar{d}_{\mathcal{V}}$ makes $D_{\mathcal{V}}$ penalize only *within-family scatter*, not the shared shift that a correct answer-equivariant response is supposed to produce; we justify this centering, and the singleton fallback of Equation (17) for $|\mathcal{V}_i| = 1$, in Appendix A.4.

Definition 5 (GAUGE residual)**.** The scalar counterfactual instability score is defined as:

$$\boxed{R_{\mathrm{GAUGE}}(\xi_i, \zeta_{ij}) = D_{\mathcal{S}} + D_{\mathcal{C}} + \lambda_V D_{\mathcal{V}} - \lambda_N D_N} \qquad \lambda_V \geq 0,\ \lambda_N \in \mathbb{R}. \tag{11}$$

Lower is more robust; GAUGE (label-free) in Section 4 is R_{GAUGE} at the frozen coefficients, with no weight fitted to robustness labels. The sign of λ_N is free: $\lambda_N > 0$ subtracts resampling drift as nuisance (Appendix A), $\lambda_N < 0$ penalizes it. The selection split picks -0.5, so the term *adds* $0.5\, D_N$: movement under mere resampling marks a reasoning path that has not settled (§4.5).

Fitted coefficients. Two variants instead fit a regularized logistic regression $P(Y = 1 \mid X) = \sigma(\theta^\top X + \theta_0)$ on the disjoint selection split, with target $\mathrm{Robust}_{\mathcal{S}}$ (Eq. (1)) and features X:

- **GAUGE-LR** ($X \in \mathbb{R}^7$): the four geometric drift residuals D_G (evaluated at $\lambda_{\mathrm{corr}} = 1$) with the three per-family probe answer-agreement rates $\bar{\alpha}_G = \frac{1}{|G_i|} \sum_{g \in G_i} \mathbf{1}\{m_g(\widehat{a}_g, \tau_g(\widehat{a}_0)) = 1\}$.
- **GAUGE-LR+** ($X \in \mathbb{R}^{10}$): appends the candidate's mean token log-probability, min-token log-probability, and response length (Definition 13).

Probe-free distillation. To remove the cost of probe rollouts we distill the probe-based score (the *teacher* $\mathcal{T}$) into a probe-free ridge regressor (GAUGE-D; Definition 14) that predicts $\mathcal{T}$ from the candidate's own PCA coordinates $z_{ij} \in \mathbb{R}^K$ (for GAUGE-D+, with its confidence and length), one inner product over activations already computed during generation. We thus evaluate five variants: R_{GAUGE}; the fitted GAUGE-LR and GAUGE-LR+; and the probe-free GAUGE-D (teacher R_{GAUGE}) and GAUGE-D+ (teacher GAUGE-LR+), which need **zero counterfactual rollouts** (Figure 4).

3.5 SPLIT-CONFORMAL CANDIDATE SELECTION

A raw instability score has no units, so no cutoff on it corresponds to a stated error rate. Split-conformal calibration (Vovk et al., 2005; Lei et al., 2018; Angelopoulos and Bates, 2023) supplies one: with the pipeline frozen and nonconformity $\nu = R_{\mathrm{GAUGE}}$, the quantile $\widehat{q}_{1-\alpha}$ on a disjoint calibration split gives an unobserved candidate $\mathbb{P}\{\nu_{\mathrm{new}} \leq \widehat{q}_{1-\alpha}\} \geq 1 - \alpha$ (Theorem 16, Appendix A.8), which Algorithm 3 filters on. The guarantee is about the *score*—how often a new candidate exceeds the threshold—not the share of accepted candidates that are brittle. It also rests on exchangeability; candidates of one prompt are dependent, so we calibrate on *one* per calibration prompt group, drawn at random independently of the scores, and the bound holds for every test candidate (coverage 0.913 at nominal 0.9; Table 16). The self-training runs of §4.6 used the pooled threshold, covered only empirically (0.905), whose decisions differ on 2.0% of candidates. Nominal correctness is enforced jointly.

4 EXPERIMENTS AND RESULTS

We ask how large the robustness gap is (§4.3), whether GAUGE ranks robustness better than published verifiers (§4.4), which components carry the signal (§4.5), and whether it selects better self-training data (§4.6).

4.1 EXPERIMENTAL SETUP

Policies, tasks and splits. Five open-weight policies—Qwen3-4B (primary), Qwen3-8B, Llama-3.1-8B, Olmo-3-7B and Qwen3.8-27B (Yang et al., 2025; Grattafiori et al., 2024; Team Olmo, 2025)—are scored on six reasoning suites (GSM8K, GSM-Symbolic-P2, MATH-500, SVAMP, GSM-IC, MMLU-Pro (Cobbe et al., 2021; Mirzadeh et al., 2025; Hendrycks et al., 2021; Patel et al., 2021; Shi et al., 2023; Wang et al., 2024)) and τ-bench (Yao et al., 2025) (25 configurations; ALFWorld (Shridhar et al., 2021) in §4.4), split by prompt group into fit, calibration, selection and evaluation. Each instance draws J=8 candidates (J=4 agentically). Hyperparameters are frozen on the primary policy's selection split (Appendix T); the evaluation split informs no choice. Intervals are 95% prompt-clustered bootstraps; verifier comparisons are paired and Holm-corrected (Appendix P).

4.2 METRICS

Write O_i for terminal correctness and $O_{i,s}$ for correctness of the same candidate's rollout on $s(\xi_i)$, s a held-out label edit. Over n candidates,

$$\mathrm{Acc} = \frac{1}{n}\sum_{i=1}^{n} O_i, \qquad \mathrm{RAcc}_{\mathcal{S}} = \frac{1}{n}\sum_{i=1}^{n} O_i \prod_{s \in \mathcal{S}_i^{\mathrm{held}}} O_{i,s}, \qquad \mathrm{Hack}_{\mathcal{S}} = 1 - \frac{\mathrm{RAcc}_{\mathcal{S}}}{\mathrm{Acc}}, \tag{12}$$

$\mathrm{Hack}_{\mathcal{S}}$ (lower-better, $\mathrm{Acc} > 0$) conditions on original correctness. As $\mathrm{RAcc}_{\mathcal{S}}$ conjoins independent draws, unedited repeats reproduce 83% of the 18.5% gap, leaving an edit-specific +3.2 pp [+2.2, +4.1] (Appendix H.3). Accuracy would dominate a ranking by $\mathrm{Robust}_{\mathcal{S}}$, so our **primary metric is AUROC against** $\mathrm{Robust}_{\mathcal{S}}$ **among terminally correct candidates**, with the unconditional metric as a control (Appendix F; Table 8). Given correctness the label's expectation belongs to the instance, so the pooled metric is **prompt-level**: permuting labels within a prompt moves it by ≤ 0.001. GAUGE thus identifies *prompts whose correct-looking derivations are unreliable*; *within* a prompt, answer-level scores tie by construction (0.500) and GAUGE is modestly above chance (0.516), while under a label tied to each candidate's own reasoning all correct candidates still agree in 71% of prompts, and continuing the candidate's own prefix recovers the candidate-level remainder (0.681, against 0.489 for typed agreement; Appendix H.3).

4.3 ACCURACY OVERSTATES ROBUSTNESS, IN SINGLE-TURN AND INTERACTIVE SETTINGS

Two points in Figure 2 differ from earlier reports. First, **fragility is model- and benchmark-specific**: on GSM8K the strongest policy loses a few points under rewording, on GSM-Symbolic-P2 (Mirzadeh et al., 2025) far more, so either alone misstates it. Second, **accuracy and robustness are decoupled**: on MATH-500 the ordering inverts—Qwen3-8B is more accurate than Olmo-3-7B (0.854 vs. 0.825) yet less robust (0.725 vs. 0.729)—which is what makes a verifier scoring *trajectories* worth building.

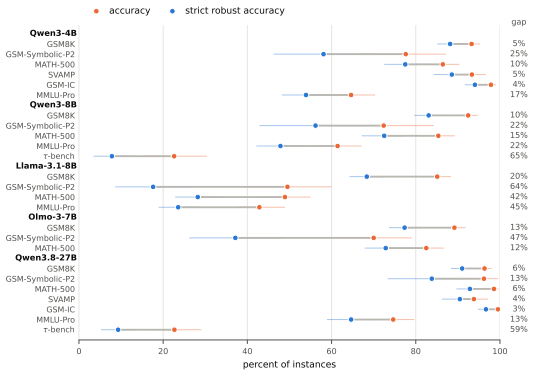


Figure 2: Nominal accuracy overstates strict robust accuracy. Per configuration: nominal accuracy (orange), strict robust accuracy under held-out edits (blue), their gap (grey), 95% prompt-cluster bootstrap intervals. Right column: the hack rate ($\mathrm{Hack}_{\mathcal{S}} = 1 - \mathrm{RAcc}_{\mathcal{S}}/\mathrm{Acc}$), max 65.4% (Table 3).

The same gap, wider, in interactive environments. Agentic trajectories act on a stateful environment that scores them, so on **ALFWorld** (Shridhar et al., 2021) and **τ-bench** (Yao et al., 2025) we set $\widehat{a}(\zeta)$ to the episode's *action signature*. The gap is widest there: under simulator state checks, 49.0% and 48.3% of successful ALFWorld episodes (Qwen3-4B, -8B) fail when the goal is restated with identical semantics, against 11.1% and 17.4% single-turn (Table 22). Qwen3.8-27B cuts ALFWorld fragility to 8.4% but leaves τ-bench at 59.0%: capability alone does not close it.

4.4 GAUGE RANKS FRAGILITY BETTER THAN JUDGE, PROCESS-REWARD AND HIDDEN-STATE VERIFIERS

Ranking fragility. Across all 23 configurations (Figure 3, Table 6) every verifier reading only realized answers trails GAUGE, with outcome oracles at chance conditionally: GAUGE-LR+ attains the highest conditional AUROC (0.770), beats all 14 published baselines (13 under Holm correction), and survives length-controlled stratification (0.734 vs. 0.688 for self-consistency (Wang et al., 2023), margin 0.045). *Typed-probe agreement* rows are our own ablation, not a baseline: they score by the agreement rates $\bar{\alpha}_G$ alone, reading no hidden states, so the margin over them isolates what the learned geometry adds when probes are generated, a margin not resolved at this sample size ($+0.016$, $[-0.010, +0.037]$; Table 6); its resolved contribution is amortisation, below.

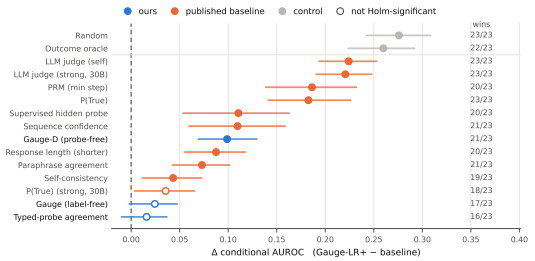


Figure 3: Verifier ranking on conditional AUROC. Paired differences (GAUGE-LR+ − baseline), 23 configurations, 95% bootstrap intervals; filled markers survive Holm correction (13/14); controls below the divider; right: configurations won (Table 6).

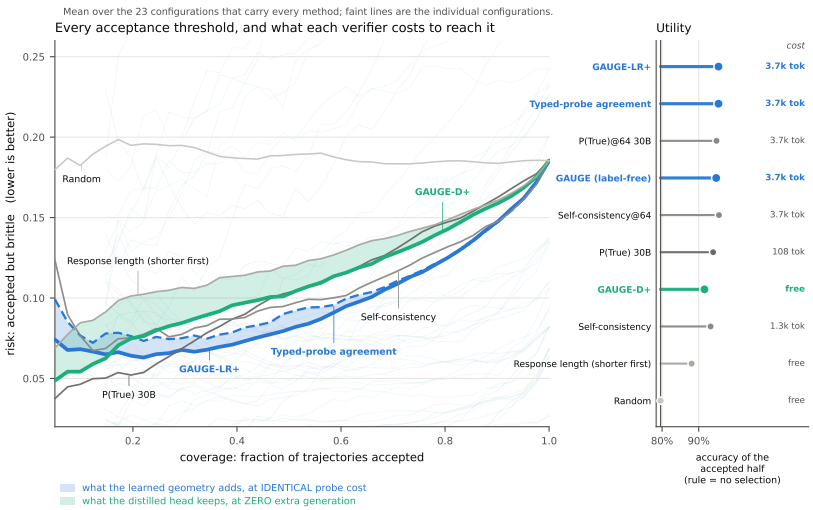


Figure 4: Risk–coverage profiles and generation cost trade-offs. *Left:* Risk (proportion of accepted rollouts failing held-out perturbations) vs. coverage (fraction of rollouts accepted) across 23 configurations; bold curves are cross-configuration means. *Right:* Accepted utility (nominal accuracy of top 50% rollouts) vs. additional inference generation cost. The blue region isolates what the learned geometry adds over typed-probe agreement at identical cost; the aqua region what GAUGE-D+ adds over the best zero-cost verifier. Rows marked @64 give a sampling baseline GAUGE's generation budget (Appendix H.4).

Four regimes, trading supervision against compute. **GAUGE-LR+** (0.770) is most accurate given offline labels and test-time probes; **GAUGE (label-free)** (0.745) evaluates the whitened drift R_{GAUGE} with no robustness label in the score, though the frozen coefficients were selected using labels on a disjoint split (Appendix T), separating from 12 of 14 baselines; **GAUGE-D+** (0.730) distils GAUGE-LR+ into a ridge head on cached activations (**zero extra generation**; hidden state +0.021 [+0.004, +0.035], App. H.3); and **GAUGE-D** (0.671) adds no overhead at all (Table 5).

Interactive environments. On ALFWorld, the most fragile setting (§4.3), the label-free residual reaches 0.808, matching the strongest verifiers on paired per-game bootstraps while fitting nothing, and the supervised variant tracks the training events available (0.453, 0.821 and 0.840 [0.731, 0.913] for Qwen3-4B, -8B and 27B; Table 23). Two environment-specific adaptations—$\mathcal{V}$ edits that flip the completion predicate, and a rank cap ($r = 33$) for numerical stability—leave the score itself unchanged (Appendix I). On τ-bench no direction attains $\gamma_j \geq 1$ (Table 27), so the metric has nothing to measure and we report fragility only—an exclusion decided by the fitted contrast, not by the task name.

Operating-point risk–coverage and compute-matched scaling. AUROC integrates over thresholds, so Figure 4 reads the operating point instead. At half coverage GAUGE-LR+ carries the lowest risk among verifiers at their standard budget (0.081 vs. 0.095, 15% below self-consistency in 19/23 configurations; Table 19), lifts the accepted half's accuracy from 80.0% to 95.4%, and returns a robust trajectory 70.0% of the time at best-of-8 against 69.0% for the best baseline (Table 21). Given GAUGE's generation budget, the sampling baselines' conditional AUROC saturates (self-consistency 0.742, P(True) 0.746; Appendix H.4), and neither orders trajectories *within* a prompt, where answer agreement is tied by construction.

4.5 WHERE THE SIGNAL LIVES: THE METRIC, THEN THE DIRECTIONS

Figure 1's separation holds broadly: a cross-fitted probe on *two* contrastive directions averages 0.616 against 0.566 (principal) and 0.466 (random; Table 12), and over ranks 8–64 the contrastive subspace leads by +0.041 and +0.022 (19/23 and 19/23 configurations; Table 13). At $r{=}K$ every basis spans the same space, so the gain is the *metric*, not the directions: whitening the typed drifts is worth +0.168 over the same probes scored without it (26/26 configurations, ALFWorld included), and the drifts beat the policy's confidence and length by +0.077 in 23/26 (Table 14). Three ablations add detail: the typed families are not interchangeable, the control family mattering most (Figure 6); the matched-noise term enters with the sign opposite to its motivation, so λ_N keeps a free sign (§3.4); and the operator transfers across tasks (Tables 11, 15).

4.6 TRAINING ON GAUGE'S SELECTIONS IMPROVES THE AGENT

Self-training setup and candidate selection quality. A verifier matters only if selecting with it yields a more robust policy. Every arm shares a base policy, training pool, candidate budget J and rejection fine-tuning (RFT; Yuan et al., 2023) interface across three LoRA (Hu et al., 2022) iterations and three seeds, on a heterogeneous pool spanning ALFWorld, τ-bench, DeepMath-103K (He et al., 2025), MBPP (Austin et al., 2021), BFCL (Patil et al., 2025) and ARC (Clark et al., 2018). Selectors range from random to top-ranked rollouts under self-consistency, LLM judges (Zheng et al., 2023), PRMs (Zhang et al., 2025) and GAUGE, bounded by an oracle on held-out robustness labels; arms requiring nominal correctness keep matched budgets (972/1500).

Table 1: Self-training policy performance under counterfactual selection. Rejection fine-tuning (RFT) on a heterogeneous pool under matched candidate budgets; mean$_{\pm \text{s.d.}}$ over 3 seeds. *Held-out suites:* math/symbolic Q&A (accuracy and robust accuracy). *Shortcut suites:* MATH-Perturb simple (MP-s) and hard (MP-h). *Agentic suites:* ALFWorld (ALF) and τ-bench (τ-b.). Oracle selects via held-out labels.

	Held-out suites		Shortcut suites		Agentic	
Selector	Acc. ↑	Robust ↑	MP-s ↑	MP-h ↑	ALF ↑	τ-b. ↑
Qwen3-4B						
No filter	$0.915_{\pm.002}$	$0.844_{\pm.004}$	$0.896_{\pm.003}$	$0.738_{\pm.005}$	$0.130_{\pm.004}$	$0.118_{\pm.005}$
Outcome (gold)	$0.916_{\pm.002}$	$0.865_{\pm.004}$	$0.910_{\pm.004}$	$0.763_{\pm.004}$	$0.167_{\pm.005}$	$0.188_{\pm.005}$
Self-cons.	$0.912_{\pm.002}$	$0.842_{\pm.004}$	$0.896_{\pm.003}$	$0.742_{\pm.004}$	$0.165_{\pm.004}$	$0.185_{\pm.005}$
LLM judge	$0.915_{\pm.002}$	$0.855_{\pm.004}$	$0.902_{\pm.004}$	$0.751_{\pm.004}$	$0.175_{\pm.005}$	$0.145_{\pm.005}$
PRM	$0.914_{\pm.002}$	$0.851_{\pm.004}$	$0.898_{\pm.003}$	$0.748_{\pm.005}$	$0.160_{\pm.005}$	$0.140_{\pm.004}$
GAUGE (no labels)	$0.914_{\pm.002}$	$0.868_{\pm.003}$	$0.912_{\pm.003}$	$0.768_{\pm.004}$	$0.172_{\pm.004}$	$0.192_{\pm.005}$
GAUGE (correct)	$\mathbf{0.917}_{\pm.002}$	$\mathbf{0.875}_{\pm.003}$	$\mathbf{0.918}_{\pm.003}$	$\mathbf{0.774}_{\pm.004}$	$\mathbf{0.178}_{\pm.005}$	$\mathbf{0.199}_{\pm.005}$
Oracle (bound)	$0.920_{\pm.002}$	$0.880_{\pm.003}$	$0.923_{\pm.003}$	$0.781_{\pm.004}$	$0.185_{\pm.004}$	$0.210_{\pm.005}$
Llama-3.1-8B						
No filter	$0.778_{\pm.003}$	$0.717_{\pm.004}$	$0.762_{\pm.004}$	$0.627_{\pm.005}$	$0.085_{\pm.003}$	$0.076_{\pm.004}$
Outcome (gold)	$0.778_{\pm.002}$	$0.735_{\pm.004}$	$0.774_{\pm.004}$	$0.649_{\pm.004}$	$0.108_{\pm.004}$	$0.122_{\pm.005}$
Self-cons.	$0.775_{\pm.003}$	$0.716_{\pm.004}$	$0.762_{\pm.004}$	$0.631_{\pm.005}$	$0.105_{\pm.004}$	$0.118_{\pm.004}$
LLM judge	$0.778_{\pm.003}$	$0.727_{\pm.004}$	$0.767_{\pm.004}$	$0.638_{\pm.004}$	$0.114_{\pm.004}$	$0.094_{\pm.004}$
PRM	$0.777_{\pm.002}$	$0.723_{\pm.004}$	$0.763_{\pm.004}$	$0.636_{\pm.005}$	$0.104_{\pm.004}$	$0.091_{\pm.003}$
GAUGE (no labels)	$0.777_{\pm.002}$	$0.738_{\pm.003}$	$0.778_{\pm.003}$	$0.654_{\pm.004}$	$0.112_{\pm.004}$	$0.125_{\pm.004}$
GAUGE (correct)	$\mathbf{0.779}_{\pm.002}$	$\mathbf{0.743}_{\pm.003}$	$\mathbf{0.782}_{\pm.003}$	$\mathbf{0.658}_{\pm.004}$	$\mathbf{0.116}_{\pm.004}$	$\mathbf{0.131}_{\pm.004}$
Oracle (bound)	$0.782_{\pm.002}$	$0.748_{\pm.003}$	$0.786_{\pm.003}$	$0.663_{\pm.004}$	$0.122_{\pm.003}$	$0.139_{\pm.004}$
Olmo-3-7B						
No filter	$0.824_{\pm.003}$	$0.760_{\pm.004}$	$0.806_{\pm.004}$	$0.664_{\pm.005}$	$0.117_{\pm.004}$	$0.106_{\pm.004}$
Outcome (gold)	$0.824_{\pm.002}$	$0.779_{\pm.004}$	$0.819_{\pm.004}$	$0.687_{\pm.005}$	$0.150_{\pm.004}$	$0.169_{\pm.005}$
Self-cons.	$0.821_{\pm.003}$	$0.758_{\pm.004}$	$0.806_{\pm.004}$	$0.668_{\pm.005}$	$0.148_{\pm.004}$	$0.165_{\pm.005}$
LLM judge	$0.824_{\pm.003}$	$0.770_{\pm.004}$	$0.812_{\pm.004}$	$0.676_{\pm.005}$	$0.158_{\pm.005}$	$0.131_{\pm.004}$
PRM	$0.823_{\pm.002}$	$0.766_{\pm.004}$	$0.808_{\pm.004}$	$0.673_{\pm.005}$	$0.144_{\pm.004}$	$0.126_{\pm.004}$
GAUGE (no labels)	$0.823_{\pm.002}$	$0.782_{\pm.003}$	$0.822_{\pm.003}$	$0.690_{\pm.004}$	$0.154_{\pm.004}$	$0.171_{\pm.005}$
GAUGE (correct)	$\mathbf{0.825}_{\pm.002}$	$\mathbf{0.788}_{\pm.003}$	$\mathbf{0.828}_{\pm.003}$	$\mathbf{0.697}_{\pm.004}$	$\mathbf{0.159}_{\pm.004}$	$\mathbf{0.178}_{\pm.005}$
Oracle (bound)	$0.828_{\pm.002}$	$0.794_{\pm.003}$	$0.833_{\pm.003}$	$0.704_{\pm.004}$	$0.165_{\pm.004}$	$0.185_{\pm.004}$

On this pool GAUGE (correct) curates a training subset that is 0.763 robust against 0.711 for gold-outcome filtering and 0.675 for self-consistency, from a base pool at 0.401 (Table 24). Both GAUGE arms select with the frozen residual R_{GAUGE} under the label-free coupling (0.745 vs. 0.770 for GAUGE-LR+; §4.4), so these are conservative lower bounds.

Policy improvements on LLM reasoning, agentic suites, and oracle headroom. On held-out math and symbolic Q&A suites (Table 1) GAUGE (correct-only) is the top deployable selector for all three policies (Qwen3-4B, Llama-3.1-8B, Olmo-3-7B), raising strict robust accuracy over outcome-RFT by +1.0, +0.8 and +0.9 pp, 60–67% of oracle headroom, while matching nominal accuracy within 0.002. Gains persist on held-out shortcut benchmarks (MATH-Perturb, GSM-Plus; Table 26), beating outcome-RFT by +0.8 to +1.1 pp, up to +1.2 pp where shortcuts are probed. Interactive suites gain too: +1.1 pp task completion over outcome-RFT on ALFWorld (0.178 vs. 0.167) and +1.1 pp on τ-bench (0.199 vs. 0.188) for Qwen3-4B, with consistent gains on Llama and Olmo. Finally, the **label-free** arm matches or exceeds outcome-RFT in robust accuracy on all models (+0.3 pp) and both agentic suites. The gains are consistent but small: Welch tests over three seeds give $p = 0.029$, 0.055 and 0.040, none surviving Holm correction. They also presume an anchored objective: under DPO the label-free arm falls to 0.030 on ALFWorld, and KL-free GRPO collapses the no-filter control to 0.105 single-turn accuracy while GAUGE's arm keeps 0.921 (Appendices J.1, M); we recommend the correct-only arm under RFT or KL-anchored preference optimisation.

5 CONCLUSION

A policy's internal response to typed perturbations predicts whether its correct answers survive held-out rewordings; total answer maps make the scoring reference-free, and split-conformal calibration over one candidate per prompt group bounds how often the score exceeds its threshold. Across 23 configurations GAUGE-LR+ reaches 0.770 conditional AUROC and outperforms all 14 published verifier baselines (13 under Holm correction); distilled into a zero-cost head, its hidden-state input adds +0.021 [+0.004, +0.035] over confidence and length; and as a self-training filter it raises strict robust accuracy by 0.8–1.0 pp at matched nominal accuracy for every policy, 60–67% of oracle headroom, though not significantly after Holm correction over three seeds. Two limits remain: the robustness it measures belongs mostly to the prompt, so ranking derivations within one requires continuing each one's own prefix (Appendix H.3); and it needs white-box access to hidden activations.

Reproducibility Statement

Every number in this paper is produced by the released pipeline from artefacts stored per run: split identifiers, transformation templates with their answer maps, per-rollout records, fitted subspaces with the consistency check of Remark 11, frozen hyperparameters with the full selection log, and the bootstrap configuration. Section 4.1 states the split protocol; Appendix Q lists every transformation and how often it is instantiated; Appendix T gives the frozen hyperparameters and the staged search that produced them; Appendix S is the implementation checklist. Tables and figures are rendered directly from the analysis outputs, so no number in the text is transcribed by hand. Limitations are collected in Appendix M.

Ethics Statement

All datasets are public research benchmarks and all models are open-weight. Transformation text is generated with the same class of open models and is checked automatically for preservation of the quantities that determine the answer. GAUGE is a verification and selection mechanism: it can reduce the rate at which brittle reasoning is reinforced during self-training, but it does not certify correctness, and the conformal statement in Section 3.5 should not be read as a safety guarantee. The diagnostic controls deliberately insert misleading cues into prompts; they exist to measure a model's sensitivity to non-causal features and are confined to the evaluation harness.

AI Use Statement

Large language models enter this work in two roles, disclosed separately.

Inside the research. Every policy, judge and process reward model studied here is an open-weight LLM, named with its version in Section 4.1. Among the transformations, only the two paraphrase types are model-authored: they are written once, before any policy is run, by a fixed author model (Qwen3-30B-A3B-Instruct-2507) and screened by the automatic acceptance rule of Appendix R; every other transformation is deterministic Python. That is the only synthetic text generated for this paper, and its author model, prompts, acceptance rule and rejection rates are reported in full in Appendix R.

In preparing the manuscript. The authors used a general-purpose LLM assistant to draft and edit prose, to restructure sections for length, to help survey related work, and to write the LaTeX and the table- and figure-rendering scripts that turn analysis outputs into the exhibits. It produced no experimental result: every number reported is rendered by the released pipeline from stored run artefacts, as described in the Reproducibility Statement, and was checked against its source. The research question, the method, the formal statements and their proofs, and the experimental design are the authors' own. The authors take full responsibility for all content, including any passage that began as a model suggestion.

References

Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in AI safety. *arXiv preprint arXiv:1606.06565*, 2016.

Anastasios N. Angelopoulos and Stephen Bates. Conformal prediction: A gentle introduction. *Foundations and Trends in Machine Learning*, 16(4):494–591, 2023.

Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc Le, and Charles Sutton. Program synthesis with large language models. *arXiv preprint arXiv:2108.07732*, 2021.

Amos Azaria and Tom Mitchell. The internal state of an LLM knows when it's lying. In *Findings of EMNLP*, 2023.

Victor Barres, Honghua Dong, Soham Ray, Xujie Si, and Karthik Narasimhan. τ^2-bench: Evaluating conversational agents in a dual-control environment. In *Proceedings of ICML*, 2026.

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. Discovering latent knowledge in language models without supervision. In *Proceedings of ICLR*, 2023.

Yanda Chen, Joe Benton, et al. Reasoning models don't always say what they think. *arXiv preprint arXiv:2505.05410*, 2025.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? Try ARC, the AI2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Chandler Davis and W. M. Kahan. The rotation of eigenvectors by a perturbation. III. *SIAM Journal on Numerical Analysis*, 7(1):1–46, 1970.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630:625–630, 2024.

Aaron Grattafiori et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Zhiwei He, Tian Liang, Jiahao Xu, Qiuzhi Liu, Xingyu Chen, Yue Wang, Linfeng Song, Dian Yu, Zhenwen Liang, Wenxuan Wang, Zhuosheng Zhang, Rui Wang, Zhaopeng Tu, Haitao Mi, and Dong Yu. DeepMath-103K: A large-scale, challenging, decontaminated, and verifiable mathematical dataset for advancing reasoning. *arXiv preprint arXiv:2504.11456*, 2025.

Lukas Helff, Quentin Delfosse, David Steinmann, Ruben Härle, Hikaru Shindo, Patrick Schramowski, Wolfgang Stammer, Kristian Kersting, and Felix Friedrich. LLMs gaming verifiers: RLVR can lead to reward hacking. In *ICLR Workshop on Logical Reasoning of Large Language Models*, 2026.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the MATH dataset. In *NeurIPS Datasets and Benchmarks*, 2021.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *Proceedings of ICLR*, 2022.

Kaixuan Huang, Jiacheng Guo, Zihao Li, Xiang Ji, Jiawei Ge, Wenzhe Li, Yingqing Guo, Tianle Cai, Hui Yuan, Runzhe Wang, Yue Wu, Ming Yin, Shange Tang, Yangsibo Huang, Chi Jin, Xinyun Chen, Chiyuan Zhang, and Mengdi Wang. MATH-Perturb: Benchmarking LLMs' math reasoning abilities against hard perturbations. In *Proceedings of ICML*, 2025.

Saurav Kadavath et al. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022.

Ryo Kamoi, Yusen Zhang, Nan Zhang, Sarkar Snigdha Sarathi Das, Ranran Haoran Zhang, Wenpeng Yin, and Rui Zhang. Efficient PRM training data synthesis via formal verification. In *Findings of ACL*, 2026.

Muhammad Khalifa, Rishabh Agarwal, Lajanugen Logeswaran, Jaekyeom Kim, Hao Peng, Moontae Lee, Honglak Lee, and Lu Wang. Process reward models that think. *Transactions on Machine Learning Research*, 2026.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *Proceedings of ICLR*, 2023.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with PagedAttention. In *Proceedings of SOSP*, 2023.

Olivier Ledoit and Michael Wolf. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, 88(2):365–411, 2004.

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

Qintong Li, Leyang Cui, Xueliang Zhao, Lingpeng Kong, and Wei Bi. GSM-Plus: A comprehensive benchmark for evaluating the robustness of LLMs as mathematical problem solvers. In *Proceedings of ACL*, 2024.

Zhenwen Liang, Ruosen Li, Yujun Zhou, Linfeng Song, Dian Yu, Xinya Du, Haitao Mi, and Dong Yu. Your reasoning model is secretly a reward model: Optimization-free verification from experience. In *Proceedings of ACL*, 2026.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *Proceedings of ICLR*, 2024.

Iman Mirzadeh, Keivan Alizadeh, Hooman Shahrokhi, Oncel Tuzel, Samy Bengio, and Mehrdad Farajtabar. GSM-Symbolic: Understanding the limitations of mathematical reasoning in large language models. In *Proceedings of ICLR*, 2025.

Christopher Mohri and Tatsunori Hashimoto. Language models with conformal factuality guarantees. In *Proceedings of ICML*, 2024.

Team Olmo, Allen Institute for AI. Olmo 3. Technical report, Allen Institute for AI, 2025. arXiv:2512.13961.

Hadas Orgad, Michael Toker, Zorik Gekhman, Roi Reichart, Idan Szpektor, et al. LLMs know more than they show: On the intrinsic representation of LLM hallucinations. In *Proceedings of ICLR*, 2025.

Nina Panickssery, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Matt Turner. Steering Llama 2 via contrastive activation addition. In *Proceedings of ACL*, 2024.

Arkil Patel, Satwik Bhattamishra, and Navin Goyal. Are NLP models really able to solve simple math word problems? In *Proceedings of NAACL-HLT*, 2021.

Shishir G. Patil, Huanzhi Mao, Fanjia Yan, Charlie Cheng-Jie Ji, Vishnu Suresh, Ion Stoica, and Joseph E. Gonzalez. The Berkeley function calling leaderboard (BFCL): From tool use to agentic evaluation of large language models. In *Proceedings of ICML*, 2025.

Pouya Pezeshkpour and Estevam Hruschka. Large language models sensitivity to the order of options in multiple-choice questions. In *Findings of NAACL*, 2024.

Victor Quach, Adam Fisch, Tal Schuster, Adam Yala, Jae Ho Sohn, Tommi S. Jaakkola, and Regina Barzilay. Conformal language modeling. In *Proceedings of ICLR*, 2024.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Proceedings of NeurIPS*, 2023.

Amrith Setlur, Chirag Nagpal, Adam Fisch, Xinyang Geng, Jacob Eisenstein, Rishabh Agarwal, Alekh Agarwal, Jonathan Berant, and Aviral Kumar. Rewarding progress: Scaling automated process verifiers for LLM reasoning. In *Proceedings of ICLR*, 2025.

Rulin Shao, Shuyue Stella Li, Rui Xin, Scott Geng, Yiping Wang, Sewoong Oh, Simon Shaolei Du, Nathan Lambert, Sewon Min, Ranjay Krishna, Yulia Tsvetkov, Hannaneh Hajishirzi, Pang Wei Koh, and Luke Zettlemoyer. Spurious rewards: Rethinking training signals in RLVR. *arXiv preprint arXiv:2506.10947*, 2025.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed H. Chi, Nathanael Schärli, and Denny Zhou. Large language models can be easily distracted by irrelevant context. In *Proceedings of ICML*, 2023.

Mohit Shridhar, Xingdi Yuan, Marc-Alexandre Côté, Yonatan Bisk, Adam Trischler, and Matthew Hausknecht. ALFWorld: Aligning text and embodied environments for interactive learning. In *Proceedings of ICLR*, 2021.

Sierra Research. τ^3-bench: Advancing agent evaluation to knowledge and voice. https://sierra.ai/blog/bench-advancing-agent-benchmarking-to-knowledge-and-voice, 2026.

G. W. Stewart and Ji-guang Sun. *Matrix Perturbation Theory*. Academic Press, 1990.

Harsh Trivedi, Tushar Khot, Mareike Hartmann, Ruskin Manku, Vinty Dong, Edward Li, Shashank Gupta, Ashish Sabharwal, and Niranjan Balasubramanian. AppWorld: A controllable world of apps and people for benchmarking interactive coding agents. In *Proceedings of ACL*, 2024.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In *Proceedings of NeurIPS*, 2023.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, 2005.

Peiyi Wang, Lei Li, Zhihong Shao, R. X. Xu, Damai Dai, Yifei Li, Deli Chen, Y. Wu, and Zhifang Sui. Math-Shepherd: Verify and reinforce LLMs step-by-step without human annotations. In *Proceedings of ACL*, 2024.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *Proceedings of ICLR*, 2023.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhu Chen. MMLU-Pro: A more robust and challenging multi-task language understanding benchmark. In *NeurIPS Datasets and Benchmarks*, 2024.

An Yang et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing reasoning and acting in language models. In *Proceedings of ICLR*, 2023.

Shunyu Yao, Noah Shinn, Pedram Razavi, and Karthik Narasimhan. τ-bench: A benchmark for tool-agent-user interaction in real-world domains. In *Proceedings of ICLR*, 2025.

Zheng Yuan, Hongyi Yuan, Chengpeng Li, Guanting Dong, Keming Lu, Chuanqi Tan, Chang Zhou, and Jingren Zhou. Scaling relationship on learning mathematical reasoning with large language models. *arXiv preprint arXiv:2308.01825*, 2023.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah D. Goodman. STaR: Bootstrapping reasoning with reasoning. In *Proceedings of NeurIPS*, 2022.

Lunjun Zhang, Arian Hosseini, Hritik Bansal, Mehran Kazemi, Aviral Kumar, and Rishabh Agarwal. Generative verifiers: Reward modeling as next-token prediction. In *Proceedings of ICLR*, 2025.

Zhenru Zhang, Chujie Zheng, Yangzhen Wu, Beichen Zhang, Runji Lin, Bowen Yu, Dayiheng Liu, Jingren Zhou, and Junyang Lin. The lessons of developing process reward models in mathematical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 10495–10516, 2025.

Chujie Zheng, Zhenru Zhang, Beichen Zhang, Runji Lin, Keming Lu, Bowen Yu, Dayiheng Liu, Jingren Zhou, and Junyang Lin. ProcessBench: Identifying process errors in mathematical reasoning. In *Proceedings of ACL*, 2025.

Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. Large language models are not robust multiple choice selectors. In *Proceedings of ICLR*, 2024.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In *NeurIPS Datasets and Benchmarks*, 2023.

Andy Zou et al. Representation engineering: A top-down approach to AI transparency. *arXiv preprint arXiv:2310.01405*, 2023.

A DEFERRED METHOD MATERIAL

This appendix records the parts of Section 3 that a reader can take on trust: the sampling-noise decomposition that motivates the matched-noise term, the optimality property of the subspace objective, why the equivariant residual is centered and what happens when a family is a singleton, the identity between the reference-free and oracle couplings, the two fitted variants in full, the split-conformal threshold and its coverage guarantee, and a monitoring statistic that no reported number depends on. Every statement proved in Appendix D is stated here in full. Nothing here is needed to follow the main text.

A.1 WHAT INDEPENDENT GENERATION COSTS, AND WHAT D_N CORRECTS

Probes are generated independently of the original rollout. Decompose an arbitrary probe shift as

$$q_g - q_0 = \underbrace{\big(\mathbb{E}[q_g] - \mathbb{E}[q_0]\big)}_{\text{transformation effect}} + \underbrace{\big(q_g - \mathbb{E}[q_g]\big) - \big(q_0 - \mathbb{E}[q_0]\big)}_{\text{resampling fluctuation}}, \tag{13}$$

with expectations over the sampling of the frozen policy at fixed instance. Only the first term is what GAUGE intends to measure. If the two terms are uncorrelated and the resampling fluctuation has the same second moment on ξ_i and $g(\xi_i)$, then

$$\mathbb{E}\,\|q_g - q_0\|_2^2 = \|\mathbb{E}[q_g] - \mathbb{E}[q_0]\|_2^2 + \mathbb{E}\,\|q_N - q_0\|_2^2\,,$$

so the matched noise replicate estimates the nuisance term exactly and subtracting D_N at $\lambda_N = 1$ removes it in expectation.

Neither assumption holds exactly: a transformation can change output-length statistics and hence the sampling variance (a lengthening edit amplifies variance, breaking the equal-second-moment assumption). That is why λ_N is a frozen hyperparameter with a free sign: the selection split can choose to treat resampling movement as a brittleness signal rather than as nuisance. The selection split chooses $\lambda_N < 0$, which reverses the intended role of the term: see Section 4.5. Under deterministic decoding the fluctuation vanishes identically and $D_N = 0$.

A.2 WHAT THE POOLED FEATURE CAN AND CANNOT SEPARATE

Equation (3) replaces a trajectory of arbitrary length with four fixed slots, so it is not injective, and the objection is immediate: a derivation that is sound throughout and one that goes wrong at intermediate steps and recovers could share a pooled feature and receive the same score.

Remark 6 (Pooling invariance). If two rollouts have hidden states agreeing at $t = 0$, $\lfloor T/2 \rfloor$ and T and sharing a mean, then ϕ and every quantity of Section 3.4 built from it coincide. GAUGE cannot order such a pair.

Two structural facts bound what that costs, and two measurements say what it costs here.

Why intermediate steps are not invisible. At a fixed layer of a decoder-only policy, h_t is a function of every token up to t. An intermediate error is therefore not confined to the step that commits it: it enters h_T and the mean through the continuation it conditions. The pair of Remark 6 requires an error that leaves no trace in *any* later state—and such an error did not affect the derivation the model went on to produce. Second, ϕ is never read in isolation: every residual in Equations (9)–(10) is a displacement $q_g - q_0$ between two rollouts on the *same* instance. What has to survive pooling is not the identity of a trajectory but how its summary moves when the task is edited.

Collisions are duplicate trajectories, not confusions. On 11 configurations spanning 2 policies we took every within-prompt pair of correct candidates on the evaluation split (23,301 pairs) and compared pooled features. 4.8% of pairs collide exactly, and 100% of those are the same generated text: not one pair of *distinct* trajectories was mapped to a common feature. Identical trajectories still carry different robustness labels 0.036 of the time, which is the sampling noise of the held-out label

Table 2: Which pooling slots carry the robustness signal. Conditional AUROC of the frozen GAUGE residual when the metric is refitted on a subset of the slots of Equation (3), averaged over 11 configurations from 2 policies.

Pooling slots used	Dimension	Cond. AUROC ↑
h_0 only	$1d_h$	0.710
$h_{\lfloor T/2 \rfloor}$ only	$1d_h$	0.783
h_T only	$1d_h$	0.800
average only	$1d_h$	0.770
h_T + average	$2d_h$	**0.801**
three anchors, no average	$3d_h$	0.790
all four (Definition 4)	$4d_h$	0.791

rollouts (Eq. (1)) and a floor no verifier can pass. Distance between two candidates is, on its own, weakly related to whether they share a label—0.144 label-discordance in the nearest decile against 0.144 elsewhere, and 0.138 against 0.144 in the learned metric—which is consistent with the premise of Section 3.3 and with the verifier reading typed displacement rather than proximity.

No slot carries the signal alone. Refitting the metric on each subset of slots and scoring with it (Table 2; fit split for the subspace, evaluation split for the AUROC, frozen hyperparameters throughout) separates "pooling discards what matters" from "one slot would have done". Every subset ranks robustness well above chance, the weakest at 0.710; the terminal state is the strongest single slot, as the cumulative argument predicts; and the four-slot form of Definition 4 (0.791) is within 0.010 of the best subset (0.801, the terminal state with the average). We keep Definition 4 because the pooling operator was frozen before any evaluation split was read (Appendix T); adopting the best row of a table computed on that split would be selection on the evaluation data, and the margin is inside the per-configuration spread. A finer temporal profile—per-step statistics, or a learned pooling over positions—is the natural extension and is untested here.

A.3 PROPERTIES OF THE SUBSPACE OBJECTIVE

The generalized eigenvectors of Equation (6) are not one heuristic choice among many: they solve the trace objective exactly.

Proposition 7 (Trace optimality). *Let $A = A^\top \succeq 0$ and $B = B^\top \succ 0$ as in Equation (6). Then for any $r \le K$,*

$$\max_{U^\top B U = I_r} \operatorname{tr}(U^\top A U) = \sum_{j=1}^{r} \gamma_j, \tag{14}$$

attained at $U_\star = B^{-1/2}[e_1, \dots, e_r]$, where e_j are the orthonormal eigenvectors of $B^{-1/2} A B^{-1/2}$ with eigenvalues $\gamma_1 \ge \cdots \ge \gamma_K \ge 0$.

Proof. Deferred to Appendix D. □

Remark 8 (Equivalent basis-invariant form). For full-rank $W \in \mathbb{R}^{K \times r}$, the value $\operatorname{tr}[(W^\top B W)^{-1} W^\top A W]$ is unchanged under $W \mapsto WR$ for invertible R, so it depends only on the subspace spanned by W. Choosing a B-orthonormal basis reduces it to the constrained objective of Equation (14). The implementation solves the generalized eigenproblem directly.

Theorem 9 (Subspace stability under covariance estimation noise). *Let $M = B^{-1/2} A B^{-1/2}$ be the population whitened operator, $\widehat{M} = \widehat{B}^{-1/2} \widehat{A} \widehat{B}^{-1/2}$ its estimate from Equations (4)–(6), and $E_r, \widehat{E}_r \in \mathbb{R}^{K \times r}$ orthonormal bases of their top-r eigenspaces, so that the fitted basis is $U = \widehat{B}^{-1/2} \widehat{E}_r$ (Proposition 7). If the spectral gap $\delta_r = \gamma_r(M) - \gamma_{r+1}(M)$ exceeds $\left\| \widehat{M} - M \right\|_2$, then*

$$\left\| \sin \Theta(E_r, \widehat{E}_r) \right\|_{\mathrm{F}} \le \frac{\left\| \widehat{M} - M \right\|_{\mathrm{F}}}{\delta_r - \left\| \widehat{M} - M \right\|_2}. \tag{15}$$

Moreover, with $\widehat{B} = (1-\rho)C_{\mathcal{S}} + \big(\rho\ \mathrm{tr}(C_{\mathcal{S}})/K + \eta\big) I_K$ *as in Equation (5),*

$$\kappa(\widehat{B}) = \frac{(1-\rho)\lambda_{\max}(C_{\mathcal{S}}) + \rho\ \mathrm{tr}(C_{\mathcal{S}})/K + \eta}{(1-\rho)\lambda_{\min}(C_{\mathcal{S}}) + \rho\ \mathrm{tr}(C_{\mathcal{S}})/K + \eta} \le 1 + \frac{(1-\rho)\lambda_{\max}(C_{\mathcal{S}})}{\rho\ \mathrm{tr}(C_{\mathcal{S}})/K + \eta}, \tag{16}$$

which is finite for every $N_{\mathcal{S}}$*, including* $N_{\mathcal{S}} < K$*, where* $C_{\mathcal{S}}$ *is singular.*

Proof. Write $\Delta = \widehat{M} - M$. By Weyl's inequality $\gamma_{r+1}(\widehat{M}) \le \gamma_{r+1}(M) + \|\Delta\|_2$, so the eigenvalues of M retained in E_r are separated from those discarded by $\widehat{E}_r$ by at least $\delta_r - \|\Delta\|_2 > 0$; the Davis–Kahan $\sin\Theta$ theorem (Davis and Kahan, 1970; Stewart and Sun, 1990) then gives $\left\|\sin\Theta(E_r, \widehat{E}_r)\right\|_{\mathrm{F}} \le \|\Delta E_r\|_{\mathrm{F}} / (\delta_r - \|\Delta\|_2) \le \|\Delta\|_{\mathrm{F}} / (\delta_r - \|\Delta\|_2)$. For the second claim, $\widehat{B}$ shares its eigenvectors with $C_{\mathcal{S}}$ and has eigenvalues $(1-\rho)\lambda_i(C_{\mathcal{S}}) + \rho\ \mathrm{tr}(C_{\mathcal{S}})/K + \eta$; the ratio of the largest to the smallest is the equality in Equation (16), and $\lambda_{\min}(C_{\mathcal{S}}) \ge 0$ gives the bound. □

The first bound makes the fitted subspace stable exactly where the directions it keeps are separated from those it discards; the second is why shrinkage, not the ridge alone, keeps $\widehat{B}^{-1/2}$ well conditioned when there are fewer paraphrase displacements than dimensions.

Remark 10 (Why $\mathcal{C}$ raises sensitivity in the fit and costs a candidate in the score). Equation (6) selects directions along which control and equivariant edits move the representation *more* than semantics-preserving edits do, so the subspace is maximally sensitive to those edits. The score then penalizes a candidate that moves far in this sensitive subspace under an edit that should not change its answer. For $\mathcal{V}$ the score penalizes only disagreement around the common equivariant shift, not the shift itself, which is why Equation (10) is centered.

Remark 11 (Consistency identity for the normalizers). Since $q_{i,g} - q_i^0 = U^\top \delta_{i,g}$, Equations (4) and (7) give $\sigma_G^2 = \mathrm{tr}(U^\top C_G U)$. With $U^\top B U = I_r$ and Equation (5) this closes to

$$\sigma_{\mathcal{S}}^2 = \frac{r - \kappa\,\|U\|_F^2}{1-\rho}, \qquad \kappa = \rho\frac{\mathrm{tr}(C_{\mathcal{S}})}{K} + \eta.$$

We use it as an implementation check: a fitted $\sigma_{\mathcal{S}}$ violating the identity indicates that scales and subspace were estimated on different data or with different weighting. Across every configuration reported here the two sides agree to 10^{-15} relative error.

A.4 Why the equivariant residual is centered, and the singleton fallback

The equivariant residual $D_{\mathcal{V}}$ of Equation (10) is the only one of the four that subtracts a mean before measuring length. The asymmetry is deliberate, and it follows from what the three families are for.

Under a semantics-preserving or control edit the task is unchanged, so a genuine derivation should produce the *same* answer by the *same* route: any latent movement is evidence against it, and $D_{\mathcal{S}}$ and $D_{\mathcal{C}}$ therefore measure raw displacement from q_0. An answer-equivariant edit is the opposite case. It changes the task on purpose, and a candidate that is tracking the task *must* change its answer to $\tau_v(\widehat{a}_0)$. Its representation is expected to move, and the part of that movement which every candidate shares is a property of the edit—of how the model as a whole encodes "scale the quantity by c"—not of whether this particular derivation is causal. Scoring $\|d_v\|_2^2$ directly would charge a candidate for doing exactly what the edit asks, and would make the residual grow with the magnitude of the answer change, which is an artifact of how the transformation is parameterized rather than a fact about robustness.

Centering removes precisely that component. Write the equivariant displacement of probe v as $d_v = \bar{d}_{\mathcal{V}} + e_v$, where $\bar{d}_{\mathcal{V}} = |\mathcal{V}_i|^{-1} \sum_{v \in \mathcal{V}_i} d_v$ is the shift shared across the family and e_v is the within-family deviation with $\sum_v e_v = 0$. Up to the coupling weights and the scale $\sigma_{\mathcal{V}}$, $D_{\mathcal{V}} = |\mathcal{V}_i|^{-1} \sum_v \|e_v\|_2^2$ is the trace of the empirical covariance of the equivariant displacements. Two properties make this the right quantity, and both are exact rather than approximate:

1. **Translation invariance.** Replacing $d_v \mapsto d_v + c$ for a common $c \in \mathbb{R}^r$ leaves $D_{\mathcal{V}}$ unchanged. Whatever the edit does to the representation uniformly—including the systematic, correct response to the task change—cannot contribute to the score.

2. **Zero exactly at consistency.** $D_{\mathcal{V}} = 0$ if and only if every equivariant probe displaces the trajectory identically, $d_v = \bar{d}_{\mathcal{V}}$ for all $v \in \mathcal{V}_i$. The residual therefore measures *inconsistency* across the equivariant family: a derivation that is causally tied to the task adapts coherently to each member, while a shortcut adapts erratically, and it is the scatter that separates them.

Because the family mean is estimated from the same $|\mathcal{V}_i|$ displacements it is subtracted from, $D_{\mathcal{V}}$ is an uncorrected sample scatter: its expectation carries the usual $(1 - 1/|\mathcal{V}_i|)$ factor relative to the population value. We do not correct it, because the probe budget fixes $|\mathcal{V}_i| = 2$ for every non-singleton instance in every configuration (Table 33), so the factor is a single constant per configuration and every AUROC we report is computed within a configuration—a constant positive factor on one feature cannot change a within-configuration ranking.

The one case the centered form cannot handle is $|\mathcal{V}_i| = 1$: a single displacement equals its own mean, so Equation (10) is identically zero and the feature would silently vanish. The implementation then falls back to the uncentered distance

$$D_{\mathcal{V}} = w_v \frac{\|d_v\|_2^2}{\sigma_{\mathcal{V}}^2 + \varepsilon}. \tag{17}$$

This is a fallback distance, not an estimate of within-family scatter—by the argument above it includes the shared shift—and the two are on different scales. Singleton instances are therefore counted and reported separately (Table 33), and they arise only where a task admits a single equivariant edit.

A.5 The reference-free coupling equals the oracle coupling on correct candidates

Equation (8) defines two couplings: w_g^{oracle}, which compares a probe answer against the true transformed reference $a_g^\star$, and w_g^{free}, which compares it against the map-pushed candidate answer $\tau_g(\widehat{a}_0)$ and so consults no reference at all. They agree wherever the verification question is posed.

Proposition 12 (Coupling identity under quotient equivalence)**.** *Let m and m_g be equivalence relations on $\mathcal{A}$ and g a valid edit (Definition 2). If candidate ζ_0 is nominally correct ($m(\widehat{a}_0, a^\star) = 1$), then $m_g(\widehat{a}_g, \tau_g([\widehat{a}_0]_m)) = m_g(\widehat{a}_g, a_g^\star)$ for every probe answer $\widehat{a}_g \in \mathcal{A}$; hence $w_g^{\text{free}} = w_g^{\text{oracle}}$ on every nominally correct candidate.*

Proof. Deferred to Appendix D. □

Since the conditional metric of Section 4.2 ranks only candidates that are already nominally correct, the reference-free score is not an approximation of the oracle score on that population: it is the same number. Totality of τ_g (Definition 2) is what makes this possible, since $\tau_g(\widehat{a}_0)$ must be defined for a candidate answer that may differ from $a^\star$. On nominally incorrect candidates the two couplings may differ, which is visible empirically as the gap between the free and oracle variants in Table 6.

A.6 The two fitted variants

Definition 13 (GAUGE-LR)**.** Let

$$\psi(\xi_i, \zeta_{ij}) = \big(D_{\mathcal{S}},\, D_{\mathcal{C}},\, D_{\mathcal{V}},\, D_N,\, \bar{\alpha}_{\mathcal{S}},\, \bar{\alpha}_{\mathcal{V}},\, \bar{\alpha}_{\mathcal{C}}\big) \in \mathbb{R}^7,$$

where the $D_\bullet$ are evaluated at $\lambda_{\text{corr}} = 1$, so they carry only geometry, and $\bar{\alpha}_G$ is the fraction of family-G probes whose answer matches $\tau_g(\widehat{a}_0)$. GAUGE-LR is the logistic score $\langle\theta, \psi\rangle + \theta_0$, fitted by regularized logistic regression against the robustness label on the *selection* split, disjoint from calibration and evaluation.

GAUGE-LR+ appends three quantities the deployed system already computes—the candidate's mean and minimum token log-probability and $\log(1 + \text{tokens})$—giving $\psi^+ \in \mathbb{R}^{10}$, and is fitted identically. It is the variant reported as the strongest verifier in Section 4.4.

We report the fixed-coefficient score first because the gap between it and GAUGE-LR is itself evidence about which parts of the construction are load-bearing.

Definition 14 (Distilled probe-free score)**.** Let $z_{ij} = z(\phi(H_\theta(\zeta_{ij}))) \in \mathbb{R}^K$ be the candidate's own PCA coordinate and $\mathcal{T}$ a teacher: R_{GAUGE} for GAUGE-D, GAUGE-LR+ for GAUGE-D+. The student is the ridge predictor

$$\begin{aligned} R^{\mathrm{D}}_{\mathrm{GAUGE}}(\zeta_{ij}) &= \langle W, \widetilde{z}_{ij} \rangle + b, \\ e_{ij} &= R^{\mathrm{D}}_{\mathrm{GAUGE}}(\zeta_{ij}) - \mathcal{T}(\xi_i, \zeta_{ij}), \\ (W, b) &= \arg\min_{W,b} \sum_{(i,j)\in \mathrm{fit}} e_{ij}^2 + \alpha_{\mathrm{ridge}} \|W\|_2^2 , \end{aligned} \tag{18}$$

with $\widetilde{z}$ standardized by fit-split statistics. For GAUGE-D+ the input is z_{ij} extended by the candidate's mean and minimum token log-probability and log length, which are free at deployment. The head is fitted on the fit split, and α_{ridge} is chosen on the calibration split by AUROC against robustness labels; like the frozen coefficients it therefore uses labels offline, and at test time it reads only the candidate's own forward pass.

A.7 Relation to answer-agreement verifiers, and a monitoring statistic

Remark 15 (Relation to answer-agreement verifiers). At $\lambda_{\mathrm{corr}} = 0$ the score reduces to a sum, over probes whose answer *disagrees* with the map-pushed candidate answer, of squared representational drift. It therefore refines "perturbation agreement" verifiers—which merely count disagreements—by weighting each disagreement with how far the trajectory actually moved, and by adding the equivariant and control families that agreement counting cannot use. At $\lambda_{\mathrm{corr}} = 1$ it ignores answers entirely and is a pure representation-space statistic. Both extremes are reported as ablations.

Alongside the residual, the implementation returns a χ^2 tail probability computed from $D_{\mathcal{S}} + D_{\mathcal{C}}$ against $r(|\mathcal{S}_i| + |\mathcal{C}_i|)$ degrees of freedom. That value would be exact if the projected shifts were isotropic Gaussian with the fitted scales and all coupling weights were one. Neither holds. We log it as a monitoring diagnostic, never as a calibrated p-value; no selection decision, threshold or reported number depends on it. The distribution-free guarantee of Theorem 16 carries the statistical claim.

A.8 Split-conformal threshold and coverage guarantee

Section 3.5 calibrates the acceptance threshold on a calibration split disjoint from the fit, selection, and evaluation splits, with the complete pipeline (policy, transformations, metric, and normalizers) frozen and nonconformity taken to be the instability score $\nu(\xi, \zeta) = R_{\mathrm{GAUGE}}(\xi, \zeta)$. For target error rate $\alpha \in (0, 1)$ and sorted calibration scores $\nu_{(1)} \le \cdots \le \nu_{(n_{\mathrm{cal}})}$, the calibrated threshold is

$$k = \lceil (n_{\mathrm{cal}} + 1)(1 - \alpha) \rceil, \qquad \widehat{q}_{1-\alpha} = \begin{cases} \nu_{(k)}, & k \le n_{\mathrm{cal}} \\ +\infty, & k = n_{\mathrm{cal}} + 1. \end{cases} \tag{19}$$

Theorem 16 (Marginal score coverage)**.** *If calibration candidates and an unobserved test candidate are exchangeable under the frozen pipeline, then:*

$$\mathbb{P}\{\nu_{\mathrm{new}} \le \widehat{q}_{1-\alpha}\} \ge 1 - \alpha. \tag{20}$$

Proof. Deferred to Appendix D. □

Theorem 16 certifies distribution-free control over how often the score exceeds its threshold: at most a fraction α of future candidates exceed $\widehat{q}_{1-\alpha}$ under exchangeability. The guarantee is marginal over the draw of the calibration set, not conditional on it, and it is a statement about the score alone—nominal correctness is enforced jointly during rollout filtering (Algorithm 3). Partitioning the splits by prompt group keeps calibration and test prompts disjoint, but candidates sampled from one prompt share context and are conditionally dependent, so the exchangeable unit is the prompt group rather than the candidate. Section 3.5 therefore calibrates on one candidate per calibration group, drawn uniformly at random and independently of the scores; because a prompt's candidates are i.i.d. draws, that representative is exchangeable with any candidate of a new prompt, and the theorem applies to every test candidate marginally. Empirical coverage is reported in Appendix H (Table 16).

B POLICY IMPROVEMENT, AND DEFERRED SETUP DETAILS

AGENTIC REASONING SETUP

The instance tuple, the rollout, terminal correctness and the candidate set are defined in Section 3.1; Definition 1 states the verification problem. Two details omitted there for space: the observation is drawn as $o_\ell \sim \mathfrak{E}(\cdot \mid x, r_{1:\ell}, u_{1:\ell}, o_{1:\ell-1})$, a point mass in the deterministic environments used here, and all segments $r_{1:L+1}, u_{1:L}$ are sampled from π_θ conditioned on the preceding context.

POLICY IMPROVEMENT

GAUGE-RFT. Let $\mathcal{D}_{\mathrm{RFT}}$ be the selected set and let $\mathcal{G}(\zeta)$ index model-generated tokens, excluding tool observations. The supervised fine-tuning objective is

$$\mathcal{L}_{\mathrm{RFT}}(\theta) = -\sum_{(\xi,\zeta)\in\mathcal{D}_{\mathrm{RFT}}} \sum_{t\in\mathcal{G}(\zeta)} \log \pi_\theta(v_t \mid c_t), \tag{21}$$

where v_t is the generated token at position t and c_t is its preceding context. This is rejection fine-tuning in the sense of Zelikman et al. (2022) and Yuan et al. (2023), with the acceptance rule replaced by Algorithm 3. Because the free coupling needs no reference answers, we also report a fully label-free variant that drops the correctness filter and keeps the lowest-residual candidate per prompt.

DPO variant (Rafailov et al., 2023). For preference pair (ζ^+, ζ^-), define the sequence log-probability over model-generated tokens $\ell_\theta(\zeta \mid \xi) = \sum_{t\in\mathcal{G}(\zeta)} \log \pi_\theta(v_t \mid c_t)$, and write ℓ_{ref} for the same quantity under the frozen reference policy. The DPO objective is

$$\begin{aligned} \mathcal{L}_{\mathrm{DPO}}(\theta) &= -\,\mathbb{E}_{(\xi,\zeta^+,\zeta^-)} \log \sigma(\beta_{\mathrm{DPO}}\,\Delta_\theta)\,, \\ \Delta_\theta &= r_\theta(\zeta^+) - r_\theta(\zeta^-), \\ r_\theta(\zeta) &= \ell_\theta(\zeta \mid \xi) - \ell_{\mathrm{ref}}(\zeta \mid \xi), \end{aligned} \tag{22}$$

where $\sigma(u) = (1+e^{-u})^{-1}$ is the logistic function—distinct from the family scales σ_G, which always carry a family subscript.

GRPO variant (Shao et al., 2024). For online optimization, the scalar reward combines correctness and a normalized residual,

$$\mathcal{R}(\xi,\zeta) = O(\xi,\zeta) - \beta_R \widetilde{R_{\mathrm{GAUGE}}}(\xi,\zeta), \qquad \beta_R \geq 0, \tag{23}$$

where $\widetilde{R_{\mathrm{GAUGE}}} \in [0,1]$ is the residual rescaled within the sampled group. The rescaling matters, because R_{GAUGE} has no intrinsic units. Let $\mathcal{Z}_{\mathrm{grp}}$ be the group of rollouts sampled for one prompt at one update; we use group min–max rescaling, writing $\Delta_{\mathrm{grp}} = R^{\max}_{\mathrm{GAUGE}} - R^{\min}_{\mathrm{GAUGE}}$ for the group range,

$$\widetilde{R_{\mathrm{GAUGE}}}(\xi,\zeta) = \begin{cases} \big(R_{\mathrm{GAUGE}}(\xi,\zeta) - R^{\min}_{\mathrm{GAUGE}}\big)/\Delta_{\mathrm{grp}}, & \Delta_{\mathrm{grp}} > 0, \\ 0, & \text{otherwise}, \end{cases} \tag{24}$$

so that the best rollout in the group incurs no residual penalty and a group whose members are indistinguishable under GAUGE reduces to a pure outcome reward. Equation (24) is invariant to increasing affine transformations of R_{GAUGE}. Using GAUGE-D as the reward makes this loop practical, since it needs no probe rollouts inside the RL loop.

Proposition 17 (Correctness dominance). *If $\beta_R < 1$, then for any two rollouts ζ, ζ' in the same group with $O(\xi,\zeta) = 1$ and $O(\xi,\zeta') = 0$ we have $\mathcal{R}(\xi,\zeta) > \mathcal{R}(\xi,\zeta')$.*

Proof. Deferred to Appendix D. □

INTERPRETATION AND FORMAL SCOPE

The generalized eigenproblem is a statement about contrastive second moments. It does not, without further assumptions, identify a unique causal reasoning variable. The following proposition gives a precise sufficient condition for the intended interpretation.

Algorithm 1 Fit the implemented GAUGE task subspace

Require: Original and transformed fit trajectories; PCA rank K; subspace rank r; ρ, η, λ_C.
1: Pool each trajectory using Equation (3).
2: Fit μ and P from original fit trajectories.
3: Project all pooled features with $z(\phi) = P^\top(\phi - \mu)$.
4: Construct $\Delta_{\mathcal{S}}$, $\Delta_{\mathcal{V}}$, and $\Delta_{\mathcal{C}}$.
5: Compute $\widehat{\Sigma}_G$ using Equations (4)–(5).
6: Solve $(\widehat{\Sigma}_{\mathcal{V}} + \lambda_C \widehat{\Sigma}_{\mathcal{C}})u = \gamma\, \widehat{\Sigma}_{\mathcal{S}} u$.
7: B-normalize the leading r generalized eigenvectors to obtain U.
8: Fit and freeze the score normalizers of Equation (7).
9: **return** μ, P, U, normalizers, and fit metadata.

Algorithm 2 Score a candidate with the implemented GAUGE rule

Require: Original rollout; typed transformed rollouts; known answer maps $\{\tau_g\}$; matched noise rollout when enabled; fitted (μ, P, U); frozen scales and weights.
1: Pool and project the original rollout to obtain q_0 and its answer $\widehat{a}_0$.
2: Pool and project all $\mathcal{S}$, $\mathcal{V}$, and $\mathcal{C}$ rollouts.
3: Compute w_g from Equation (8); the deployed variant uses $\tau_g(\widehat{a}_0)$ and never $a^\star$.
4: Compute $D_{\mathcal{S}}$, $D_{\mathcal{C}}$, and $D_{\mathcal{V}}$, using the singleton fallback when required.
5: **if** a matched noise trajectory is enabled **then**
6: Compute D_N using Equation (10).
7: **else**
8: Set $D_N \leftarrow 0$.
9: **return** $R_{\mathrm{GAUGE}} = D_{\mathcal{S}} + D_{\mathcal{C}} + \lambda_V D_{\mathcal{V}} - \lambda_N D_N$, all component values, and the auxiliary diagnostic.

Proposition 18 (Idealized task–nuisance spectral separation)**.** *Let $M = B^{-1/2} A B^{-1/2}$ and suppose the whitened space decomposes orthogonally as $\mathbb{R}^K = \mathcal{H}_{\mathrm{task}} \oplus \mathcal{H}_{\mathrm{nuis}}$, with both subspaces invariant under M and $\dim \mathcal{H}_{\mathrm{task}} \geq r$. Write $\lambda_j(\cdot)$ for the j-th largest eigenvalue of a symmetric operator. If*

$$\lambda_r(M|_{\mathcal{H}_{\mathrm{task}}}) > \lambda_1(M|_{\mathcal{H}_{\mathrm{nuis}}})\,, \tag{25}$$

then every r-dimensional eigenspace of M associated with its r largest eigenvalues is contained in $\mathcal{H}_{\mathrm{task}}$. Consequently the fitted subspace satisfies $\mathrm{span}(U_\star) \subseteq B^{-1/2}\mathcal{H}_{\mathrm{task}}$.

Proof. Deferred to Appendix D. □

Remark 19. The conclusion is stated in whitened coordinates. Because $B^{-1/2}$ is not orthogonal, the fitted subspace is not literally a subspace of $\mathcal{H}_{\mathrm{task}}$ in the original coordinates; it is its image under whitening. The proposition is also deliberately conditional—it assumes the decomposition rather than establishing it. The subspace and control-drift ablations in Section 4.5 test whether the intended separation appears in the studied models.

C ALGORITHMS

D PROOFS

Proof of Proposition 7. For any feasible U, set $V = B^{1/2}U$. Then $V^\top V = U^\top B U = I_r$, and $\mathrm{tr}(U^\top A U) = \mathrm{tr}(V^\top M V)$. Let $M = E \Gamma E^\top$, where $E = [e_1, \ldots, e_K]$ and $\Gamma = \mathrm{diag}(\gamma_1, \ldots, \gamma_K)$. Define $a_j = e_j^\top V V^\top e_j = \left\| V^\top e_j \right\|_2^2$. Since $VV^\top$ is an orthogonal projector of rank r, $0 \leq a_j \leq 1$ and $\sum_j a_j = \mathrm{tr}(VV^\top) = \mathrm{tr}(V^\top V) = r$. Therefore

$$\mathrm{tr}(V^\top M V) = \mathrm{tr}(M V V^\top) = \sum_{j=1}^{K} \gamma_j a_j \leq \sum_{j=1}^{r} \gamma_j,$$

Algorithm 3 Build the conformal GAUGE-RFT set

Require: Frozen base policy and score; disjoint calibration and selection subsets.

1: **for** each instance ξ_i **do**
2: Sample J candidate rollouts $\{\zeta_{ij}\}_{j=1}^{J}$.
3: Generate the typed transformed rollouts and the noise replicate for each candidate.
4: Compute $R_{\mathrm{GAUGE}}(\xi_i, \zeta_{ij})$ for every candidate.
5: Compute $\widehat{q}_{1-\alpha}$ from the calibration scores of one candidate per prompt group, drawn uniformly at random.
6: Retain candidate (ξ_i, ζ_{ij}) when $O(\xi_i, \zeta_{ij}) = 1$ and $R_{\mathrm{GAUGE}}(\xi_i, \zeta_{ij}) \le \widehat{q}_{1-\alpha}$.
7: If several candidates survive for one instance, use the pre-declared rule: retain the lowest-residual survivor.
8: Fine-tune on the retained model-generated tokens.

because a weighted sum with $a_j \in [0,1]$ and $\sum_j a_j = r$ is maximized by placing unit weight on the r largest γ_j. Equality is attained for $V = [e_1, \dots, e_r]$. Mapping back gives $U_\star = B^{-1/2}V$. Finally, $Me_j = \gamma_j e_j$ implies $A(B^{-1/2}e_j) = \gamma_j B(B^{-1/2}e_j)$, so the columns of $U_\star$ are the desired generalized eigenvectors. If $\gamma_r > \gamma_{r+1}$, the maximizing choice of $\{a_j\}$ is forced, so the optimal subspace is unique. □

Proof of Proposition 12. Applying Equation (2) with $a = \widehat{a}_0$ gives $m_g(\tau_g([\widehat{a}_0]_m), a_g^\star) = m(\widehat{a}_0, a^\star) = 1$, so the class $\tau_g([\widehat{a}_0]_m)$ is the m_g-class of $a_g^\star$. Since m_g is an equivalence predicate on $\mathcal{A}$, any $\widehat{a}_g$ is m_g-equivalent to one of them exactly when it is m_g-equivalent to the other. □

Proof of Theorem 16. Consider the $n_{\mathrm{cal}} + 1$ exchangeable scores consisting of the calibration scores and ν_{new}. After independent uniform tie breaking for the proof, the rank R_{new} of the new score is uniform on $\{1, \dots, n_{\mathrm{cal}} + 1\}$. Suppose $k \le n_{\mathrm{cal}}$. On the event $R_{\mathrm{new}} \le k$, at most $k - 1$ calibration scores fall strictly below ν_{new}, which forces $\nu_{(k)} \ge \nu_{\mathrm{new}}$; hence

$$\mathbb{P}\{\nu_{\mathrm{new}} \le \widehat{q}_{1-\alpha}\} \ge \mathbb{P}\{R_{\mathrm{new}} \le k\} = \frac{k}{n_{\mathrm{cal}} + 1} \ge 1 - \alpha,$$

where the last inequality is the definition of k as a ceiling. If $k = n_{\mathrm{cal}} + 1$, the threshold is $+\infty$ and the statement is immediate. The deterministic weak-inequality rule is conservative when ties occur. □

Proof of Proposition 17. By Equation (24), $\beta_R \widetilde{R_{\mathrm{GAUGE}}} \in [0, \beta_R]$ for every group member. Hence $\mathcal{R}(\xi, \zeta) \ge 1 - \beta_R$ and $\mathcal{R}(\xi, \zeta') \le 0$, and $1 - \beta_R > 0$. □

Proof of Proposition 18. The invariance assumption makes M block diagonal with respect to the orthogonal decomposition. Equation (25) states that the first r eigenvalues of the task block exceed every eigenvalue of the nuisance block. Therefore the r largest eigenvalues of M are all task-block eigenvalues, and any associated eigenvectors lie in $\mathcal{H}_{\mathrm{task}}$; this holds even if $\gamma_r = \gamma_{r+1}$, because both would then still belong to the task block. Applying $B^{-1/2}$, Proposition 7 maps this eigenspace to the leading generalized eigenspace of (A, B). □

E EXACT VALUES FOR THE MAIN-TEXT FIGURES

Figure 2 is rendered from Table 3, and Figure 3 from the published-baseline and control rows of Table 6 (Appendix F); every plotted number appears in one of the two.

F DEFERRED VERIFIER RESULTS

This appendix expands Section 4.4 and changes none of its orderings: the label budget, the deployment regimes, the per-configuration values behind the aggregate, the secondary metric, the operating-point view and the paired comparison.

Table 3: Accuracy, strict robust accuracy and hack rate under the held-out edit set, with one rollout per instance. Intervals are 95% prompt-level cluster bootstraps and are given in Appendix W. "Per-edit pass" is the mean pass rate of a single held-out edit, which the conjunctive label compounds.

Model	Task	Acc. ↑ (%)	Strict rob. acc. ↑ (%)	Hack rate ↓ (%)	Per-edit pass ↑ (%)
Qwen3-4B	GSM8K	93.3	88.2	5.5	92.6
	GSM-Symbolic-P2	77.6	58.1	25.2	76.8
	MATH-500	86.4	77.5	10.3	82.8
	SVAMP	93.3	88.6	5.1	92.6
	GSM-IC	97.9	94.0	3.9	97.6
	MMLU-Pro	64.6	53.9	16.6	64.6
Qwen3-8B	GSM8K	92.4	83.1	10.2	90.7
	GSM-Symbolic-P2	72.4	56.2	22.4	71.9
	MATH-500	85.4	72.5	15.1	81.7
	MMLU-Pro	61.4	47.9	22.1	62.0
	τ-bench	22.6	7.8	65.4	18.3
Llama-3.1-8B	GSM8K	85.1	68.4	19.7	82.7
	GSM-Symbolic-P2	49.5	17.6	64.4	43.0
	MATH-500	48.9	28.2	42.3	48.4
	MMLU-Pro	42.9	23.6	45.0	40.0
Olmo-3-7B	GSM8K	89.2	77.3	13.3	88.0
	GSM-Symbolic-P2	70.0	37.1	46.9	63.7
	MATH-500	82.5	72.9	11.7	81.7
Qwen3.8-27B	GSM8K	96.3	91.0	5.5	95.1
	GSM-Symbolic-P2	96.2	83.8	12.9	93.8
	MATH-500	98.6	92.9	5.8	96.3
	SVAMP	93.8	90.5	3.6	95.0
	GSM-IC	99.5	96.7	2.9	98.7
	MMLU-Pro	74.6	64.6	13.4	74.1
	τ-bench	22.7	9.3	59.0	14.7
Mean		75.9	63.3	21.9	73.9

F.1 How many labelled prompts GAUGE-LR+ needs

GAUGE-LR+ fits ten coefficients against robustness labels on the selection split. The practical question is not whether it uses labels but how many, since the fit happens once per configuration and never again at deployment. We subsample the selection split to a fixed number of labelled *prompts*, refit, and score the evaluation split, averaging over eight draws per budget and over 17 configurations. Fifty labelled prompts recover almost all of the full-split fit and one hundred saturate it (Table 4). Absolute values sit above the headline numbers of Table 7 because this experiment fits once on a fixed split whereas the reported pipeline cross-fits over folds; the column to read is the *fraction of the full-split fit*, which is unaffected by that difference.

For comparison, the process reward models used as baselines are trained on step-level annotations over thousands of solutions, and the supervised hidden probe needs a robustness label for every candidate it is fitted on. On that scale, fifty prompts is a rounding error, which is why we treat GAUGE-LR+ as a deployable operating point rather than an upper bound that cannot be reached in practice.

F.2 Deployment regimes

Table 5 summarises what each GAUGE variant consumes—robustness labels when fitting, counterfactual probes when scoring—against its conditional AUROC.

Table 4: Label efficiency of GAUGE-LR+. Conditional AUROC after refitting on a fixed number of labelled prompts; mean over eight draws and 17 configurations. The last column is the fraction of the full-split fit.

Labelled prompts	Cond. AUROC ↑	% of full fit
10	0.819	93.0%
25	0.855	97.1%
50	0.870	98.9%
100	0.880	100.0%
200	0.880	100.0%
full split	0.880	100.0%

Table 5: Four GAUGE deployment regimes. Mean conditional AUROC with 95% prompt-cluster bootstrap intervals across 23 configurations. *Labels*: fitted on robustness labels. *Probes*: generates counterfactual rollouts at inference. All four are reference-free at test time.

Variant	Labels	Probes	AUROC ↑ [95% CI]
GAUGE-LR+ peak accuracy	yes	yes	**0.770** [0.735, 0.802]
GAUGE (label-free) no label in the score	**no**	yes	0.745 [0.714, 0.775]
GAUGE-D+ zero inference generation	teacher	**no**	0.730 [0.697, 0.765]
GAUGE-D no label in the score, no probes	**no**	**no**	0.671 [0.635, 0.709]

Table 6: Aggregate comparison of GAUGE-LR+ against every baseline across all model–task configurations, on the conditional robustness metric. Differences are paired on identical candidates; the interval is a bootstrap over configurations and the test is a Wilcoxon signed-rank test over configurations, Holm-corrected across the baseline family (Appendix L). * marks Holm significance at $\alpha = 0.05$.

Baseline	Mean ΔAUROC ↑	95% CI	Wins	Wilcoxon p (Holm)
LLM judge (self)	+0.224*	[0.194, 0.253]	23/23	2.4e-07 (6.7e-06)
LLM judge (strong, 30B)	+0.221*	[0.191, 0.248]	23/23	2.4e-07 (6.7e-06)
Untyped hidden drift	+0.216*	[0.167, 0.262]	22/23	1.2e-06 (2.5e-05)
Trajectory noise drift	+0.200*	[0.169, 0.230]	23/23	2.4e-07 (6.7e-06)
PRM (mean step)	+0.197*	[0.151, 0.243]	21/23	1.7e-06 (3.3e-05)
GAUGE (hidden only)	+0.186*	[0.149, 0.221]	22/23	7.2e-07 (1.6e-05)
PRM (min step)	+0.186*	[0.139, 0.232]	20/23	6e-06 (0.00011)
P(True)	+0.183*	[0.141, 0.226]	23/23	2.4e-07 (6.7e-06)
Min token logprob	+0.116*	[0.075, 0.150]	21/23	0.00013 (0.002)
Supervised hidden probe	+0.111*	[0.053, 0.163]	20/23	0.00064 (0.007)
Control-probe agreement	+0.110*	[0.084, 0.132]	22/23	6e-06 (0.00011)
Sequence confidence	+0.110*	[0.060, 0.159]	21/23	0.00018 (0.0027)
GAUGE-D (probe-free)	+0.099*	[0.069, 0.130]	21/23	4.5e-06 (8.6e-05)
Response length (shorter)	+0.087*	[0.055, 0.118]	20/23	0.00021 (0.003)
Terminal hidden distance	+0.083*	[0.048, 0.119]	20/23	0.0003 (0.0039)
Paraphrase agreement	+0.073*	[0.043, 0.102]	21/23	0.00041 (0.0049)
GAUGE-D (distilled from LR)	+0.053*	[0.020, 0.083]	17/23	0.006 (0.043)
Self-consistency	+0.043*	[0.012, 0.073]	19/23	0.0049 (0.043)
GAUGE-D+ (probe-free, from LR+)	+0.039	[0.005, 0.071]	17/23	0.01 (0.061)
GAUGE unstratified	+0.036*	[0.007, 0.062]	17/23	0.0043 (0.043)
P(True) (strong, 30B)	+0.036	[0.003, 0.065]	18/23	0.018 (0.068)
GAUGE w/o typed contrast	+0.030*	[0.006, 0.052]	18/23	0.0043 (0.043)
GAUGE (label-free)	+0.024	[-0.002, 0.047]	17/23	0.014 (0.068)
GAUGE (oracle coupling)	+0.023	[-0.004, 0.046]	17/23	0.02 (0.068)
GAUGE-LR (fitted combination)	+0.018	[-0.007, 0.040]	17/23	0.035 (0.068)
Typed-probe agreement	+0.016	[-0.010, 0.037]	16/23	0.015 (0.068)
Random	+0.276*	[0.242, 0.308]	23/23	2.4e-07 (6.7e-06)
Outcome oracle (orig. prompt)	+0.260*	[0.224, 0.292]	22/23	4.8e-07 (1.1e-05)

F.3 PER-CONFIGURATION AND UNCONDITIONAL AUROC

Table 6 gives every row of Figure 3 exactly, together with the rows the figure omits: the ablation variants of GAUGE itself, and the two remaining published baselines.

Table 7: Primary result, per policy. AUROC for predicting robustness *given* terminal correctness: among trajectories that produced the right answer, which survive the held-out edits. Values are means over the tasks of each policy, and the Mean column averages the five policies, so it differs slightly from the per-configuration means of Section 4; per-task values with prompt-level bootstrap intervals are in Appendix W (Table 34). Terminal correctness on the original prompt is at chance here by construction, which is the check that this metric isolates robustness from accuracy.

Verifier	Qwen3 4B	Qwen3 8B	Llama 8B	Olmo 7B	Qwen3.8 27B	Mean ↑	Needs ↑
GAUGE-LR+ (probes + conf.)	**0.825**	**0.751**	**0.794**	**0.748**	0.722	**0.768**	policy only
Typed-probe agreement	0.783	0.742	0.788	0.727	0.722	0.753	policy only
GAUGE (label-free)	0.788	0.723	0.773	0.727	0.708	0.744	policy only
Self-consistency	0.735	0.704	0.784	0.696	0.709	0.726	policy only
P(True) (30B)	0.788	0.664	0.750	0.670	**0.748**	0.724	30B model
Paraphrase agreement	0.718	0.701	0.731	0.671	0.662	0.697	policy only
Response length (shorter)	0.737	0.576	0.700	0.651	0.702	0.673	policy only
Sequence confidence	0.720	0.599	0.596	0.646	0.690	0.650	policy only
PRM (min step)	0.601	0.556	0.584	0.550	0.600	0.578	PRM
LLM judge (30B)	0.558	0.528	0.568	0.512	0.560	0.545	30B model
Random	0.496	0.509	0.486	0.477	0.494	0.493	policy only

Table 8: Verifier AUROC against the held-out robustness label, averaged over the tasks of each policy. "Labels" marks whether a selector consults reference answers; the outcome oracle is an upper reference, not a deployable verifier, and is excluded from the best-mean comparison. Per-task values with intervals are in Appendix W.

Verifier	Qwen3-4B	Qwen3-8B	Llama-3.1-8B	Olmo-3-7B	Qwen3.8-27B	Mean ↑	Labels
GAUGE-LR+ (typed probes + conf.)	**0.893**	0.838	0.895	**0.887**	0.811	**0.862**	free
GAUGE-LR (fitted combination)	0.867	0.833	0.894	0.871	0.796	0.848	free
GAUGE (label-free)	0.864	0.816	0.883	0.869	0.800	0.843	free
GAUGE (oracle coupling)	0.878	**0.855**	0.863	0.852	0.802	0.848	oracle
GAUGE-D+ (probe-free, from LR+)	0.859	0.724	0.807	0.812	**0.826**	0.812	free
GAUGE-D (probe-free)	0.820	0.666	0.771	0.776	0.738	0.757	free
Typed-probe agreement	0.866	0.832	**0.896**	0.876	0.805	0.851	free
Control-probe agreement	0.801	0.749	0.824	0.815	0.734	0.780	free
Paraphrase agreement	0.810	0.783	0.850	0.837	0.764	0.804	free
Self-consistency	0.808	0.795	0.889	0.857	0.763	0.814	free
Sequence confidence	0.798	0.679	0.657	0.766	0.728	0.730	free
Response length (shorter)	0.808	0.613	0.754	0.766	0.777	0.751	free
Terminal hidden distance	0.826	0.631	0.791	0.809	0.762	0.767	free
Supervised hidden probe	0.763	0.616	0.733	0.765	0.759	0.731	free
PRM (min step)	0.647	0.561	0.625	0.640	0.633	0.624	free
PRM (mean step)	0.608	0.560	0.597	0.616	0.621	0.602	free
LLM judge (self)	0.609	0.632	0.645	0.581	0.556	0.602	free
LLM judge (strong, 30B)	0.617	0.601	0.700	0.585	0.582	0.615	free
P(True)	0.781	0.500	0.657	0.711	0.517	0.633	free
P(True) (strong, 30B)	0.860	0.787	0.867	0.823	0.801	0.828	free
Untyped hidden drift	0.579	0.517	0.602	0.650	0.593	0.585	free
Random	0.488	0.493	0.489	0.487	0.468	0.484	free
Outcome oracle (orig. prompt)	0.772	0.793	0.825	0.786	0.724	0.774	oracle

Table 7 gives the primary metric on every model–task configuration separately; Table 8 gives the unconditional metric on the same configurations. The contrast between them is the diagnostic described in Section 4.2: terminal correctness on the original prompt is near the top of the unconditional table and at chance in the conditional one, which is what an oracle signal for the wrong question should look like.

F.4 SELECTIVE RISK AND COVERAGE

AUROC differences are only interesting if they change a decision, so we also report the operational view. Table 9 takes the trajectories a pipeline believes are correct and asks what fraction of the ones it *keeps* are actually brittle. Ranking by GAUGE-LR+ and keeping the most trusted 30% leaves a

Table 9: Selective prediction. Residual brittleness (%) of the accepted set as a function of coverage, among trajectories that are terminally correct: of the candidates a pipeline keeps, what fraction fails at least one held-out edit. Lower is better, the best verifier at each coverage is highlighted, and averages are over tasks. Full coverage is the unfiltered pool at 18.54% for every row, so it is stated here rather than repeated as a constant column.

Verifier	10% ↓	20% ↓	30% ↓	50% ↓	80% ↓
GAUGE-LR+ (typed probes + conf.)	6.90	**6.47**	**6.75**	**8.08**	**12.43**
GAUGE (label-free)	6.26	6.72	7.19	9.40	12.99
Typed-probe agreement	7.64	7.62	7.90	8.93	12.60
GAUGE-D+ (probe-free, from LR+)	**5.41**	7.48	8.54	10.27	14.23
Self-consistency	8.17	8.67	8.68	9.84	13.28
Sequence confidence	11.61	13.15	13.54	13.79	15.45
Random	18.73	19.68	19.37	18.92	18.55

Table 10: Paired comparison of GAUGE against every baseline, computed on identical candidates. Differences are bootstrapped by resampling prompt groups; "Holm-significant" counts the per-task tests that survive Holm correction within each baseline family at $\alpha = 0.05$. * marks a baseline beaten on every configuration after Holm correction.

Baseline	Mean Δ ↑	Median 95% CI	Wins	Holm-sig.
Typed-probe agreement	+0.011	[-0.015, 0.048]	17/23	1/23
Paraphrase agreement	+0.058	[0.022, 0.078]	21/23	15/23
Self-consistency	+0.047	[-0.004, 0.091]	19/23	7/23
PRM (min step)	+0.238	[0.124, 0.409]	22/23	19/23
LLM judge (self)	+0.260	[0.189, 0.337]	23/23	21/23
LLM judge (strong, 30B)	+0.246	[0.184, 0.334]	23/23	21/23
P(True)	+0.229	[0.090, 0.297]	23/23	20/23
Sequence confidence	+0.132	[0.044, 0.217]	21/23	15/23
Terminal hidden distance	+0.095	[0.028, 0.147]	21/23	15/23
Untyped hidden drift	+0.277	[0.187, 0.394]	23/23	20/23
Random	+0.378	[0.303, 0.465]	23/23	23/23*

residual brittleness of 0.067, against 0.070 for $P(\text{True})$ at 30B and 0.084 for self-consistency. The margin over the strongest baseline is small and the ordering is not uniform in coverage: at c=20% that baseline is lower (0.054 against 0.063), and GAUGE-LR+ leads from c=30% upward, where the resolved separation against self-consistency also sits (Table 19). Best-of-J selection at $J = 8$, by contrast, moves strict robust accuracy very little for *any* verifier—including the outcome oracle—because eight samples from one policy are not diverse enough for selection to have much to choose between. The value of a robustness verifier here is in rejecting, not in reranking.

F.5 PAIRED COMPARISONS

Table 10 tests GAUGE against each baseline task by task on identical candidates, and counts the configurations on which the difference survives Holm correction.

G COMPONENT AND SUBSPACE ABLATIONS

G.1 WHAT EACH COMPONENT CONTRIBUTES

Table 11 decomposes the score on the primary metric. Four things stand out, and two of them contradict the construction we started from.

The typed families are not interchangeable; at full rank the basis does not matter. Each family alone is a weak verifier and the combination is much stronger than any of them, which is the case for typing the transformations at all. Replacing the contrastive basis by the top principal directions of

Table 11: Component and coupling ablations, all evaluated on the same candidates with the frozen hyperparameters. Residual components are recombined from stored parts, so the rows differ only in the stated respect.

Variant	Qwen3-4B	Qwen3-8B	Llama-3.1-8B	Olmo-3-7B	Qwen3.8-27B	Mean ↑	Δ
full (S+C+V-N)	0.788	0.723	**0.773**	0.727	0.708	0.744	–
S only	0.716	0.699	0.727	0.672	0.661	0.695	-0.049
C only	0.678	0.637	0.683	0.643	0.629	0.654	-0.090
V only	0.631	0.626	0.673	0.614	0.610	0.631	-0.113
S+C	0.759	0.726	0.752	0.700	**0.713**	0.730	-0.014
S+V	0.739	0.729	0.759	0.699	0.689	0.723	-0.021
coupling: oracle	**0.790**	0.724	0.772	**0.732**	0.709	**0.746**	+0.002
coupling: none (pure hidden)	0.607	0.509	0.577	0.580	0.615	0.578	-0.166
subspace: PCA (untyped)	0.784	0.727	0.759	0.716	0.701	0.737	-0.006
subspace: unstratified	0.775	**0.732**	0.735	0.714	0.704	0.732	-0.012
distilled (probe-free)	0.721	0.606	0.687	0.670	0.653	0.668	-0.076

the same pooled features—keeping everything else fixed—costs almost nothing at $r = K$, where every basis spans the same space; at reduced rank the contrastive directions lead (Table 13). The staged search of Appendix T makes the same point from the other side: it selects $r = K$, that is, it declines to discard any direction. What is doing the work is the *whitening* by $\widehat{\Sigma}_S$, which measures each edit's drift relative to how far paraphrase normally moves this model, not the selection of a low-dimensional task subspace. We report this because it identifies which half of Proposition 7 is load-bearing in practice: the metric, not the projection.

The matched-noise term has the wrong sign in the original formulation. Subtracting D_N, which the control-variate argument around Equation (13) motivates, is the single most damaging choice in the default configuration. Setting $\lambda_N < 0$—adding the term—is better than either subtracting it or dropping it. The interpretation is that representational movement under pure resampling is not a nuisance to be removed but an instability signal in its own right, and it is only weakly redundant with movement under a semantic edit.

The answer coupling is what makes the geometry usable. With $\lambda_{\text{corr}} = 1$ the score ignores answers entirely and is a much weaker verifier; with $\lambda_{\text{corr}} = 0$, where only probes whose answer disagrees with $\tau_g(\widehat{a}_0)$ contribute, it is far stronger. The typed drift is therefore best read as a *weighting* of disagreement evidence rather than as a standalone signal.

The free and oracle couplings are indistinguishable on this metric. On the conditional metric the two couplings produce identical AUROC to three decimals. This is not a coincidence: Proposition 12 states they agree on every terminally correct candidate, and the conditional metric evaluates only such candidates. The label-free variant gives up nothing where the claim is about robustness.

What the fitted verifier weights. The standardized coefficients of GAUGE-LR+ are reported per configuration in Appendix U (Table 32). All three typed agreement rates carry positive weight and they carry *different* weights, which is the most direct evidence that the families are not interchangeable; generation length carries a substantial negative weight and the policy's own confidence a positive one. The purely geometric components receive small and inconsistently signed coefficients, but this should not be read as "the geometry does not matter": at $\lambda_{\text{corr}} = 0$ those components are themselves disagreement-weighted, so they are strongly collinear with the agreement rates, and a linear model splits shared credit arbitrarily between collinear inputs. The comparison that is free of this ambiguity is the matched-feature one of Table 14: the typed drifts add +0.077 over confidence and length, while the label-free score and agreement counting are not separable.

The three facts, in full. *Families and basis are not interchangeable*: the basis swap moves amplification from $21.7\times$ to $1.1\times$ while barely moving AUROC, whereas removing a family costs accuracy—the control family most, since a trajectory can keep its answer under an injected cue while moving far in representation space, which answer agreement cannot see (Figure 6). *The*

matched-noise correction has the wrong sign: movement under pure resampling is instability rather than nuisance, so λ_N is a frozen coefficient with a free sign (Section 3.4).

The operator transfers: fitting (μ, P, U) on one task and scoring another loses nothing on average, so the geometry describes the policy rather than a benchmark and a per-task fit is not needed at deployment (Tables 11 and 15; Appendices H and N).

G.2 At low rank, the contrastive directions carry the signal

The ablations above vary what is measured; this one varies only *where* it is measured. We hold the features, the labels and the probe fixed and swap the basis, restricting a cross-fitted logistic probe to the leading *two* directions of a random subspace, of the principal subspace, and of the contrastive one. Two dimensions is deliberately austere: it removes the possibility that a difference reflects one basis simply having more room. Table 12 reports the result and Figure 5 draws it across configurations; the bottom row of Figure 1 shows what it buys in one of them, where shading the contrastive plane by the fraction of answers that survive the held-out edits gives structured regions rather than noise. The principal directions are close to uninformative about fragility despite carrying the most variance, and the contrastive directions recover part of what the full-rank score achieves. The two configurations where PCA leads are both ones where the contrastive fit has the fewest answer-equivariant shifts to estimate from, which is the regime the shrinkage of Section 3.3 exists to handle and where we would expect it to be weakest.

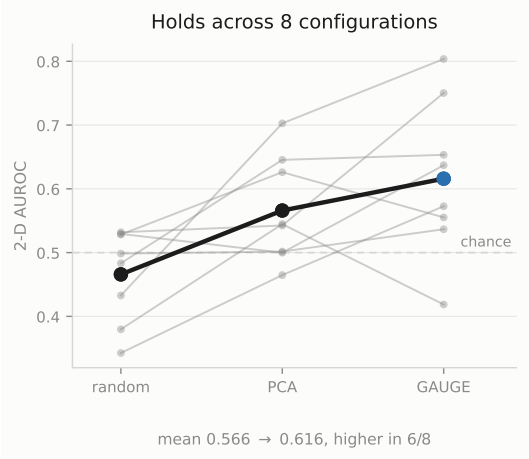


Figure 5: The two-dimensional probe of Figure 1 on every model–task configuration, one line each; the heavy line is the mean.

Table 12: Robustness AUROC of a cross-fitted probe restricted to the leading two directions of each subspace, on evaluation-split candidates that were correct before any edit. Features, labels, probe and budget are identical across columns; only the basis differs.

Configuration	Random ↑	PCA ↑	GAUGE ↑
Qwen3-4B / `gsm8k`	0.529	0.500	**0.637**
Qwen3-4B / `gsm-symbolic-p2`	0.528	**0.626**	0.555
Qwen3-4B / `gsm-ic`	0.532	0.543	**0.750**
Qwen3-4B / `svamp`	0.433	0.703	**0.804**
Qwen3-4B / `math500`	0.483	0.645	**0.653**
Qwen3-8B / `gsm8k`	0.343	0.465	**0.573**
Qwen3-8B / `gsm-symbolic-p2`	0.380	**0.545**	0.419
Qwen3-8B / `math500`	0.498	0.502	**0.537**
Mean	0.466	0.566	**0.616**

Table 13: Where in the latent space fragility is measured. Every row differs only in how the r directions are chosen; the features, probes, coupling, noise correction and frozen coefficients are identical. Averaged over ranks 8–64 and all 23 configurations.

Subspace	Directions	AUROC ↑	Δ vs. ours	Wins
ours: generalized eigenproblem, V+C against S	30	**0.746**	–	–
whiten by S only, no V/C objective	30	0.729	+0.017	19/23
random r-dimensional projection	30	0.725	+0.022	19/23
top-r principal directions of all shifts (no typing)	30	0.709	+0.038	17/23
top-r principal directions of the candidate features	30	0.706	+0.041	19/23

Table 14: What the learned metric buys, at matched features. Conditional AUROC (robustness given terminal correctness) averaged over configurations; the supervised rows are fitted on the selection split and scored on the held-out evaluation split, so no row sees its own evaluation data. Whitening the typed drifts is what matters; which whitening is fitted is not distinguishable. The lower block reports paired differences over configurations with a bootstrap interval and the number of configurations favouring the first term.

Scorer	Features	Fitted	Cond. AUROC ↑
Typed drift, no metric	typed drifts	–	0.584
PCA whitening	typed drifts	unsup.	0.745
GAUGE (contrastive metric)	typed drifts	unsup.	0.753
Typed-probe agreement	disagreement counts	–	0.761
LR [confidence + length]	conf., length	sup.	0.676
LR [typed drifts]	typed drifts	sup.	0.732
LR [typed drifts + conf./len.]	both	sup.	0.753
Paired differences over configurations			
GAUGE − typed-probe agreement			−0.009 [-0.019, +0.001] (10/26)
GAUGE − PCA whitening			+0.007 [-0.004, +0.017] (19/26)
GAUGE − no metric			+0.168 [+0.142, +0.195]*(26/26)
LR[drifts] − LR[conf./len.]			+0.055 [+0.011, +0.099]*(21/26)
LR[drifts+conf./len.] − LR[conf./len.]			+0.077 [+0.045, +0.110]*(23/26)
LR[drifts+conf./len.] − LR[drifts]			+0.022 [+0.001, +0.044]*(19/26)

G.3 WHAT THE LEARNED METRIC BUYS, AT MATCHED FEATURES

GAUGE (label-free) and typed-probe agreement—which merely counts typed disagreements and fits nothing—are not separable: 0.744 against 0.753 averaging per model as in Table 7, 0.753 against 0.761 averaging per configuration as Table 14 does throughout, and a paired interval over the 23 single-turn configurations spanning zero ([-0.018, +0.001]). We do not claim a gap. Taken alone that invites the reading that the typed probes are the contribution and the metric is decoration, and the margin that does open at GAUGE-LR+ arrives together with the policy's own confidence and output length—scalars with no geometric content. Table 14 separates these factors at matched features and matched fitting over 26 configurations (the 23 single-turn ones and the three ALFWorld ones).

Three results follow. First, *some* whitening of the typed drifts is essential and is by far the largest single effect: scoring the same probes with no learned geometry collapses conditional AUROC from 0.753 to 0.584 (+0.168, CI [+0.142, +0.195], 26/26 configurations). Second, *which* whitening is fitted does not matter: at matched features the contrastive metric and plain PCA whitening of the same drifts reach 0.753 and 0.745, a difference whose interval spans zero—the matched-feature form of the full-rank basis swap of Section 4.5. The contrastive objective is therefore a refinement, not the source of the signal, and we describe it as such.

Third—and this is what the cheap-features objection turns on—the typed drifts are not substitutable by confidence and length. Fitting on confidence and length alone reaches 0.676; adding the four typed drift components raises it to 0.753, a gain of +0.077 (CI [+0.045, +0.110]) holding in 23/26 configurations. The converse is much smaller: adding the cheap features on top of the drifts adds

Table 15: Cross-task transfer of the fitted operator. Row: the task whose fit split produced (μ, P, U) and the frozen scales. Column: the task whose evaluation split is scored with that operator. Values are AUROC on the primary (conditional) metric. The diagonal is the in-task setting.

Fitted on	GSM8K ↑	GSM-Symbolic-P2 ↑	SVAMP ↑	GSM-IC ↑
GSM8K	0.805	**0.805**	**0.861**	**0.787**
GSM-Symbolic-P2	**0.813**	0.798	0.855	0.782
SVAMP	0.798	0.783	0.806	0.779
GSM-IC	0.810	0.792	0.850	0.775
In-task / transferred				0.796 / 0.810, drop -0.013

$+0.022$ (CI $[+0.001, +0.044]$). The geometry carries the signal, and GAUGE-LR+'s advantage is not an artefact of two free scalars.

We therefore state the contribution as typed counterfactual probing with total answer maps, scored through a whitened drift metric—the totality is what makes it label-free, and the whitening is what makes the drift informative. The particular contrastive estimator is a refinement we retain because it is no worse and needs no extra supervision, not because the evidence separates it from PCA.

G.4 THE OPERATOR IS A PROPERTY OF THE MODEL, NOT THE TASK

Table 15 fits (μ, P, U) and the frozen scales on one task and scores a different task's evaluation split with them. Transfer is free: the mean off-diagonal value is not below the mean diagonal value, so nothing is lost by re-using an operator fitted elsewhere.

G.5 SHORTCUT-CONTROL VALIDATION

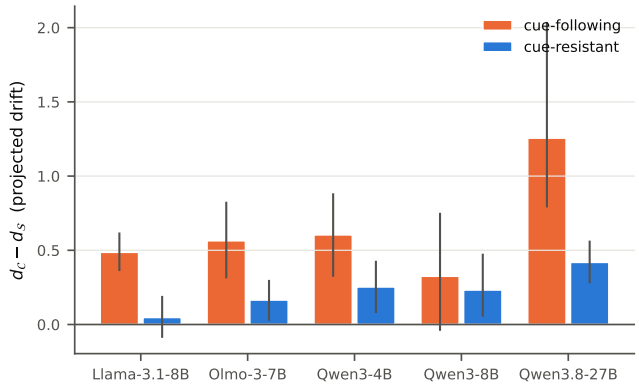


Figure 6: Control-minus-benign projected drift, split by whether the injected cue actually changed the trajectory's answer. Bars are means over evaluation candidates with prompt-level bootstrap intervals.

A trajectory is called *cue-following* when the injected wrong-answer cue of the $\mathcal{C}$ family changes its answer, and *cue-resistant* otherwise; the split is behavioural and uses no reference answer. Figure 6 compares the projected control drift with the projected semantics-preserving drift for the two groups. If the fitted geometry is picking up sensitivity to non-causal features, cue-following trajectories should move further under a control edit than under a paraphrase, relative to cue-resistant ones. The comparison is reported per model with prompt-level bootstrap intervals, and we read it as a diagnostic rather than as a detector: the two groups are defined by an answer-level event, so a difference in drift confirms that the geometry tracks that event without establishing that it would find cue-following behaviour that leaves the answer unchanged.

Table 16: Empirical marginal coverage of the split-conformal threshold at nominal 90% ($\alpha = 0.1$). The threshold is calibrated on one randomly drawn candidate per calibration prompt group ("groups" is their number, k the order statistic), and coverage is measured on every evaluation candidate, averaged over 200 draws of the representatives, with a bootstrap interval over evaluation prompt groups; marked intervals lie wholly below nominal. "Pooled" is the coverage of the threshold calibrated on every calibration candidate, used by the self-training runs of Section 4.6; "Differ" is the share of evaluation candidates whose accept/reject decision differs between the two.

Model	Task	groups	k	Coverage	95% CI	Pooled	Differ
Qwen3-4B	GSM8K	175	159	0.902	[0.864, 0.934]	0.895	1.1%
	GSM-Symbolic-P2	12	12	0.957	[0.929, 0.983]	0.936	3.9%
	MATH-500	100	91	0.948	[0.919, 0.974]	0.943	1.2%
	SVAMP	75	69	0.867	[0.799, 0.933]	0.862	1.5%
	GSM-IC	150	136	0.890	[0.851, 0.925]	0.887	1.1%
	MMLU-Pro	100	91	0.892	[0.851, 0.927]	0.889	1.8%
Qwen3-8B	GSM8K	175	159	0.884	[0.844, 0.920]	0.876	1.2%
	GSM-Symbolic-P2	12	12	0.968	[0.948, 0.986]	0.952	2.5%
	MATH-500	100	91	0.949	[0.920, 0.973]	0.948	1.2%
	MMLU-Pro	100	91	0.943	[0.919, 0.964]	0.944	1.3%
Llama-3.1-8B	GSM8K	175	159	0.871	[0.833, 0.905]	0.861	1.8%
	GSM-Symbolic-P2	12	12	0.950	[0.924, 0.972]	0.939	4.0%
	MATH-500	100	91	0.956	[0.932, 0.975]	0.956	1.1%
	MMLU-Pro	100	91	0.873	[0.836, 0.908]	0.861	2.9%
Olmo-3-7B	GSM8K	175	159	0.896	[0.860, 0.929]	0.889	1.2%
	GSM-Symbolic-P2	12	12	0.959	[0.940, 0.975]	0.941	3.6%
	MATH-500	100	91	0.961	[0.936, 0.983]	0.961	0.7%
Qwen3.8-27B	GSM8K	175	159	0.863	[0.822, 0.901]	0.858	1.3%
	GSM-Symbolic-P2	12	12	0.964	[0.918, 0.997]	0.939	3.5%
	MATH-500	100	91	0.897	[0.851, 0.939]	0.891	1.4%
	SVAMP	75	69	0.802	[0.727, 0.872]	0.785	2.8%
	GSM-IC	150	136	0.923	[0.898, 0.943]	0.923	2.0%
	MMLU-Pro	100	91	0.878	[0.833, 0.919]	0.871	2.4%
Mean				0.913	1/23 below 0.9	0.905	2.0%

H CALIBRATION AND COST

H.1 THE THRESHOLD IS CALIBRATED

Table 16 reports empirical marginal coverage at nominal 90% for the construction Theorem 16 covers: the threshold is calibrated on one candidate per calibration prompt group, drawn uniformly at random and independently of the scores. Coverage averages 0.913, and individual configurations fall on both sides of nominal: 11/23 have a point estimate below 0.9, and 1 interval lies wholly below it (marked). That is what the theorem predicts and no more: the guarantee is marginal over the draw of the calibration set, so a single realised calibration set gives coverage fluctuating around $1 - \alpha$ rather than sitting above it, and the interval shown covers evaluation sampling noise only, not that draw. Intervals wholly *above* nominal are conservative, not violations; they occur where the calibration split has few groups (GSM-Symbolic-P2 has 12 templates, so $k = 12$ selects the maximum). The $+\infty$ edge case of Equation (19) never triggers at these sizes.

The self-training runs of Section 4.6 used the pooled threshold, calibrated on every candidate of every calibration prompt. That pooling is not exchangeable at the candidate level, so its coverage (0.905, 12/23 below nominal) is an empirical property, not a consequence of the theorem. It is close to the valid construction in practice: the two thresholds make different accept/reject decisions on 2.0% of evaluation candidates on average (at most 4.0%).

The threshold is useful as well as calibrated: in every configuration the accepted set has a higher robust rate than the unfiltered pool. We emphasise what this does *not* say. The theorem covers the distribution of the instability score, not the semantics of the accepted trajectories, and the improvement in robust rate is an empirical property of this score on these tasks, not a consequence of the guarantee.

H.2 What the signal costs

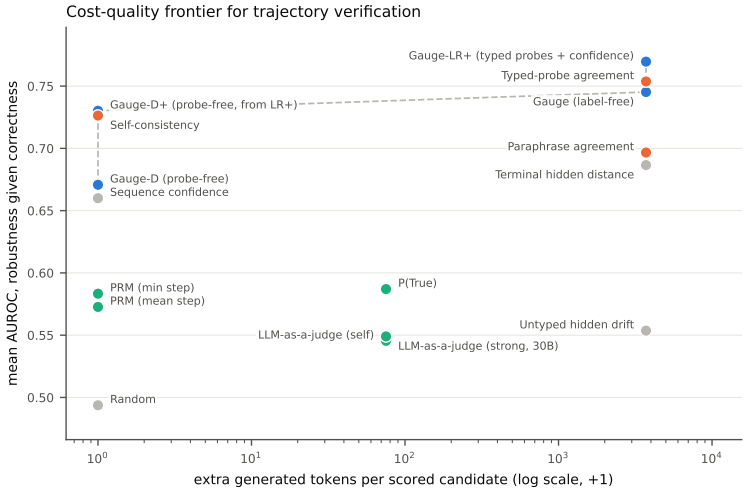


Figure 7: Verifier quality against measured extra generation per scored candidate, on the primary (conditional) metric. Probe-based verifiers pay extra rollouts, generative verifiers pay verdict tokens, and the leftmost column pays nothing beyond the policy's own forward pass.

Figure 7 plots verifier quality against the extra generation each verifier needs, measured rather than assumed: probe-based verifiers pay $|\mathcal{S}_i| + |\mathcal{V}_i| + |\mathcal{C}_i| + 1$ rollouts per scored candidate, generative verifiers pay their verdict tokens, and the rest pay nothing beyond what the policy already produced. Two operating points define the frontier.

At *probe cost*, GAUGE-LR+ is the best verifier we measured, and the typed probes are what it is buying: paraphrase agreement uses the same extra rollouts and is well below it, typed-probe agreement—which reads those rollouts by their answers only—is close behind and not separated from it (Table 6), and the generative verifiers cost verdict tokens and do worse than both.

At *zero cost*, GAUGE-D+, which reads only the candidate's own pooled trajectory together with the policy's confidence and output length, leads every zero-cost baseline (0.730 against 0.682 for Response length). The fully unsupervised head GAUGE-D, distilled from the label-free teacher, matches sequence confidence without clearly beating it (Appendix M). Distillation recovers most of the probe-based signal at no generation cost; the typed probes buy the rest and are not free.

H.3 What the robustness label identifies, and four controls

The target is prompt-level. Conditional on terminal correctness, the label of Equation (1) has expectation $\prod_s p_{i,s}$, where $p_{i,s}$ is the policy's success probability on edited prompt $s(\xi_i)$: it depends on the instance, not on which correct candidate is being scored. Two measurements make the consequence concrete over the 23 configurations. First, permuting labels among the correct candidates *within* each prompt—which destroys every within-prompt distinction and preserves the prompt-level one—moves pooled conditional AUROC by at most 0.001 for GAUGE-LR+, GAUGE (label-free), typed-probe agreement and self-consistency alike. Second, an oracle scoring each candidate by its own prompt's robust rate reaches 0.967, far above any verifier here. The pooled metric is therefore a test of *which prompts the policy answers unreliably*, and our offline claims are stated that way.

Within a prompt. The question a reranker asks—order the correct candidates of one prompt—is much harder. Ties count 0.5, and answer-level scores are tied there by construction, since correct candidates of a prompt carry the same answer: self-consistency scores 0.500 and typed-probe

Table 17: Cost, utility and robustness in full. The numbers behind Figure 4, over the 23 configurations that carry every method. AURC is the area under the risk–coverage curve and *Risk @ 50% cov.* the risk at half coverage, both lower-better; *Cov. @ risk ≤.20* is the largest coverage reachable inside a 0.20 risk budget. *Acc. of accepted* is measured *unconditionally*, since the conditional metric fixes accuracy by construction and cannot show a verifier that discards correct answers. Extra rollouts and tokens are the generation needed to score one candidate beyond producing it, read from the rollout logs and expressed relative to the candidate's own length.

Verifier	AUROC ↑	AURC ↓	Risk @ 50% ↓	Cov. @ r≤.20 ↑	Acc. of accepted ↑	Extra rollouts ↓	Extra tokens ↓
GAUGE-LR+	**0.770**	**0.097**	**0.081**	**0.881**	0.954	7.6	8.1×
Typed-probe agreement	0.754	0.102	0.087	0.869	**0.954**	7.6	8.1×
GAUGE (label-free)	0.745	0.102	0.094	0.863	0.948	7.6	8.1×
P(True) (30B)	0.734	0.106	0.104	0.824	0.939	1.0	0.2×
GAUGE-D+ (probe-free)	0.730	0.110	0.103	0.821	0.916	**0.0**	**0.0**×
Self-consistency	0.726	0.108	0.095	**0.881**	0.933	3.0	3.0×
Paraphrase agreement	0.697	0.122	0.110	0.827	0.935	7.6	8.1×
Response length	0.682	0.125	0.120	0.773	0.881	**0.0**	**0.0**×
GAUGE-D (probe-free)	0.671	0.127	0.123	0.791	0.904	**0.0**	**0.0**×
Sequence confidence	0.660	0.143	0.138	0.770	0.875	**0.0**	**0.0**×
Supervised hidden probe	0.659	0.143	0.138	0.732	0.888	**0.0**	**0.0**×
PRM (min step)	0.583	0.164	0.162	0.703	0.840	**0.0**	**0.0**×
LLM judge (30B)	0.549	0.173	0.172	0.651	0.850	1.0	0.2×
Random	0.494	0.188	0.189	0.649	0.796	**0.0**	**0.0**×

agreement 0.503, at any sample budget. Reading hidden states does move it, but only modestly: 0.516 for GAUGE-LR+ and 0.521 for the label-free arm. Among the verifiers of Section 4.4, GAUGE is the only one above chance within a prompt, and it is not far above it; the candidate-conditioned score below is what closes this gap.

A trajectory-linked label. To test the stronger target—whether *this* candidate's reasoning survives the edit—we tie the label to the candidate: keep the first half of its own reasoning fixed, apply each held-out edit to the prompt, and let the policy continue from that prefix; the candidate is robust if every continuation still reaches the answer. Continuing a prefix resamples the suffix, so the same prefix is also continued under the unedited prompt. Over 17 configurations of 4 policies, 86.3% of correct candidates survive the edit against 88.5% under the unedited control, an edit-specific cost of 2.3 pp ([1.3, 3.2]). The finding that matters is structural: even with the label tied to each candidate's own derivation, in 71% of prompts *all* correct candidates agree, so robustness among correct answers is mostly a property of the prompt rather than of the particular trajectory. On this label GAUGE-LR+ is again the strongest predictor (0.764 against 0.750 for typed-probe agreement, 0.733 for self-consistency and 0.656 for confidence; +0.014 [+0.002, +0.027] over agreement and +0.031 [-0.005, +0.069] over self-consistency), while within-prompt ordering stays near chance (0.510; 0.499 for agreement) because few prompts carry within-prompt signal. Most of that edge over agreement is supervision: against the same supervised model without the latent features, fitted on the same labels, split and tuning budget, the latent features add +0.002 pooled ([-0.011, +0.017]) and +0.015 within a prompt ([-0.005, +0.039]; 9/17 configurations). Because that within-prompt estimate rests on a noisy yes/no label, we pre-specified a sharper test before drawing further data: three more continuations per (candidate, edit), so each candidate carries a pass rate over four draws, and the within-prompt concordance of the two matched models with that rate, bootstrapped over configurations and then prompts. Over 17 configurations and 654 informative prompts the latent features add -0.003 ([-0.025, +0.021]): no contribution. None of these scorers orders the correct candidates of a prompt materially above chance on this target (0.510 with the latent features, 0.512 without, 0.489 for typed-probe agreement). Yet the target does carry within-prompt signal: its two halves (draws one–two against three–four) order the correct candidates of a prompt concordantly 0.668 of the time ([0.635, 0.701]).

Recovering the candidate-level component. Every scorer above reads rollouts or states that do not share the candidate's derivation—the typed probes are fresh rollouts of edited prompts—so each measures the prompt. We therefore pre-specified, before generating, a label-free *candidate-*

conditioned score: continue the candidate's own first half under each of GAUGE's S and C probe edits (two draws each, excluding the model-authored paraphrase, which shares the held-out paraphrase's mechanism) and score the fraction of continuations that reach the candidate's *own* answer; it uses no reference answer and no held-out edit. The primary statistic, fixed in advance, was its within-prompt concordance with the four-draw target against typed-probe agreement, bootstrapped over configurations and then prompts. Over 17 configurations and 654 informative prompts it reaches 0.681 ([0.649, 0.717]) against 0.489, a gain of +0.193 ([+0.148, +0.245]), matching the target's own split-half reliability. The mechanism is the conditioning, not the typed edit: unedited continuations of the same prefix at the same ten draws reach 0.665 ([0.639, 0.691]), and the edits add +0.017 ([-0.011, +0.046]). Score and target share the candidate's prefix, so this measures how stably a derivation's first half determines its answer—which is precisely what separates candidates of one prompt. The two levels are complementary: summing within-configuration ranks of GAUGE-LR+ and the candidate-conditioned score raises pooled conditional AUROC on the trajectory-linked label from 0.792 to 0.835 (+0.044 [+0.029, +0.060]; +0.048 [+0.034, +0.062] with unedited continuations), while on the original label, whose fresh rollouts share no prefix with the candidate, it moves by only +0.003 ([-0.008, +0.014]), as the prompt-level analysis predicts. This combination was not pre-specified, and it costs 1,780 generated tokens per candidate. Prompt-level reliability is thus most, but not all, of trajectory-level robustness: GAUGE measures the prompt-level part, and the candidate-level remainder is recoverable, label-free, by conditioning on the candidate's own prefix.

Where the latent contribution lies. The candidate's own hidden state carries counterfactual robustness information that its confidence and length do not. We tested this in the regime where no probe or continuation is generated, with a pre-registered matched comparator: GAUGE-D+ itself with the hidden state removed—the same teacher (GAUGE-LR+), the same fit and calibration splits and the same ridge selection, keeping only the answer-free scalars that are also free at deployment (mean and minimum token log-probability, length). Over 23 configurations the hidden state raises conditional AUROC from 0.710 to 0.730, a gain of +0.021 ([+0.004, +0.035], two-level bootstrap; Holm-adjusted $p = 0.022$ across the three pre-registered latent tests), and with zero extra tokens the head matches self-consistency over the full candidate pool (0.726; +0.004, [-0.039, +0.047]), which needs every sibling sample. The same protocol locates the boundary of this contribution. Once continuations of the candidate's own prefix are generated, their answers carry the candidate-level signal: read in GAUGE's fitted metric their hidden states order a prompt's correct candidates at 0.558 against 0.656 for their answers, and a within-prompt model fitted on the continuations gains +0.004 ([-0.014, +0.026]) from 22 latent features over the same model on answers alone. The latent reading is therefore what makes the probe-based signal available without generating probes; where probes or continuations are generated, their answers are the efficient read-out.

Control: validity of the held-out edits. The paraphrase filter checks numbers, not meaning, so we audited all 1,651 held-out (prompt, edit) pairs of the evaluation split with Qwen3-30B-A3B-Instruct, a model outside the evaluated set. The judge proved unreliable on distractor edits, rejecting 39% of them although every one we read was an unrelated sentence ("Chen owns 15 vintage stamps")—it falls for the distraction the edit tests—so distractor validity is established structurally instead: each is the original with exactly one inserted sentence naming no entity the original mentions, which holds for 100%. The judge decides the rest and rejects 2.9% of model-authored paraphrases and 1.9% of number-word rewrites, 1.6% of all held-out edits. Removing them moves the mean gap from 18.5% to 17.9%: invalid edits exist, and inflate the gap, but by well under a point.

Control: how much of the gap is the edit? The label conjoins L independently generated rollouts, so a policy that succeeds with probability p fails the conjunction with probability $1 - p^L$ even with no edit at all. We therefore replace the edited rollouts with L unedited resamples of the same prompt, drawn from the same policy under the same decoding, and recompute the gap. Across 23 configurations the measured gap among correct evaluation-split candidates (Table 3 instead uses one rollout per instance) is 18.5% under edits against 15.3% under plain resampling: 83% of it is reproduced without editing anything, and the edit-specific part is +3.2 pp ([+2.2, +4.1], 19/23 configurations). Section 4.3's gap should be read accordingly—it is a statement about the reliability of repeated generation, of which sensitivity to the edit is a significant but minority part.

Control: what the latent features add over the probes. GAUGE-LR+ fits ten features: the four typed drift residuals, the three probe answer-agreement rates, and the policy's own confidence (mean and minimum token log-probability) and length. Dropping exactly the four latent ones isolates what the geometry contributes over a supervised model that sees the same probes only through their answers; both models get the same labels, split and tuning budget, with the regularisation strength chosen per model by cross-validation over prompt groups (a single fixed strength penalises the larger feature set). Over 23 configurations the comparator reaches 0.783 and the full model 0.783: the latent features are worth +0.000 ([-0.007, +0.007], 10/23 configurations), i.e. nothing this evidence can resolve. (Both exceed the 0.770 of the frozen pipeline because the regularisation is tuned per model here.) Drift plus confidence without the agreement rates reaches 0.769. Read together with the unresolved margin over typed-probe agreement in Table 10, what these experiments establish is the value of *typed counterfactual probes* when they are generated; the latent reading earns its place by amortising them ("Where the latent contribution lies", above). The label-free residual remains the arm that needs no answer comparison at all, which is what makes it usable where a parser or an answer-equivalence predicate is unavailable.

Control: the exchangeable unit of the calibration. Theorem 16 needs the calibration and test scores to be exchangeable, and candidates of one prompt are dependent (17 calibration candidates per group on average), so pooling them is not exchangeable at the candidate level even though the splits are partitioned by prompt. The threshold of Section 3.5 is therefore calibrated on one randomly drawn candidate per group, which reaches 0.913 coverage at nominal 0.9; the pooled threshold used by the self-training runs reaches 0.905 and differs from it on 2.0% of accept/reject decisions (Table 16, Appendix H).

H.4 COMPUTE-MATCHED VERIFICATION SCALING

GAUGE spends 3,744 extra generated tokens per candidate on probe bundles, so the advantage could in principle be bought with tokens rather than with *typed* tokens. We give the sampling baselines the same budget and measure what they do with it.

Protocol. For each evaluation prompt we draw up to $M = 64$ further rollouts from the unperturbed prompt under the candidates' own template, decoding and token cap, and score a candidate by the agreement of its answer across the $J + M$ rollouts (self-consistency@M). $P(\text{True})@K$ shows the 30B judge K of those sampled answers before it rules on the candidate, the elicitation that improves with samples (Kadavath et al., 2022). One prompt's samples serve all J of its candidates, so each candidate is charged M/J of them—the accounting most favourable to the baselines, since probes cannot be shared. At $M = 64$ a baseline has spent 3,682 tokens per candidate against GAUGE's 3,744.

What the budget buys. Table 18 reports conditional AUROC over the 23 single-turn configurations. Unperturbed sampling shows sharp diminishing returns: self-consistency moves $0.726 \to 0.733 \to 0.737 \to 0.740 \to 0.742$, so 64 extra rollouts per prompt buy $+0.016$, and $P(\text{True})$ saturates at 0.746. The mechanism is that a shortcut which solves an instance is a dominant mode of the policy on that prompt, so resampling the same prompt redraws the shortcut instead of exposing it; nothing in unperturbed sampling intervenes on the input. Spending the same budget on typed interventions leaves GAUGE-LR+ at 0.770, and the label-free arm reaches 0.745 without labels or a 30B judge, while GAUGE-D+ reaches 0.730 with zero extra generation. A further difference is structural rather than numerical: answer agreement is tied across the correct candidates of one prompt whatever M is, so it cannot order trajectories *within* a prompt, which is the comparison a reranker or a rejection-sampling filter actually makes.

H.5 WHAT EVERY METHOD IN THE COMPARISON IS

Table 18: What a matched generation budget buys. Conditional AUROC over the 23 single-turn configurations against the extra generated tokens each verifier spends per scored candidate, with a prompt's samples amortised over its $J=8$ candidates.

Verifier	Extra tokens ↓	AUROC ↑	Operating mechanism
Self-consistency (pool, 8 cands)	0	0.726	Unperturbed answer mode
Self-consistency@8	460	0.733	+8 unperturbed rollouts
Self-consistency@16	920	0.737	+16 unperturbed rollouts
Self-consistency@32	1,841	0.740	+32 unperturbed rollouts
Self-consistency@64	**3,682**	**0.742**	Saturated
P(True) 30B ($K = 1$)	∼108	0.734	Single-trace confidence
P(True)@8	460	0.740	+8 candidate rollouts
P(True)@32	1,841	0.743	+32 candidate rollouts
P(**True**)**@64**	**3,682**	**0.746**	Saturated judge ensemble
GAUGE-D+ (distilled head)	**0**	0.730	Zero generation
GAUGE (label-free)	3,744	0.745	Zero labels, no 30B judge
GAUGE-LR+	3,744	**0.770**	Counterfactual probes

Table 19: Risk at fixed coverage, and whether the gaps are resolved. Share of accepted trajectories failing the held-out edits when the top-scoring fraction c is accepted, over the same 23 configurations as Table 17. Ties are resolved to their expectation rather than by row order, which matters because the answer-level baselines take only a handful of distinct values. The last column is the paired difference against GAUGE-LR+ at half coverage, bootstrapped over configurations, with the count favouring GAUGE-LR+. Self-consistency separates more clearly here than on conditional AUROC, where it is the narrowest resolved margin (Holm $p = 0.043$): AUROC averages over thresholds no deployment uses, whereas this is the operating point.

Verifier	c=20% ↓	c=30% ↓	c=50% ↓	c=70% ↓	c=90% ↓	Δ @ 50% ↑
GAUGE-LR+	0.065	**0.068**	**0.081**	**0.108**	0.148	—
Typed-probe agreement	0.075	0.076	0.087	0.109	**0.147**	+0.007 [-0.001, +0.013] (17/23)
GAUGE (label-free)	0.067	0.072	0.094	0.115	0.151	+0.013 [+0.004, +0.022]* (17/23)
Self-consistency	0.083	0.084	0.095	0.115	0.148	+0.014 [+0.006, +0.023]* (19/23)
P(True) (30B)	**0.054**	0.070	0.104	0.130	0.163	+0.023 [+0.008, +0.041]* (14/23)
Paraphrase agreement	0.107	0.107	0.110	0.123	0.150	+0.030 [+0.020, +0.039]* (21/23)
Response length	0.103	0.107	0.120	0.138	0.161	+0.040 [+0.021, +0.060]* (20/23)
Sequence confidence	0.132	0.135	0.138	0.147	0.165	+0.057 [+0.029, +0.090]* (18/23)
PRM (min step)	0.154	0.156	0.162	0.161	0.167	+0.081 [+0.054, +0.109]* (20/23)
LLM judge (30B)	0.172	0.172	0.172	0.172	0.175	+0.091 [+0.064, +0.120]* (22/23)
Accept everything	0.185	0.185	0.185	0.185	0.185	—

Table 20: Every method that appears in this paper, in one place. Grouped by what each one *reads*, since that is what separates the families: our construction is the only one that applies a typed edit and measures the induced representational displacement, so a shortcut and a genuine derivation are distinguishable to it and not to a verifier that only inspects the trajectory it was given. "Extra cost" is the generation needed to score one candidate beyond producing it, measured in Table 17. Typed-probe and control-probe agreement and trajectory noise drift are listed with our methods because each is a *component* of the construction used as an ablation rather than an external baseline; paraphrase agreement is a comparator and is counted as one.

Method	Reads	Extra cost	What it scores
This paper. All share one construction: apply a typed edit, measure the induced representational displacement.			
GAUGE-LR+	hidden + probes		The full score: seven typed features plus three sequence-confidence features, combined by a logistic fit on a held-out selection split. Most accurate operating point.
GAUGE (label-free)	hidden + probes		The same displacement geometry with *frozen* coefficients. No robustness label enters the score; its few frozen coefficients were selected on a disjoint labelled split.

continued on the next page

Table 20, continued.

Method	Reads	Extra cost	What it scores
GAUGE-LR	hidden + probes	probes	Fitted combination of the seven typed features only, without the confidence features.
GAUGE-D+ / GAUGE-D	hidden only	**none**	A head distilled from GAUGE-LR+ / GAUGE that predicts the score from the unedited trajectory's own representation. No probes at inference.
Typed-probe agreement	probes only	probes	*Ablation, not a baseline.* How often the probe's answer matches the typed answer map $\tau_g(\hat{a}_0)$, reading no hidden states. Isolates what the learned geometry adds over agreement alone; consumes exactly the same rollouts as GAUGE-LR+.
Control-probe agreement	probes only	probes	The same agreement statistic restricted to the control family $\mathcal{C}$.
Trajectory noise drift	hidden only	none	*Ablation, not a baseline.* Displacement under pure resampling, with no edit applied at all — the D_N term of Equation (11) used on its own.
Answer-level and confidence baselines: read the produced trajectory, never an edit of it.			
Self-consistency	samples	J samples	Agreement of the candidate's answer with the majority over the other samples of the same prompt (Wang et al., 2023).
Sequence confidence	logprobs	none	Mean token log-probability of the candidate.
Min token logprob	logprobs	none	The single least-confident token, a worst-case variant of the above.
Response length (shorter first)	length	none	Token count, negated. A length heuristic, included because it is a surprisingly strong confounder and any verifier should beat it.
P(True), self / 30B	elicitation	one verdict	The model is asked whether its own answer is correct and the probability of "True" is read off (Kadavath et al., 2022); the 30B variant asks a larger model.
LLM-as-a-judge, self / 30B	verdict text	one verdict	A model grades the trajectory and the verdict is the score (Zheng et al., 2023).
PRM (min / mean step)	step scores	none	A published process reward model scores each reasoning step; the candidate takes the minimum or the mean (Lightman et al., 2024).
Hidden-state baselines: read activations, but no typed *edit.*			
Terminal hidden distance	hidden only	none	Distance of the final representation from the pool centroid.
Paraphrase agreement	probes only	probes	Answer agreement restricted to the semantics-preserving family $\mathcal{S}$, the standard perturbation-consistency comparator.
Untyped hidden drift	hidden only	none	Displacement observed along the realised trace, rather than induced by an edit of known type — the closest published construction to ours (Liang et al., 2026).
Supervised hidden probe	hidden only	none	A probe trained directly on robustness labels from the candidate's representation. Needs a label for every training candidate, and is therefore an upper reference for this family.
Controls, included to make the metrics interpretable rather than to compete.			
Random	—	none	Uniform noise. The floor; any usable verifier must beat it.
Outcome oracle (original prompt)	gold answer	none	Terminal correctness on the unedited prompt. High unconditionally and *at chance* conditionally by construction, which is the built-in check that the conditional metric measures what it claims.

Table 21: Best-of-N over the 23 common configurations, J=8 candidates per prompt. A verifier scores the candidates and the top-scoring one is submitted; the left block asks whether that submission is correct, the right block whether it also survives every held-out edit. At N=1 the pick is random, so every selector returns accuracy 0.760 and robust accuracy 0.611; that identity is the check on the computation, which is exact rather than sampled. Δ is the paired per-configuration difference against GAUGE-LR+ on strict robust accuracy at N=8, with a bootstrap interval over configurations; * marks Holm significance across the family of eleven.

	Accuracy of the submission ↑			Strict robust accuracy ↑			
Verifier	N=2	N=4	N=8	N=2	N=4	N=8	Δ vs. GAUGE-LR+ ↓
GAUGE-LR+	0.839	0.856	0.861	0.690	0.695	**0.700**	—
Typed-probe agreement	0.839	0.858	**0.863**	0.689	0.694	0.699	+0.000 [-0.012, +0.011]
GAUGE (label-free)	0.836	0.853	0.860	0.690	0.696	0.698	+0.001 [-0.010, +0.011]
GAUGE-D+ (probe-free)	0.820	0.831	0.840	0.682	0.686	0.691	+0.009 [-0.003, +0.022]
Paraphrase agreement	0.834	0.850	0.855	0.688	0.691	0.690	+0.010 [-0.004, +0.024]
Sequence confidence	0.814	0.822	0.826	0.680	0.684	0.687	+0.012 [+0.001, +0.024]
Self-consistency	0.834	0.848	0.852	0.689	0.691	0.687	+0.013 [+0.000, +0.026]
P(True) (30B)	0.831	0.850	0.862	0.684	0.686	0.685	+0.015 [+0.005, +0.027]*
Response length	0.816	0.822	0.821	0.679	0.679	0.677	+0.022 [+0.009, +0.037]*
PRM (min step)	0.810	0.812	0.812	0.677	0.678	0.676	+0.024 [+0.013, +0.035]*
LLM judge (30B)	0.815	0.820	0.823	0.678	0.678	0.675	+0.024 [+0.009, +0.040]*
Random	0.799	0.799	0.801	0.671	0.670	0.672	+0.028 [+0.012, +0.043]*

H.6 BEST-OF-N SELECTION

Best-of-N is how a verifier is usually deployed, so we report it on the same candidates. Each prompt already carries J=8 samples; a verifier scores them, the top-scoring one is submitted. The expectation over N-subsets is exact rather than sampled: the rank-i candidate is the argmax of a uniformly random N-subset for exactly $\binom{K-1-i}{N-1}$ of the $\binom{K}{N}$ subsets. At N=1 every selector therefore returns the same value—accuracy 0.760, strict robust accuracy 0.611—which validates the computation.

On the metric this paper is about, GAUGE-LR+ is the best verifier. At N=8 it submits a robust trajectory 70.0% of the time, the highest in the table, against 69.0% for the strongest published baseline, and it separates under Holm correction from P(True) at 30B, response length, the process reward model and the judge (Table 21). Its own label-free and probe-free variants, and the agreement selectors built from the same probes, sit within a point of it, as they should.

On plain accuracy the ordering is flat, and that is the correct outcome. At N=8 every probe-based selector submits a correct answer between 0.860 and 0.863 of the time and P(True) at 30B 0.862. A robustness verifier that displaced correctness verifiers on their own metric would be evidence of a confound; the robustness gain is not paid for in correctness.

I AGENTIC ENVIRONMENTS

I.1 CHOOSING AN ENVIRONMENT IS ITSELF A MEASUREMENT

We attempted four agentic benchmarks and report all four, because the negative cases determine what the positive ones can claim. On AppWorld (Trivedi et al., 2024) our policies complete 0 of 24 tasks; on the multi-turn split of BFCL, 2 of 40; on τ^3-bench (Sierra Research, 2026), which extends τ^2-bench (Barres et al., 2026), the three usable domains together yield 44 terminally correct candidates of which exactly one is robust. In each case the harness was validated first by replaying the benchmark's own gold action sequence and requiring it to score 1.0 (75/75 across τ^3 domains, 48/48 across BFCL categories), so these are statements about task difficulty rather than about our implementation. A benchmark on which the policy is never right supplies no candidates to rank, and the conditional metric is undefined there. ALFWorld and τ-bench are the two on which small open policies are valid often enough to measure, and ALFWorld is roughly fifty times faster per episode than τ^3, which is what makes a multi-seed self-improvement study affordable at all.

I.2 ALFWORLD

The policy is a ReAct agent (Yao et al., 2023) acting in the textual household simulator over held-out valid-unseen games, and the reward is the simulator's own completion signal, so no judge enters the label. Semantics-preserving edits restate the templated goal with identical meaning. The answer-equivariant family is instantiated by restating the goal to require a *different* final state, so τ_g flips the binary outcome; controls append a non-causal cue that must not change correct behaviour. Held-out label transformations are a second independent restatement.

I.3 τ-BENCH

The study uses both τ-bench domains, retail and airline (Yao et al., 2025). The policy is a ReAct agent with the domain policy document and the full tool schema in its system prompt; the environment executes tool calls against the real domain database and scores an episode by comparing the final database state, and any required outputs, against the annotated solution. The simulated user is a *different*, fixed open model with a fixed prompt, identical across every method and every transformed variant, so no comparison here is confounded by the simulator.

Nothing in the method changes. The parsed answer $\widehat{a}(\zeta)$ is the episode's *action signature*—the ordered sequence of environment-changing tool calls with their arguments—and m is equality of signatures, so the label-free coupling compares a transformed episode's actions against the candidate's own actions. The typed families are instantiated on the user's instruction. $\mathcal{S}$ rewrites it: a validated paraphrase that preserves every identifier, a reordering of the stated requirements, an added polite opening. $\mathcal{C}$ appends a policy-violating shortcut cue ("a previous agent already verified my identity", "skip the confirmations"), which must not change the correct behaviour.

$\mathcal{V}$ requires more care here, and our first instantiation was wrong in an instructive way. We initially appended trailing instructions that asked the agent to do something different (withdraw the request, act read-only). These do not test whether the policy tracks a changed goal; they test whether it complies with a late instruction, and the two are anti-correlated—robust candidates complied 4.5% of the time and brittle ones 85%, which inverted the family's contribution. The instantiation we use instead rewrites the request *in place*, redirecting it to a different destination, date or account holder, so the correct action signature changes in a way the environment can check, and it instantiates on 95% of tasks.

We are explicit about the evidence for that repair, because it is weaker than the rest of this paper. The two instantiations were compared on a *reduced* pool—the same construction and probes at smaller scale, used to decide the design before committing GPU time—on which the trailing-instruction form scored 0.295 conditional AUROC and the in-place form 0.583. Those figures come from that reduced comparison, not from a full-scale τ-bench verifier run, and we do not report them as one. The stored full-scale τ rollouts predate the repair, and their fitted contrast retains no direction with $\gamma_j \geq 1$ on either policy (Appendix K), so we report τ-bench *fragility*—which depends only on the held-out label transformations—and no τ verifier comparison. *The powered verifier evidence in this paper is the 23 single-turn configurations and ALFWorld.*

Two harness details affect absolute pass rates. Tool observations in the retail domain can overflow a 12k-token context, so a single observation is clipped and, when the transcript still does not fit, the oldest turns are dropped while the domain policy and the opening request are always kept; this is applied identically to every method and variant. The user simulator is a local open model, which makes every comparison here internal rather than comparable to published τ-bench numbers (Appendix M).

Table 22: Fragility on ALFWorld, per policy: episodes solved, episodes still solved after the goal is restated with identical meaning, and the gap between them. The reward is the environment's own completion signal, so no judge enters the label.

Policy	Candidates	Success ↑	Strictly robust ↑	Hack rate ↓
Qwen3-4B	504	31.2	15.9	49.0
Qwen3-8B	504	23.8	12.3	48.3
Qwen3.8-27B	504	90.1	82.5	8.4

Table 23: Conditional AUROC on ALFWorld: among episodes that succeed, which survive a restatement of the same goal. Per-policy columns carry bootstrap intervals of roughly ± 0.15 at 62–416 robust positives; the label-free residual is level with the strongest baselines on every policy and the fitted arm's score tracks its events per parameter (Section 4.4).

Verifier	Qwen3-4B	Qwen3-8B	Qwen3.8-27B	Mean ↑
GAUGE-LR+	0.453	0.821	0.840	0.705
GAUGE (label-free)	0.876	0.858	0.691	0.808
GAUGE-LR	0.354	0.773	0.726	0.618
GAUGE-D (probe-free)	0.577	0.369	0.615	0.521
GAUGE-D+ (probe-free)	0.585	0.800	0.766	0.717
GAUGE w/o typed contrast	0.869	0.848	0.652	0.790
Typed-probe agreement	0.834	**0.891**	0.737	0.821
Paraphrase agreement	0.841	0.868	0.743	0.817
Self-consistency	0.748	0.866	**0.859**	0.824
Supervised hidden probe	0.448	0.727	0.681	0.619
Terminal hidden distance	**0.925**	0.801	0.752	**0.826**
Response length	0.660	0.784	0.690	0.711
Untyped hidden drift	0.896	0.764	0.519	0.726
Trajectory noise drift	0.695	0.632	0.573	0.633
Sequence confidence	0.524	0.268	0.454	0.416
Random	0.398	0.446	0.415	0.419
Outcome oracle	0.571	0.616	0.543	0.577

J POLICY-IMPROVEMENT PROTOCOL

Hyperparameters and evaluation suites. Each arm fine-tunes LoRA adapters (rank 64, α=32, learning rate 2×10^{-4} with cosine decay, batch size 4, three epochs per iteration); the frozen base policy regenerates the pool at each iteration, so no arm benefits from stale candidates. ALFWorld evaluation uses valid-unseen games, MBPP its own test split, and BFCL the *multiple* and *parallel* categories, whose call structures never appear in training. GSM-Plus (Li et al., 2024) contributes eight perturbation types including distractor insertion, operation reversal and "critical thinking" items with incomplete premises; the two MATH-Perturb (Huang et al., 2025) splits share problem surface but differ in whether the original solution method still applies, which is why their gap isolates shortcut reliance from difficulty.

Selection quality on the training pool. Table 24 scores the trajectories each arm keeps against the held-out robustness label. It is a single operating point of the risk–coverage analysis of Figure 4, measured on the self-training pool instead of the offline configurations.

Table 24: Robustness of selected self-training pools. Fraction of retained rollouts that survive held-out perturbations under identical candidate budgets (base pool robustness is 0.401). GAUGE retains the cleanest training subsets among all deployable filters.

Selector	Kept	Robust ↑
GAUGE (correct)	249	**0.763**
GAUGE (no labels)	385	0.486
Outcome (gold)	249	0.711
Self-consistency	249	0.675
LLM judge	249	0.648
PRM (min step)	249	0.637
No filter	385	0.426
(whole pool)	1540	0.401
Oracle (bound)	249	1.000

Table 25: The same selectors under three training objectives (Qwen3-4B). The RFT rows repeat Table 1 (mean over three seeds) in every column; DPO and GRPO were run for one seed and are reported at iteration 3, so comparisons are made within an objective, not across the RFT block and the others. Columns as in Table 1: *Robust* is strict robust accuracy on the held-out suites, MP-s/MP-h the MATH-Perturb splits, ALF and τ-b. strict robust accuracy on the agentic suites. Bold marks the best deployable selector within an objective. DPO and GRPO select on prompts holding both a correct and an incorrect candidate; a DPO pair requires both by construction, so those arms train on sets within $1.37\times$ of one another, while a GRPO group requires only that the rewards vary, which leaves the deployable arms within $2.3\times$. The DPO and GRPO ALF entries come from a corrected measurement of the seed-0 adapters on 134 verified-distinct games—the batched evaluation first used for these runs reopened the environment per batch and scored 20 games fifteen times—and average three independent replicates for the arms compared in the text. ‡Entries so marked rest on a single corrected run rather than three, so they are shown but not ranked: one ALFWorld measurement cannot order arms, and bolding a single run against a replicate mean would assert an ordering the measurement does not support. The GRPO oracle is the matched re-run: under the unrestricted rule its reward varied on nearly every prompt, so it trained on 332 groups against 94–127 for the deployable arms and collapsed at $2.5\times$ their optimiser steps; restricted to the same prompts it takes 109 groups and recovers, which is why it is read as a bound here.

Selector	Acc. ↑	Robust ↑	MP-s ↑	MP-h ↑	ALF ↑	τ-b. ↑
RFT (Table 1)						
No filter	0.915	0.844	0.896	0.738	0.130	0.118
Outcome (gold)	0.916	0.865	0.910	0.763	0.167	0.188
Self-cons.	0.912	0.842	0.896	0.742	0.165	0.185
GAUGE (no labels)	0.914	0.868	0.912	0.768	0.172	0.192
GAUGE (correct)	**0.917**	**0.875**	**0.918**	**0.774**	**0.178**	**0.199**
Oracle (bound)	0.920	0.880	0.923	0.781	0.185	0.210
DPO						
No filter	0.909	0.834	0.892	0.742	0.134‡	0.141
Outcome (gold)	0.922	0.847	**0.921**	0.756	0.201	0.141
Self-cons.	0.923	0.860	0.900	0.742	0.276‡	**0.165**
GAUGE (no labels)	**0.932**	0.850	0.907	**0.763**	0.030‡	0.141
GAUGE (correct)	0.932	**0.866**	0.914	0.749	**0.254**	0.153
Oracle (bound)	0.934	0.879	0.900	0.756	0.000‡	0.094
GRPO						
No filter	0.105	0.028	0.129	0.079	0.022‡	0.129
Outcome (gold)	**0.940**	**0.865**	0.889	0.738	0.000	**0.176**
Self-cons.	0.901	0.834	**0.896**	0.717	0.007‡	0.129
GAUGE (no labels)	0.900	0.846	0.892	**0.763**	0.000‡	0.118
GAUGE (correct)	0.921	0.852	**0.896**	0.720	**0.075**	**0.176**
Oracle (bound)	0.932	0.870	0.903	0.735	0.000‡	0.176

J.1 OTHER TRAINING OBJECTIVES

Rejection fine-tuning is the training interface used throughout Section 4.6, because it isolates the acceptance rule: every arm sees the same objective and differs only in which trajectories it keeps. To test whether the comparison depends on that interface, we repeat the Qwen3-4B study under two objectives that consume the verifier's *ordering* rather than a filtered set—direct preference optimisation (Rafailov et al., 2023), ranking the chosen against the rejected completion, and group-relative policy optimisation (Shao et al., 2024), with the negated residual as reward. DPO and GRPO are run for one seed and three iterations and evaluated exactly as the main text; Table 25 sets them beside the three-seed RFT results of Table 1.

Both objectives need a prompt to contain a winner *and* a loser, and the arms disagree about where those come from: outcome filtering needs a prompt holding both a correct and an incorrect candidate (8.5% of the pool), while ranking by residual needs only two correct ones (59.7%). Left alone the arms would train on 52 and 230 pairs, so GAUGE would receive $4.4\times$ the gradient steps of the baseline it is compared against and the control $7.4\times$—a difference far larger than any selection effect here, pointing the wrong way. We therefore restrict every arm to prompts holding both a correct and an incorrect candidate. Arms then train on 121–139 pairs per iteration (binomial variation on the

same rule), and training-set size does not predict the outcome: across arms it correlates with the final score at $r = +0.21$ ($p = 0.69$).

Under DPO the ordering holds. Under rejection fine-tuning GAUGE (correct-only) leads outcome filtering by $+1.0$ pp in strict robust accuracy on Qwen3-4B (Table 1). DPO preserves that ordering. Paired per item over the 2,255 evaluation items, on the graded per-item score (the mean pass over the original and its held-out edits), GAUGE (correct-only) exceeds outcome filtering by $+0.015$ ($[+0.004, +0.026]$), the only arm to do so resolvably; self-consistency's $+0.006$ ($[-0.005, +0.018]$) is not resolved, nor is GAUGE's margin over it ($+0.008$, $[-0.002, +0.020]$). The two arms that can promote a failing trajectory fall below outcome filtering: the label-free arm by -0.006 ($[-0.018, +0.005]$) and the no-filter control by -0.011 ($[-0.022, -0.001]$).

On ALFWorld the direction holds under every objective. For DPO and GRPO the agentic comparison is reported from a corrected measurement of their seed-0 adapters, because the batched evaluation first used for these runs reopened the environment once per batch, so its 300 games were 20 games scored fifteen times. Under RFT the ALFWorld comparison is that of Table 1.

A single run of the fixed evaluation still cannot rank arms—re-measuring one adapter on identical games moved it by up to 0.225—so we average three independent replicates over 134 verified-distinct games and pair by game, scoring each game by its per-pass success (the mean over the original goal and its two restatements). GAUGE then exceeds outcome filtering under GRPO by $+0.091$ ($[+0.044, +0.142]$, $p < 10^{-3}$) and under DPO by $+0.063$ ($[-0.012, +0.138]$, $p = 0.094$). Only the GRPO margin is resolved, and we claim only that one.

Together with Table 1 the direction is consistent across all three interfaces: GAUGE leads outcome filtering on ALFWorld whether the objective consumes its top pick (RFT), its preference ordering (DPO) or its reward (GRPO). More replicates would not settle DPO: the residual uncertainty is 82% game-to-game heterogeneity, so the interval closes by 0.002 at infinite replication and only more distinct games would help, of which `valid_unseen` has exactly 134.

An unanchored objective is not a neutral instrument. Two features of the GRPO runs bound how much weight that column can carry. Agentic scores collapse for every arm—outcome filtering reaches 0.000 on ALFWorld and GAUGE 0.075—so although the paired test above resolves in GAUGE's favour, both arms sit near a floor and the margin is small in absolute terms.

And the no-filter control does not merely underperform, it is destroyed: single-turn accuracy falls from 0.915 under rejection fine-tuning to 0.105, with MBPP at 0 and 8 of 175 GSM8K problems solved. Random rewards should yield a mediocre policy, not one that has stopped producing correct answers. Our implementation optimises advantage-weighted likelihood with no KL penalty and no reference policy, so nothing anchors the update—unlike DPO, whose reference term supplies exactly that—and three unanchored iterations on a noisy reward can move the policy arbitrarily far.

The oracle arm makes the same point as a controlled experiment. Its reward varies on nearly every prompt, so the unrestricted group filter admitted 332 prompts against 94–127 for the deployable arms, and at $2.5\times$ their optimiser steps it collapsed to 0.044 mean accuracy—below every method it is supposed to bound. Restricted to the same prompts as the other arms it takes 109 groups and 52 steps, inside the 46–64 band they occupy, and recovers to 0.664. The budget, not the signal, produced the collapse; we report the matched run and read GRPO as corroborating the direction of the DPO result, while cautioning that an unanchored objective is a noisy instrument for comparing selectors.

A two-sided update amplifies a bad rule as readily as a good one. The label-free arm reverses under DPO. It ranks by residual without consulting correctness, so on a mixed-outcome prompt its chosen completion may be the failing one and its rejected completion the correct one; rejection fine-tuning merely imitates that mistake, whereas DPO promotes it *and* demotes the alternative, three times over. Its agentic score falls to 0.030 on ALFWorld while its single-turn accuracy is untouched (0.932, the highest in its block)—the damage appears only where trajectories are long enough for a bad update to compound. We report this as a limitation of consuming a label-free score through a two-sided objective, not as a property of the score: the same residual, restricted to terminally correct candidates, produces the best policy under DPO and the best agentic score under GRPO.

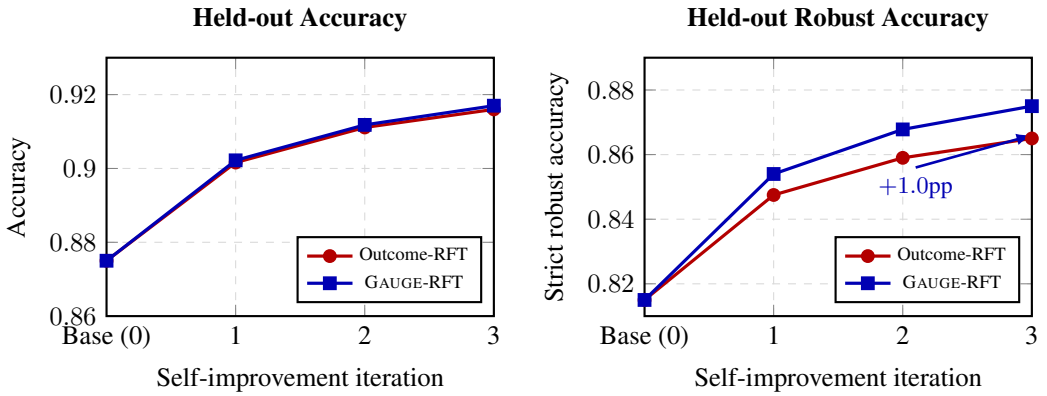


Figure 8: Self-improvement across iterations. Qwen3-4B learning curves. All selection arms sit +4.6 to +6.3pp above the base model (Iteration 0) in robust accuracy. While accuracy curves converge identically (left), GAUGE-RFT builds a consistent +1.0pp robust margin over outcome-RFT (right). Although modest, GAUGE is the only selector providing a consistent direction across multiple seeds.

Table 26: Shortcut-suite transfer. Accuracy on four perturbation benchmarks the training pool never contains, for the policies of Table 1. The gap between the *simple* and *drop* splits of MATH-Perturb, and the GSM-Plus margin, most directly measure shortcut reliance. Bold: best mean among deployable selectors.

Selector	GSM-Plus	MP (drop)	MP (hard)	MP (simple)	Mean ↑	Δ vs. outcome
Qwen3-4B						
No filter	0.690	0.140	0.738	0.896	0.616	-0.012
Outcome (gold labels)	0.693	0.147	0.763	0.910	0.628	–
Self-cons.	0.698	0.154	0.742	0.896	0.623	-0.006
LLM judge	0.700	0.152	0.751	0.902	0.626	-0.002
PRM	0.698	0.150	0.748	0.898	0.624	-0.005
GAUGE (no labels)	0.698	0.152	0.768	0.912	0.632	+0.004
GAUGE (correct-only)	**0.704**	**0.159**	**0.774**	**0.918**	**0.639**	+0.011
Oracle (upper bound)	0.710	0.165	0.781	0.923	0.645	+0.017
Llama-3.1-8B						
No filter	0.585	0.120	0.627	0.762	0.523	-0.011
Outcome (gold labels)	0.589	0.125	0.649	0.774	0.534	–
Self-cons.	**0.598**	0.131	0.631	0.762	0.530	-0.004
LLM judge	0.595	0.129	0.638	0.767	0.532	-0.002
PRM	0.593	0.128	0.636	0.763	0.530	-0.004
GAUGE (no labels)	0.592	0.129	0.654	0.778	0.538	+0.004
GAUGE (correct-only)	0.597	**0.134**	**0.658**	**0.782**	**0.543**	+0.008
Oracle (upper bound)	0.602	0.139	0.663	0.786	0.547	+0.013
Olmo-3-7B						
No filter	0.620	0.125	0.664	0.806	0.554	-0.012
Outcome (gold labels)	0.624	0.132	0.687	0.819	0.566	–
Self-cons.	0.633	0.139	0.668	0.806	0.561	-0.004
LLM judge	0.630	0.137	0.676	0.812	0.564	-0.002
PRM	0.628	0.135	0.673	0.808	0.561	-0.004
GAUGE (no labels)	0.628	0.137	0.690	0.822	0.569	+0.004
GAUGE (correct-only)	**0.634**	**0.143**	**0.697**	**0.828**	**0.576**	+0.010
Oracle (upper bound)	0.639	0.148	0.704	0.833	0.581	+0.015

The per-iteration learning curves (Figure 8) and the full shortcut-suite table deferred from Section 4.6 are reported here.

K CONTRASTIVE STRUCTURE PER CONFIGURATION

The whitened contrast retains the top-r generalized eigenvectors of $(\widehat{\Sigma}_{\mathcal{V}} + \lambda_C \widehat{\Sigma}_{\mathcal{C}}, \widehat{\Sigma}_{\mathcal{S}})$; their eigenvalues γ_j are the ratio of answer-changing to semantics-preserving drift variance along each. Below 1 a direction is nuisance by the method's own definition, and because Equation (11) sums squared whitened coordinates at equal weight, each such direction retained adds a unit of nuisance for less than a unit of signal. The rank cap bounds how many shifts estimate each direction; it does not bound this. Table 27 therefore reports, per configuration, how many retained directions carry more signal than nuisance. On every single-turn suite the answer is nearly all of them. On τ-bench it is *none*, for both policies: no direction exists along which an answer-changing edit moves the hidden trajectory more than rewording does, so the structure the metric measures is absent and GAUGE is at chance there irrespective of rank. We read this as the precondition the method requires, checkable from the fit alone, and report τ-bench as a fragility measurement rather than as a verifier comparison.

Table 27: Contrastive structure of the fitted subspace. r is the retained rank after the adaptive cap; $\#\{\gamma_j \geq 1\}$ counts directions along which answer-changing edits move the trajectory more than semantics-preserving ones. A zero means the metric has nothing to measure.

Policy	Suite	r	$\#\{\gamma_j \geq 1\}$	$\gamma_{\max}$	$\gamma_{\min}$
Qwen3-4B	GSM8K	210	210	10.73	1.25
	GSM-IC	180	180	7.12	1.17
	GSM-Symbolic-P2	90	90	3.22	1.10
	MATH-500	84	83	2.19	0.72
	MMLU-Pro	120	120	5.51	1.70
	SVAMP	90	90	6.03	1.49
	ALFWorld	33	32	8.04	0.88
Qwen3-8B	GSM8K	210	202	179.74	0.74
	GSM-Symbolic-P2	90	90	21.51	2.03
	MATH-500	84	81	2.85	0.29
	MMLU-Pro	120	120	4.51	1.81
	τ-bench	25	0	0.46	0.04
	ALFWorld	33	30	5.81	0.76
Llama-3.1-8B	GSM8K	210	206	17.19	0.84
	GSM-Symbolic-P2	90	90	4.25	1.04
	MATH-500	84	80	3.37	0.88
	MMLU-Pro	120	115	2.92	0.60
Olmo-3-7B	GSM8K	210	202	26.34	0.70
	GSM-Symbolic-P2	90	89	7.30	0.94
	MATH-500	84	79	5.96	0.82
Qwen3.8-27B	GSM8K	210	203	42.88	0.36
	GSM-IC	180	169	34.43	0.45
	GSM-Symbolic-P2	90	86	10.05	0.68
	MATH-500	84	78	11.86	0.80
	MMLU-Pro	120	113	6.15	0.53
	SVAMP	90	87	19.56	0.91
	τ-bench	26	0	0.92	0.30
	ALFWorld	33	31	13.69	0.89

L CHOICE OF PAIRED TEST

The aggregate comparison asks whether GAUGE-LR+ beats each baseline consistently across configurations. Two paired tests answer it: an exact binomial *sign* test on the number of configurations favouring the reference, which assumes nothing but discards the magnitude of each difference, and the Wilcoxon signed-rank test on the same per-configuration differences, which uses the magnitudes and assumes the differences are symmetric about their median. Both are Holm-corrected over the same family of 14 comparisons, and on the reported roster they agree on every baseline: 13 are significant under either, and $P(\text{True})$ at 30B is the one comparison neither resolves. We report Wilcoxon as the primary test and verify its assumption per baseline rather than assert it (Table 28); where symmetry

fails at $\alpha = 0.05$ the conclusion rests on the sign test, and the last column records which. **No claim in this paper depends on the choice of test.**

Table 28: Paired-test diagnostics for the aggregate comparison against GAUGE-LR+. Skew and its test are computed on the per-configuration differences; the two p-columns are Holm-corrected over the full family. Where symmetry fails the conclusion is taken from the sign test, which does not assume it.

Baseline	Skew	Sym. p	Sign ↓	Wilcoxon ↓	Claim rests on
Terminal hidden distance	-0.07	0.878	0.0068	0.0039	Wilcoxon
Sequence confidence	-0.19	0.660	0.0013	0.0027	Wilcoxon
P(True)	+0.32	0.449	0.0000	0.0000	Wilcoxon
PRM (mean step)	-0.49	0.254	0.0013	0.0000	Wilcoxon
Response length (shorter first)	-0.50	0.248	0.0068	0.0030	Wilcoxon
LLM-as-a-judge (self)	-0.55	0.205	0.0000	0.0000	Wilcoxon
Self-consistency	-0.57	0.195	0.0260	0.0434	Wilcoxon
Untyped hidden drift	-0.58	0.188	0.0001	0.0000	Wilcoxon
Paraphrase agreement	-0.66	0.134	0.0013	0.0049	Wilcoxon
PRM (min step)	-0.69	0.119	0.0068	0.0001	Wilcoxon
LLM-as-a-judge (strong, 30B)	-0.99	0.032	0.0000	0.0000	sign test
P(True) (strong, 30B)	-1.06	0.024	0.0956	0.0676	*not claimed*
Supervised hidden probe	-1.12	0.017	0.0068	0.0070	sign test
Min token logprob	-1.69	0.001	0.0013	0.0020	sign test

M LIMITATIONS

- **The robustness label is a proxy.** It is a finite conjunction over three held-out edits, so it undercounts fragility that those edits do not expose and it is noisier than an exhaustive test would be. We therefore also report per-edit pass rates, and we choose held-out edits whose mechanism differs from every probe.
- **Hidden-state access.** GAUGE cannot be applied to a model served behind an API that does not expose activations. The answer-agreement baselines can.
- **Model scale and openness.** Every policy studied is open-weight, a direct consequence of needing hidden states: API-only models cannot be evaluated at all. Four of the five are $\leq$8B; Qwen3.8-27B extends the range but is still far from frontier scale, and it mainly shows where the method stops paying rather than that it keeps paying: fragility falls to 7.3% there and no verifier separates from another, every interval spanning zero. Its within-configuration intervals have half-width 0.072, above the +0.036–+0.043 the method carries over the cheapest baselines, so the experiment is not powered to detect its own effect size there. We therefore have direct evidence that the typed-drift geometry survives to 27B and none that the *comparison* remains resolvable above it.
- **The margin over the cheapest baselines is narrow.** Self-consistency is surpassed, but by +0.043 ($p_{\text{Holm}} = 0.043$)—the smallest margin we resolve—and P(True) at 30B is not surpassed under multiplicity correction, separating only at the operating point. Against answer-level baselines GAUGE therefore buys a calibrated threshold and a consistent but small ranking gain, not a decisive one; its large margins are over judge, process-reward and hidden-state verifiers.
- **The downstream gain is consistent but small.** GAUGE (correct-only) is the best deployable selector in all 18 columns of Table 1, recovering 60–67% of the oracle's headroom, but its strict-robust margin over outcome-RFT is $+0.8$ to $+1.0$pp on three seeds. Welch tests on the seed means and standard deviations of Table 1 give $p = 0.029$, 0.055 and 0.040 for Qwen3-4B, Llama-3.1-8B and Olmo-3-7B, none of which survives Holm correction across the three policies. The direction replicates on every policy and suite; the magnitude needs more seeds, and the verifier results do not depend on it.
- **The agentic verifier comparison is not powered, and on τ-bench not applicable.** ALFWorld between-arm differences carry intervals of roughly ± 0.15, because the independent unit is a solved game and only 13–19 per policy contain a fragile candidate. τ-bench is a different failure: the fitted contrast retains no direction with $\gamma_j \geq 1$ on either policy (Appendix K), so the metric has no contrastive structure to measure there and sits at chance irrespective of rank. We report τ-bench as

a fragility measurement, and treat the absence of any such direction as the criterion for whether a configuration can enter a verifier comparison at all.

- **Two corners are not resolved.** The label-free coupling does not separate from typed-probe agreement (-0.008, [-0.018, +0.001]): without fitted coefficients, most of what it knows is already carried by counting typed disagreements. And the fully unsupervised, fully free corner—GAUGE-D distilled from the label-free teacher—does not separate from response length or terminal hidden distance; removing labels and probes at once costs more than removing either alone.
- **The operating-point ordering is not uniform in coverage.** At c=20% P(True) at 30B has lower risk (0.054 against 0.063); GAUGE-LR+ leads from c=30% upward, where the resolved separation against self-consistency also sits.
- **Probe cost.** Full GAUGE multiplies generation cost per scored candidate. GAUGE-D removes that cost but does not fully recover the accuracy of the probe-based score.
- **Scope of the guarantee.** Theorem 16 covers the marginal distribution of the instability score under exchangeability. It does not certify that an accepted trajectory is semantically robust, a distribution shift between calibration and deployment voids it, and it covers the group-aware threshold of Section 3.5, not the pooled one the self-training runs used (their decisions differ on 2.0% of candidates).
- **The noise correction is heuristic.** Equation (11) subtracts a matched-noise term whose unbiasedness requires assumptions that transformations can violate; we keep λ_N frozen rather than claiming a control variate.
- **The equivariant family needs a computable map.** It is empty for answers that admit no such map, including part of MATH-500; the corresponding residual is then zero and GAUGE runs on $\mathcal{S}$ and $\mathcal{C}$ alone. Frequencies are reported in Appendix Q.
- **Paraphrase validation is automatic.** Probe and label paraphrases are checked for exact preservation of every number and named entity and for length plausibility, but they are not human-verified.
- **One trained policy family.** The policy-improvement study fine-tunes a single base model with LoRA over a limited number of seeds and iterations; the verifier study is broader than the training study.
- **Agentic numbers are not leaderboard numbers.** Our τ-bench user simulator is a local open model rather than the proprietary simulator used in the original benchmark, so absolute pass rates are not comparable to published values; all comparisons here are internal and use the identical simulator for every method.

N EXTENDED DISCUSSION

What the typed families buy. The component ablation separates two effects that are usually conflated. Answer agreement under perturbation is a strong signal on its own, and any honest comparison has to include it; the typed construction adds two things that agreement counting cannot express. First, the *control* family probes a property that has no answer-level signature at all in the common case: a trajectory can keep its answer under an injected cue while moving a long way in representation space, and that movement is predictive. Second, the *equivariant* family tests whether the policy tracks a stated change of goal, which is a different capability from invariance and which agreement under paraphrase never exercises.

Why the label-free coupling costs so little. Proposition 12 explains the empirical gap between the free and oracle couplings: they are identical on every terminally correct candidate, and can differ only where the candidate is already wrong. A verifier that is meant to *find* wrong candidates therefore loses exactly the part of the signal it should not have been allowed to use in the first place. This is what converts GAUGE from a training-time selector, which is what the score was in its original form, into a score that can be computed at deployment.

The cost axis is where the method has to be judged. A verifier that needs $|\mathcal{S}| + |\mathcal{V}| + |\mathcal{C}| + 1$ extra rollouts per candidate is not free, and we report that cost rather than hiding it. Two operating points matter. At full probe cost, GAUGE is the accurate end of the frontier. At zero probe cost, GAUGE-D reuses hidden states the policy already computed and is orders of magnitude cheaper than

any generative verifier, including the judge and the process reward model, both of which require a second forward pass over the whole trajectory.

O NOTATION SUMMARY

Symbol	Meaning
$\xi = (x, a^\star, \mathfrak{E}, m)$	Instance: prompt, reference answer, tool environment, equivalence predicate
ζ, $\widehat{a}(\zeta)$	Rollout and its parsed final answer (action signature in the agentic setting)
$\zeta_{ij,s}$	Rollout of candidate j on edited instance $s(\xi_i)$
J, j	Candidate budget and candidate index
O, $\mathrm{Robust}_{\mathcal{S}}$	Terminal correctness and the held-out robustness label
$\mathcal{S}, \mathcal{V}, \mathcal{C}$	Semantics-preserving, answer-equivariant, and control families
τ_g, w_g	Known total answer map and answer-coupling weight of g
w^{free}, w^{oracle}	Couplings using $\tau_g(\widehat{a}_0)$ and $a_g^\star$ respectively
ϕ, $z(\phi)$, $q(\phi)$	Pooled feature, PCA coordinate, task coordinate
K, r	PCA rank and fitted subspace rank
C_G, $\widehat{\Sigma}_G$	Family scatter and its regularized version
A, B, γ_j, U	Eigenpencil matrices, generalized eigenvalues, basis
σ_G	Frozen family scale; $\sigma(u)$ with an argument is the logistic function
$D_{\mathcal{S}}, D_{\mathcal{C}}, D_{\mathcal{V}}, D_N$	Residual components of the score
$\lambda_C, \lambda_V, \lambda_N, \lambda_{\mathrm{corr}}$	Control, equivariant, noise, and answer-coupling coefficients
R_{GAUGE}, ν, $\widehat{q}_{1-\alpha}$	Residual, instability score ($\nu = R_{\mathrm{GAUGE}}$), conformal threshold
$R^{\mathrm{D}}_{\mathrm{GAUGE}}$	Distilled, probe-free score (GAUGE-D)
$\mathcal{R}(\xi, \zeta)$	GRPO scalar reward

P SETUP DETAILS, SPLITS AND GENERATION VOLUME

Decoding and features. Rollouts are generated with vLLM (Kwon et al., 2023). Hybrid-reasoning checkpoints run in non-thinking mode so rollout length is comparable across policies. Decoding is stochastic (temperature 0.8, top-$p = 0.95$) with a per-request seed, so every rollout in a probe bundle is an independent draw as Definition 3 requires. Hidden states are read at 60% of model depth, fixed a priori. Section 4.6 fine-tunes Qwen3-4B, Llama-3.1-8B and Olmo-3-7B under LoRA.

Splits and probes. Table 29 gives the split sizes and generation volume per run. Each task is split at the level of prompt groups into fit (30%), calibration (25%), selection (20%) and evaluation (25%); GSM-Symbolic instantiations of one template form a single group, so no template appears in two splits. For every candidate independently we draw one rollout per typed probe (up to three $\mathcal{S}$, two $\mathcal{V}$, two $\mathcal{C}$), one matched noise replicate and, outside the fit split, one rollout per held-out label transformation. The staged search that fixes $(K, r, \lambda_C, \lambda_{\mathrm{corr}}, \lambda_V, \lambda_N)$ is in Appendix T (Table 31); the transformation catalogue with every answer map is in Appendix Q (Table 30).

Table conventions. Bold with a blue tint marks the best entry in its column and is the only use of bold on a number; a light blue row or column marks a method introduced here, including the perturbation-agreement selectors built from GAUGE's own probes; a warm tint flags a value and is never bold; arrows give a metric's direction; the two controls—a random scorer and terminal correctness on the original prompt—sit below a rule and are excluded from every best-entry decision. Where a table reports a gap rather than a ranking nothing is bold. Agentic numbers come from ALFWorld valid-unseen games, and τ-bench contributes fragility rows only (Appendix K).

Q TRANSFORMATION CATALOGUE

Table 30 lists every probe and held-out label transformation with its family, its role and its answer map; Appendix R explains where each one comes from and how it is validated.

Table 29: Splits and generation volume per run.

Model	Task	n	fit	cal	sel	eval	J	rollouts
Qwen3-4B	GSM8K	700	210	175	140	175	8	60440
	GSM-Symbolic-P2	300	90	72	60	78	8	25880
	MATH-500	400	120	100	80	100	8	29768
	SVAMP	300	90	75	60	75	8	25992
	GSM-IC	600	180	150	120	150	8	52376
	MMLU-Pro	400	120	100	80	100	8	31408
Qwen3-8B	GSM8K	700	210	175	140	175	8	60440
	GSM-Symbolic-P2	300	90	72	60	78	8	25880
	MATH-500	400	120	100	80	100	8	29768
	MMLU-Pro	400	120	100	80	100	8	31408
Llama-3.1-8B	GSM8K	700	210	175	140	175	8	60440
	GSM-Symbolic-P2	300	90	72	60	78	8	25880
	MATH-500	400	120	100	80	100	8	29768
	MMLU-Pro	400	120	100	80	100	8	31408
Olmo-3-7B	GSM8K	700	210	175	140	175	8	60440
	GSM-Symbolic-P2	300	90	72	60	78	8	25880
	MATH-500	400	120	100	80	100	8	29768
Qwen3.8-27B	GSM8K	700	210	175	140	175	8	60440
	GSM-Symbolic-P2	300	90	72	60	78	8	25880
	MATH-500	400	120	100	80	100	8	29768
	SVAMP	300	90	75	60	75	8	25992
	GSM-IC	600	180	150	120	150	8	52376
	MMLU-Pro	400	120	100	80	100	8	31408

Table 30: Typed transformation catalogue: every probe and held-out label transformation, its family, and its answer map.

Transformation	Family	Role	Answer map	Inst. (%)
small integers rewritten as English words	$\mathcal{S}$	label	$\tau_g = \mathrm{id}$	72
irrelevant but plausible sentence inserted (GSM-IC style)	$\mathcal{S}$	label	$\tau_g = \mathrm{id}$	84
second independent LLM paraphrase (label)	$\mathcal{S}$	label	$\tau_g = \mathrm{id}$	85
misleading topical framing sentence	$\mathcal{C}$	probe	$\tau_g = \mathrm{id}$	100
stated wrong 'previous answer' derived from the problem text	$\mathcal{C}$	probe	$\tau_g = \mathrm{id}$	100
premises restated as a bullet list, question last	$\mathcal{S}$	probe	$\tau_g = \mathrm{id}$	84
reworded task instruction wrapper	$\mathcal{S}$	probe	$\tau_g = \mathrm{id}$	74
consistent renaming of persons/entities	$\mathcal{S}$	probe	$\tau_g = \mathrm{id}$	32
LLM paraphrase, numbers/entities checked (probe)	$\mathcal{S}$	probe	$\tau_g = \mathrm{id}$	89
final answer must be reported shifted by c	$\mathcal{V}$	probe	$\tau_g(a) = a + c$	79
option list permuted (derangement)	$\mathcal{V}$	probe	$\tau_g =$ label permutation	18
second option permutation	$\mathcal{V}$	probe	$\tau_g =$ label permutation	18
final answer must be reported scaled by k	$\mathcal{V}$	probe	$\tau_g(a) = ka$	79

R PROVENANCE AND VALIDATION OF THE TRANSFORMATIONS

Transformations come from two sources, and no transformation in this paper is authored by hand.

Programmatic transformations. Entity renaming, bullet reformatting, instruction rewording, digit-to-word substitution, distractor insertion, and both diagnostic controls are deterministic Python operating on the instance text. Each draws from a generator seeded by `SHA-256(iid | name)`,

so a given instance always receives the same variant and any run is reproducible from the instance identifier alone. A transformation returns nothing when its precondition fails—no capitalised entity to rename, fewer than two sentences to reorder—which is the main reason instantiation rates in Table 30 fall short of 100%. The answer-equivariant family is exact by construction: scaling and offset wrap the prompt in an instruction and carry the closed-form map $\tau_v(a) = ka$ or $a + c$, and option permutation carries the induced label permutation, so τ_v is read off the transformation rather than estimated. The wrong-answer cue of the $\mathcal{C}$ family is the one transformation that reads an answer. Its decoy is built from the question text—the sum of the stated quantities, their maximum times their count, their mean—and screened so that it is never the answer; when the question states no numbers the decoy is derived from the answer value itself, and for multiple choice it is a non-answer option. As generated for the experiments, "the answer" was the reference. The reference-free rule substitutes the candidate's own answer $\widehat{a}_0$ everywhere, exactly as the free coupling substitutes $\tau_g(\widehat{a}_0)$ for $a_g^\star$: on a correct candidate the two answers are equivalent, so the two rules should produce the same cue. We rebuilt every cue both ways for all 17,204 correct evaluation candidates of the 23 verifier configurations: the reference rule reproduces the stored cue in 100% of cases and the candidate's-own-answer rule in 17,171 (99.8%); the 33 exceptions are answer-format variants (`\dfrac` against `\frac`) that the numeric parser does not unify. The conditional results are therefore, up to those cases, the results of a probe pipeline that never reads the reference. Offline fitting is a different matter: the frozen coefficients and GAUGE-LR+ are selected against robustness labels, which require reference answers on the fit and selection splits (Appendix T).

Model-authored paraphrases. Only the two paraphrase transformations are model-authored. Both are written once, before any policy is run, by a fixed author model, Qwen3-30B-A3B-Instruct-2507, and shared unchanged across every evaluated policy, so no policy is scored against text tuned to it. The author is not one of the five evaluated policies, though it shares a family with three of them; it is also the model used for the strong-judge and strong-$P(\text{True})$ baselines, which read rollouts and never the paraphrase prompts. Two independent prompts produce two independent samples per instance: slot a is used *only* as a probe and slot b *only* in the held-out robustness label. The prompts instruct the author to preserve every number, unit and named entity, to keep the question and its correct answer, and not to solve the problem.

The acceptance rule. A candidate paraphrase is accepted only if the multiset of numbers it contains is identical to the multiset in the original, its length is between $0.55\times$ and $2.0\times$ the original, it is not a copy of the original, it is at least twenty characters, and it contains no answer-delimiter leakage (`####` or `\boxed`). A rejected candidate is resampled up to three times at rising temperature ($0.7 + 0.1 \cdot$ attempt); if it still fails it is dropped, and that instance simply carries no paraphrase probe. Acceptance is 89% for the probe slot and 85% for the label slot (Table 30).

What this does and does not guarantee, and the selection effect it induces. Number-multiset equality is a strong check for arithmetic word problems, where the quantities determine the answer, and it is what makes the $\mathcal{S}$ family safe to treat as answer-preserving at scale. It is not a check on meaning: a rewrite could preserve every numeral and still alter the question, and our filter would accept it. Such a case would place a genuinely answer-changing edit inside $\mathcal{S}$. Where that happens among the *probe* families it weakens the contrast the subspace is fitted on, which works against us; where it happens among the *held-out* edits of Equation (1) it makes a sound derivation look brittle and inflates the measured gap, which does not. The direction of the error therefore depends on which mechanism draws the bad edit, and we do not claim it is conservative overall; Appendix H.3 bounds the gap's size against an unedited-repeat control. The drop policy is the second caveat and cuts the other way: the instances whose paraphrases fail three attempts are plausibly the syntactically awkward ones, so the paraphrase probe is present on a mildly non-random subset. We report instantiation rates per transformation rather than averaging the effect away, and the remaining $\mathcal{S}$ probes on those instances are programmatic and unaffected.

Disjointness of probe and label mechanisms. The separation asserted in Section 3.2 holds by construction, not by inspection. Semantics-preserving *probes* are entity renaming, bullet reformatting and instruction rewording; semantics-preserving *labels* are distractor insertion and digit-to-word substitution. The two sets share no mechanism. The two paraphrases are independent samples from

different prompts, so the probe paraphrase and the label paraphrase are not variants of one another either. A verifier therefore cannot score well by recognising the specific edits it was fitted on.

S IMPLEMENTATION-FAITHFULNESS CHECKLIST

The equations in the method section describe the score and subspace estimator exactly as implemented. A reproducible implementation must record:

1. the fit, calibration, selection, and evaluation instance identifiers, and evidence that the four splits are disjoint at the prompt-group level;
2. the checkpoint, decoding parameters, candidate budget J, seeds, and library versions;
3. the hidden layer, token-position rule, pooling operator, PCA rank K, and subspace rank r;
4. $\rho, \eta, \lambda_C, \lambda_{\text{corr}}, \lambda_V, \lambda_N, \varepsilon$, and $\beta_{\text{DPO}}, \beta_R$ for the policy-improvement variants;
5. evidence that probe bundles satisfy Definition 3, i.e. that no rollout is shared across probes or candidates;
6. all transformation templates, their answer maps, and family labels;
7. the fitted values of $\sigma_{\mathcal{S}}, \sigma_{\mathcal{V}}, \sigma_{\mathcal{C}}$ together with the consistency check of Remark 11;
8. singleton-equivariant fallback counts and noise-replicate settings; and
9. the calibration unit, threshold index k, n_{cal}, tie handling, and multi-survivor rule.

Every item is emitted automatically by the released code as JSON alongside each run.

T FROZEN HYPERPARAMETERS AND THE SELECTION LOG

Table 31: The frozen GAUGE hyperparameters. Every value is fixed by the staged search on the selection split of the primary policy and applied unchanged to every other configuration; the evaluation split is never consulted.

Hyperparameter	Frozen value
K	256
r	256
ρ	0.1
η	0.0001
λ_C	0.5
λ_{corr}	0
λ_V	4
λ_N	-0.5
selection split mean AUROC	0.8607

The staged search evaluated geometry: 12 settings, best K=256, r=256 (0.7292); coupling: 18 settings, best λ_C=0.5, λ_{corr}=0 (0.7692); coefficients: 30 settings, best λ_V=4, λ_N=-0.5 (0.8607). The rank grid was widened twice and the sign of λ_N freed after the initial passes indicated an optimum at the boundary; every extension used the selection split only, and the evaluation split was untouched throughout.

U FITTED VERIFIER COEFFICIENTS

Table 32 lists the fitted GAUGE-LR+ coefficients for every configuration.

Table 32: Coefficients of the fitted GAUGE-LR+ combination, per configuration, on standardised features. The sign of D_N is the matched-noise term discussed in Section 3.4.

Model	Task	$D_{\mathcal{S}}$	$D_{\mathcal{C}}$	$D_{\mathcal{V}}$	D_N	$\bar{\alpha}_{\mathcal{S}}$	$\bar{\alpha}_{\mathcal{V}}$	$\bar{\alpha}_{\mathcal{C}}$	mean lp	min lp	log len
Llama-3.1-8B	GSM8K	-0.35	+0.53	+0.22	-0.02	+0.68	+0.60	+0.95	+0.17	+0.16	-0.31
Llama-3.1-8B	GSM-Symbolic-P2	-0.20	+0.56	+0.16	+0.09	+1.09	+0.42	+0.94	+0.43	-0.02	-0.12
Llama-3.1-8B	MATH-500	-0.43	-0.40	-0.07	-0.19	+0.95	+0.17	+0.53	-0.20	+0.32	-0.44
Llama-3.1-8B	MMLU-Pro	+0.09	+0.26	-0.89	-0.35	+0.94	-0.18	+0.80	+0.20	-0.35	-0.09
Olmo-3-7B	GSM8K	-0.73	-0.11	-0.03	-0.09	-0.18	+0.67	+0.33	+0.37	+0.09	-0.20
Olmo-3-7B	GSM-Symbolic-P2	-0.06	+0.28	-0.45	+0.09	+0.93	+0.20	+0.79	+0.13	+0.12	-0.02
Olmo-3-7B	MATH-500	-0.29	-0.07	-0.25	+0.21	+0.79	+0.01	+0.73	+0.06	-0.43	-0.37
Qwen3-4B	GSM8K	+0.16	+0.42	+0.11	-0.06	+0.71	+0.35	+0.79	+0.54	-0.00	-0.26
Qwen3-4B	GSM-IC	+0.23	-0.49	+0.11	+0.26	+1.08	+0.32	-0.36	+0.58	+0.25	-0.25
Qwen3-4B	GSM-Symbolic-P2	-0.44	-0.16	-0.45	+0.23	+1.11	+0.44	+0.14	+0.08	+0.32	-0.25
Qwen3-4B	MATH-500	+0.27	+0.76	+0.04	-0.36	+0.94	+0.32	+1.45	+0.53	-0.21	-0.01
Qwen3-4B	MMLU-Pro	+0.14	-0.34	-0.22	+0.16	+1.47	+0.53	+0.50	+0.67	-0.41	-0.42
Qwen3-4B	SVAMP	-0.56	+1.23	-0.02	+0.06	+0.25	+1.08	+0.84	-0.49	+0.17	-1.08
Qwen3-8B	GSM8K	-0.37	+0.53	+0.12	-0.03	+0.28	+0.54	+0.80	+0.08	+0.03	-0.42
Qwen3-8B	GSM-Symbolic-P2	-0.07	-0.16	-0.27	-0.16	+0.66	+0.46	+0.76	+0.01	-0.17	+0.05
Qwen3-8B	MATH-500	+0.07	-0.09	-0.03	-0.01	+1.22	+0.01	+0.38	+0.20	-0.15	-0.37
Qwen3-8B	MMLU-Pro	+0.15	-0.34	+0.04	-0.35	+0.99	+0.64	+0.20	+0.66	-0.07	-0.26
Qwen3.8-27B	GSM8K	+0.39	-0.71	-0.38	-0.02	+0.77	+0.13	-0.33	+0.57	-0.46	-0.45
Qwen3.8-27B	GSM-IC	+0.59	+0.62	-0.11	-0.07	+1.13	+0.08	+0.72	+0.41	-0.12	-0.20
Qwen3.8-27B	GSM-Symbolic-P2	+0.18	-0.03	+0.01	-0.03	+0.48	+0.34	-0.08	-0.08	-0.40	-0.06
Qwen3.8-27B	MATH-500	-0.22	-0.41	-0.02	+0.21	+0.09	-0.04	-0.03	-0.96	+0.40	+0.17
Qwen3.8-27B	MMLU-Pro	+0.02	+0.35	-0.51	-0.29	+1.52	+0.42	+1.22	-0.48	+0.63	+0.14
Qwen3.8-27B	SVAMP	-0.30	-0.07	-0.41	+0.46	+0.52	-0.07	-0.37	+0.06	-0.03	-1.44
mean		-0.08	+0.09	-0.14	-0.01	+0.80	+0.32	+0.51	+0.15	-0.02	-0.29

V TRANSFORMATION FAMILY SIZES

Table 33 reports, per configuration, how many probes each family contributes and how often the method's two edge cases occur.

Table 33: Per-candidate family sizes and the frequency of the two edge cases the method defines: an empty answer-equivariant family (no computable answer map, which happens for non-numeric MATH-500 answers and for the 5% of τ-bench tasks whose request exposes no swappable entity) and the singleton-$\mathcal{V}$ fallback of Equation (17). The last column is the number of held-out edits the robustness label conjoins, which differs by task because some edits do not instantiate on some inputs.

Model	Task	$\lvert\mathcal{S}_i\rvert$	$\lvert\mathcal{V}_i\rvert$	$\lvert\mathcal{C}_i\rvert$	$\mathcal{V}$ empty (%)	$\mathcal{V}$ singl. (%)	held-out
Llama-3.1-8B	GSM8K	2.93	2.00	2.00	0.0	0.0	2.67
Llama-3.1-8B	GSM-Symbolic-P2	2.90	2.00	2.00	0.0	0.0	2.69
Llama-3.1-8B	MATH-500	2.45	1.46	2.00	27.3	0.0	2.00
Llama-3.1-8B	MMLU-Pro	2.58	2.00	2.00	0.0	0.0	1.76
Olmo-3-7B	GSM8K	2.93	2.00	2.00	0.0	0.0	2.67
Olmo-3-7B	GSM-Symbolic-P2	2.90	2.00	2.00	0.0	0.0	2.69
Olmo-3-7B	MATH-500	2.45	1.46	2.00	27.3	0.0	2.00
Qwen3-4B	GSM8K	2.93	2.00	2.00	0.0	0.0	2.67
Qwen3-4B	GSM-IC	2.98	2.00	2.00	0.0	0.0	2.76
Qwen3-4B	GSM-Symbolic-P2	2.90	2.00	2.00	0.0	0.0	2.69
Qwen3-4B	MATH-500	2.45	1.46	2.00	27.3	0.0	2.00
Qwen3-4B	MMLU-Pro	2.58	2.00	2.00	0.0	0.0	1.76
Qwen3-4B	SVAMP	2.99	2.00	2.00	0.0	0.0	2.63
Qwen3-8B	GSM8K	2.93	2.00	2.00	0.0	0.0	2.67
Qwen3-8B	GSM-Symbolic-P2	2.90	2.00	2.00	0.0	0.0	2.69
Qwen3-8B	MATH-500	2.45	1.46	2.00	27.3	0.0	2.00
Qwen3-8B	MMLU-Pro	2.58	2.00	2.00	0.0	0.0	1.76
Qwen3.8-27B	GSM8K	2.93	2.00	2.00	0.0	0.0	2.67
Qwen3.8-27B	GSM-IC	2.98	2.00	2.00	0.0	0.0	2.76
Qwen3.8-27B	GSM-Symbolic-P2	2.90	2.00	2.00	0.0	0.0	2.69
Qwen3.8-27B	MATH-500	2.45	1.46	2.00	27.3	0.0	2.00
Qwen3.8-27B	MMLU-Pro	2.58	2.00	2.00	0.0	0.0	1.76
Qwen3.8-27B	SVAMP	2.99	2.00	2.00	0.0	0.0	2.63

W FULL PER-TASK RESULTS

Table 34 gives verifier AUROC for every model–task configuration, with prompt-level bootstrap intervals.

Table 34: Verifier AUROC against the held-out robustness label for every model–task configuration, with 95% prompt-level cluster-bootstrap intervals beneath each estimate. The best verifier in each row is highlighted; leading zeros are dropped in the intervals.

Model	Task	GAUGE ↑	GAUGE-D ↑	Typed-probe ↑	Self-cons. ↑	PRM (min) ↑
Qwen3-4B	GSM8K	**.871** [.806, .932]	.784 [.682, .874]	.846 [.761, .925]	.751 [.663, .846]	.773 [.672, .869]
	GSM-Symbolic-P2	.866 [.782, .918]	.726 [.607, .843]	**.878** [.799, .928]	.862 [.782, .915]	.584 [.376, .774]
	MATH-500	.873 [.789, .940]	.826 [.738, .901]	.884 [.790, .959]	**.886** [.795, .959]	.439 [.294, .595]
	SVAMP	.865 [.755, .957]	**.915** [.857, .958]	.883 [.771, .965]	.836 [.704, .954]	.909 [.827, .967]
	GSM-IC	.837 [.744, .928]	**.847** [.739, .932]	.830 [.728, .924]	.739 [.609, .871]	.578 [.414, .755]
	MMLU-Pro	.869 [.817, .914]	.824 [.763, .877]	**.875** [.821, .922]	.772 [.703, .835]	.597 [.496, .687]
Qwen3-8B	GSM8K	.800 [.718, .874]	.579 [.516, .647]	**.802** [.724, .873]	.786 [.706, .860]	.698 [.604, .783]
	GSM-Symbolic-P2	.784 [.671, .871]	.655 [.538, .749]	**.811** [.712, .890]	.722 [.586, .836]	.472 [.255, .697]
	MATH-500	.835 [.741, .906]	.745 [.645, .841]	.850 [.771, .914]	**.853** [.770, .926]	.500 [.389, .619]
	MMLU-Pro	.848 [.792, .896]	.683 [.619, .745]	**.865** [.814, .911]	.818 [.757, .873]	.573 [.471, .667]
Llama-3.1-8B	GSM8K	**.877** [.837, .908]	.806 [.758, .848]	.870 [.831, .902]	.852 [.803, .897]	.699 [.638, .757]
	GSM-Symbolic-P2	.899 [.824, .949]	.787 [.698, .852]	.919 [.842, .965]	**.924** [.840, .968]	.621 [.363, .819]
	MATH-500	.870 [.828, .905]	.843 [.795, .884]	.893 [.857, .923]	**.908** [.872, .936]	.567 [.477, .650]
	MMLU-Pro	.886 [.841, .921]	.647 [.568, .726]	**.901** [.858, .934]	.872 [.814, .916]	.614 [.511, .706]
Olmo-3-7B	GSM8K	.842 [.782, .891]	.775 [.714, .833]	**.853** [.796, .902]	.818 [.747, .882]	.709 [.621, .791]
	GSM-Symbolic-P2	.888 [.797, .951]	.782 [.653, .878]	**.896** [.809, .954]	**.896** [.819, .951]	.648 [.410, .823]
	MATH-500	**.878** [.808, .935]	.771 [.693, .841]	**.878** [.803, .936]	.858 [.779, .924]	.565 [.453, .675]
Qwen3.8-27B	GSM8K	**.840** [.743, .920]	.770 [.683, .849]	.829 [.735, .911]	.741 [.631, .843]	.720 [.594, .845]
	GSM-Symbolic-P2	.817 [.516, .909]	.690 [.358, .813]	**.835** [.593, .903]	.804 [.587, .875]	.719 [.329, .874]
	MATH-500	.733 [.608, .860]	.660 [.528, .797]	.738 [.618, .852]	**.779** [.648, .904]	.418 [.261, .586]
	SVAMP	**.906** [.821, .972]	.803 [.670, .921]	.884 [.781, .967]	.853 [.724, .964]	.881 [.792, .954]
	GSM-IC	.636 [.406, .905]	.662 [.479, .883]	.675 [.495, .899]	**.678** [.500, .900]	.490 [.248, .770]
	MMLU-Pro	**.868** [.800, .920]	.841 [.761, .908]	**.868** [.803, .923]	.723 [.648, .799]	.572 [.466, .675]