PROSPECTIVE GROUNDING: WHEN PROXY DIAGNOSTICS WARRANT CONTROL OVER INFERENCE-TIME SEARCH

Anonymous authors
Paper under double-blind review

ABSTRACT

Inference-time search can increase proxy reward while losing content already satisfied by the current output. When should a diagnostic of this risk guide acceptance, verification, or refinement? We introduce the *Prospective Grounding Certificate* (PGC), a protocol for qualifying diagnostics before they control search. PGC tests predictive improvement, agreement between automatic outcomes and human judgment, and the semantic value of a specified policy change. We show that the same observable proxy process can support opposite semantic risks, establishing why repeatability alone cannot justify diagnostic use. Before collecting labels, exact checks rule out comparisons that cannot attain the required predictive improvement or change a policy action. Independent outcomes then test the three claims with family-wise false-authorization control. Controlled component deletions show how each test blocks a distinct failure that the other two miss. A text-to-image analysis uses 1,000 cases sampled from a 1,227-case frame and 3,000 blinded paired ratings. The target-attribute diagnostic clears all three criteria for the registered finite-frame policy comparison. The certificate links a diagnostic's predictive contribution to the semantic value of the search decisions it changes.

1 INTRODUCTION

"A map is not the territory."
Alfred Korzybski, New Orleans, 28 December 1931

Inference-time search navigates a space of candidate outputs using proxy scores (Snell et al., 2025; Huang et al., 2025; Zhang et al., 2026). Optimizing an imperfect proxy can improve its score while degrading the intended outcome (Gao et al., 2023; Skalse et al., 2022; Pan et al., 2022; Khalaf et al., 2025). In text-to-image generation, a replacement may receive a higher reward while omitting an object explicitly requested by the user. The score improves, but the output loses something worth preserving.

Diagnostics of this risk could help a search policy decide whether to accept a candidate, request verification, or continue refinement. A decrease in the candidate's target-attribute score, for example, could signal the loss of requested content. The central question is when such a diagnostic deserves that role. A signal may be stable across repeated measurements yet fail to predict semantic loss; even an informative signal may leave the preferred action unchanged.

We introduce the *Prospective Grounding Certificate* (PGC) to turn diagnostic authority into a testable claim about a fixed search policy. PGC joins predictive improvement, semantic measurement validity, and policy value under a false-authorization guarantee. The certificate establishes both what the diagnostic predicts and whether its induced policy change is worthwhile. Before collecting labels, exact support and action checks identify comparisons that cannot attain the requested gain or alter a decision.

The key distinction is between how often search replaces the current output and whether a replacement loses required content. We call these replacement exposure and semantic susceptibility. The proxy process identifies exposure, but requires an outcome coupling to identify susceptibility. Without that coupling, the same proxy process can support opposite semantic risks.

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

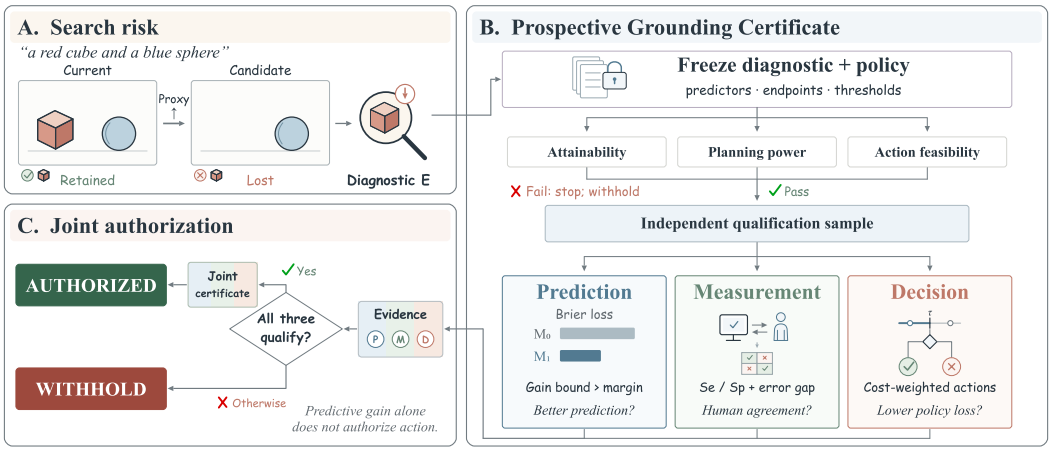


Figure 1: From search risk to joint authorization. (A) A higher-scoring replacement loses a required object, motivating prospective qualification of the diagnostic. (B) PGC fixes the diagnostic–policy tuple and checks attainability, planning power, and action feasibility before independent qualification. Prediction requires a lower confidence bound on gain above the margin; measurement tests sensitivity (Se), specificity (Sp), and between-stratum error gaps; decision tests cost-weighted semantic policy loss. (C) All three claims must qualify to issue a joint certificate and authorize the fixed tuple; otherwise, policy use is withheld. The example and miniatures illustrate the protocol schematically.

Controlled component deletions isolate the contribution of each qualification axis. The text-to-image study uses 3,000 blinded paired ratings to assess all three claims, contrasting successful qualification with a repeatable but ungrounded diagnostic. Figure 1 summarizes the protocol.

2 Related Work

Search and proxy optimization. Test-time scaling allocates additional computation to sampling, verification, or search (Snell et al., 2025; Huang et al., 2025; Zhang et al., 2026). Universal Guidance steers diffusion with pretrained guidance functions (Bansal et al., 2023). DOODL optimizes diffusion latents through the generated image (Wallace et al., 2023). Initial-noise optimization, constrained sampling, reward alignment, and seed pruning steer inference-time generation (Tang et al., 2025; Song et al., 2025; Kim et al., 2026; Guimaraes & Perona, 2026). Reliable-seed mining instead curates generated training sets for fine-tuning (Li et al., 2024). Optimizing an imperfect reward can improve the proxy while degrading the target (Gao et al., 2023; Skalse et al., 2022; Pan et al., 2022; Khalaf et al., 2025), although auxiliary proxies can improve preference learning under additional structure (Zhu et al., 2025). Pessimistic selection and image-space constraints address this risk during inference (Yu et al., 2026; Zhai et al., 2025); reward-sharpness-aware fine-tuning instead addresses reward sensitivity during model adaptation (Kim & Sim, 2026). PGC addresses the evidence needed to let a risk diagnostic govern these search decisions.

Identification and independent evaluation. Partial-identification analyses make explicit which conclusions follow from observations and which require additional assumptions (Tamer, 2010). Transport guarantees similarly require stable data-generating relations (Subbaswamy et al., 2019). For search, the unresolved quantity is the risk of semantic loss conditional on replacement. Our factorization separates this risk from the observable probability of replacement. Estimating the semantic relation also requires an evaluation design that controls feedback from method development. Unrestricted adaptive reuse of holdout feedback can induce overfitting; protected reuse requires additional mechanisms (Dwork et al., 2015; Blum & Hardt, 2015). PGC uses independent outcomes after fixing the diagnostic and its predictors.

Prediction and decision guarantees. Proper scoring rules support honest comparison of probabilistic forecasts (Brier, 1950; Gneiting & Raftery, 2007). Selective prediction learns an absten-

tion mechanism (Geifman & El-Yaniv, 2019), whereas risk-control methods calibrate procedures to bound an observed loss (Bates et al., 2021; Angelopoulos et al., 2025; 2024). Learn-then-Test also controls multiple risk requirements jointly (Angelopoulos et al., 2025, Section 2.4). Decision-focused learning connects predictive models to downstream utility (Wilder et al., 2019). PGC asks whether adding a diagnostic improves a fixed predictor and leads to a valuable policy change. Its exact support and action checks identify infeasible comparisons before collecting qualification outcomes.

Semantic measurement. Reliability and validity are different measurement properties (Schmidt et al., 2000), and benchmark scores can lack the construct validity needed for their intended interpretation (Blodgett et al., 2021; Bean et al., 2025; Caban, 2026). Question-answering and object-focused evaluators systematize text–image alignment (Hu et al., 2023; Ghosh et al., 2023; Kamath et al., 2025), while LLM-judge studies document human agreement and systematic bias (Liu et al., 2023; Zheng et al., 2023; Wang et al., 2024). Reward-model benchmarks and text–image metric studies expose related evaluator gaps (Kim et al., 2025; Ross et al., 2024).

PGC makes evaluator validity part of the authorization rule, alongside predictive gain and policy value. A shared paired audit can test the evaluator's semantic errors and the policy's human-assessed loss. This joins the three claims in a certificate for the diagnostic's use after selection. Appendix H compares the evaluation targets and guarantees.

3 Selection-Conditioned Grounding

Search state and semantic outcome. Fix a prompt c. Let A denote the observed candidate set, r the proxy reward, and $z_A^\star$ the winner under a fixed tie rule. The operator $\mathcal{S}$ uses unrevealed randomness ω to produce a future winner $z'^\star$. Let $x_A^\star$ and $x'^\star$ be their decoded images. For a binary attribute a, the rule $h_a^\star(c, x)$ denotes latent semantic correctness. We define the current winner, replacement event, and semantic downside jointly as

$$\begin{aligned} z_A^\star &= \arg\max_{z \in A} r(z), \quad z'^\star = \mathcal{S}(c, A, z_A^\star; \omega), \\ C &= \mathbf{1}\{z'^\star \neq z_A^\star\}, \quad D^\star = C\,\mathbf{1}\{h_a^\star(c, x_A^\star) = 1\}\mathbf{1}\{h_a^\star(c, x'^\star) = 0\}. \end{aligned} \tag{1}$$

The same notation covers pool expansion, local refinement, pruning, and verification.

Empirical qualification operationalizes the intended construct $D^\star$ as a human-semantic endpoint D^H fixed before sampling. Prediction is scored against the registered automatic endpoint $\widetilde{D}$, measurement qualifies $\widetilde{D}$ against D^H, and decision uses D^H directly. This separates scalable automatic prediction scoring from the human judgment that defines policy value.

Exposure–susceptibility factorization. Let X denote the prespecified ordinary covariates, E the candidate diagnostic, and $T^\star = \mathbf{1}\{h_a^\star(c, x_A^\star) = 1\}\mathbf{1}\{h_a^\star(c, x'^\star) = 0\}$ semantic loss conditional on replacement. Because $D^\star = CT^\star$, conditional downside satisfies

$$\mathbb{P}(D^\star = 1 \mid X, E) = \underbrace{\mathbb{P}(C = 1 \mid X, E)}_{\text{replacement exposure}} \underbrace{\mathbb{P}(T^\star = 1 \mid C = 1, X, E)}_{\text{semantic susceptibility}}. \tag{2}$$

The observable proxy law identifies replacement exposure under the declared search mechanism. Outcome data are needed to identify susceptibility. To quantify what the diagnostic adds, consider any registered binary outcome D (with $D = \widetilde{D}$ in the empirical prediction axis), let $\eta_X = \mathbb{P}(D = 1 \mid X)$ and $\eta_{X,E} = \mathbb{P}(D = 1 \mid X, E)$. Under Brier loss (Brier, 1950; Gneiting & Raftery, 2007), the incremental Bayes value of E is

$$\mathcal{R}^\star(X) - \mathcal{R}^\star(X, E) = \mathbb{E}[(\eta_{X,E} - \eta_X)^2]. \tag{3}$$

The value in Equation (3) is positive exactly when the diagnostic changes conditional outcome risk beyond X. This Bayes quantity motivates grounding; PGC tests incremental predictive value for a fixed predictor pair. The factorization and projection identity are proved in Appendices B.2 and B.1.

Theorem 1 (Assumption-minimal selection-conditioned non-identifiability)**.** *Let O collect the proxy-side records, including candidate images, scores, search selections, and diagnostic responses,*

but excluding semantic outcomes. Consider the class of semantic models that imposes no restriction linking the latent correctness rule to the observable proxy process. This class contains an observable law with positive-probability replacement and two coherent semantic correctness rules $h^\star_{a,0}, h^\star_{a,1}$ such that $\mathcal{L}_{\mathbb{P}_0}(O) = \mathcal{L}_{\mathbb{P}_1}(O)$ but the conditional downside laws differ on eligible replacements. Downside is not uniformly identified from proxy-side observations over this assumption-minimal class.

Appendix B.3 constructs both semantic worlds. The consequence for search control is that identical proxy evidence can justify opposite policy choices, as the following corollary shows.

Corollary 1 (No uniform proxy-only policy authority)**.** *The class in Theorem 1 contains an observable law and a nontrivial proxy-only accept–reject policy $\psi(O) \in \{0, 1\}$ whose risk ranking against ordinary search $\psi_0 = 1$ reverses between two compatible semantic laws. For every $\lambda > 0$, use*

$$\ell_\lambda(u, d) = ud + \lambda(1 - u)(1 - d), \qquad \mathfrak{R}_{\lambda,\mathbb{P}}(\psi) = \mathbb{E}_{\mathbb{P}}[\ell_\lambda(\psi(O), D) \mid C = 1]. \tag{4}$$

The two semantic laws have the same observable distribution, but ψ has lower replacement-conditioned risk in one and higher risk in the other.

Appendix B.5 gives the proof.

A sufficient coupling condition. Grounding becomes possible when the semantic coupling is explicit.

Proposition 1 (Stable grounding across prompt laws)**.** *Suppose $g(X, E)$ satisfies $\mathbb{E}[g(X, E) \mid X] = 0$ and $0 < \mathbb{E}[g(X, E)^2] < \infty$, and*

$$\mathbb{P}(D = 1 \mid X, E) = m(X) + \beta g(X, E) \tag{5}$$

for $\beta \neq 0$, with the right-hand side in $[0, 1]$. Then the incremental Bayes Brier value is $\beta^2\mathbb{E}[g(X, E)^2] > 0$. Under a new prompt law $\mathbb{Q}$, the value remains positive if the same conditional relation holds with $\beta \neq 0$, $\mathbb{E}_{\mathbb{Q}}[g(X, E) \mid X] = 0$, and $0 < \mathbb{E}_{\mathbb{Q}}[g(X, E)^2] < \infty$.

The proposition concerns Bayes information under an invariant conditional law. We test finite source-trained predictors separately by fixing the predictors and coupling direction in one setting, then evaluating them under a new prompt law. This parallels transport analyses that require a stable outcome relation across environments (Subbaswamy et al., 2019). The proof and an exact construct-positive world are in Appendices B.4 and B.6.

4 Prospective Grounding Certificate

The coupling condition explains when a diagnostic can carry outcome information. PGC tests whether a particular predictor and policy can use that information. It fixes a diagnostic, predictor pair, outcome instrument, and policy before qualification; we call this collection a *frozen tuple*. Feasibility screens determine whether the proposed comparison can succeed, and independent outcomes test its three claims.

Definition 1 (Diagnostic authority)**.** *For a frozen tuple, fix a baseline predictor $\mathcal{M}_0(X)$ and an augmented predictor $\mathcal{M}_1(X, E)$ that differs only through declared transformations of diagnostic E. Also fix an automatic endpoint $\widetilde{D}$, an operational human-semantic endpoint D^H, a measurement partition, two induced policies, and their semantic loss. The tuple has* diagnostic authority *when adding E clears the predictive margin for $\widetilde{D}$, the measurement claim linking $\widetilde{D}$ to D^H qualifies on the declared partition, and the augmented policy clears the semantic decision margin. Diagnostic authority is indexed by this tuple.*

4.1 Qualification sequence

Before outcome access, the freeze fixes the diagnostic family, predictors, margins, units, error allocation, measurement design, and policy costs. Table 1 orders the feasibility screens and qualification tests. A family that clears the screens is evaluated once on independent outcomes; Appendices C.1 and C.6 give the full specification.

For accept/reject policies, let $a = 1$ mean accept, p_j be predictor j's risk estimate, and $c_{\mathrm{FA}}, c_{\mathrm{FR}} > 0$ be false-accept and false-reject costs. The fixed policies use

$$L(a, d) = c_{\mathrm{FA}} ad + c_{\mathrm{FR}}(1 - a)(1 - d),$$
$$a_j = \mathbf{1}\{p_j < \tau\}, \qquad \tau = \frac{c_{\mathrm{FR}}}{c_{\mathrm{FA}} + c_{\mathrm{FR}}}.$$

Decision qualification requires a lower bound on $\mathbb{E}[L(a_0, D^H) - L(a_1, D^H)]$ above its margin. Before outcomes, action feasibility checks whether the fixed policies disagree. For prediction margin $m_P > 0$ and planned upper bound U_P, planning requires $\mathbb{P}_{\mathrm{plan}}(U_P < -m_P) \geq \pi_{\min}$, together with registered event-count requirements. The planning law and target power $\pi_{\min}$ are prespecified.

Table 1: PGC at a glance. Pre-outcome screens establish feasibility; independent outcomes qualify the three claims. Planning power is evaluated under the declared planning model.

Stage	What must pass?	Result
Before outcome collection		
Attainability	Target and planning effect within **exact support**	Effect feasible
Planning power	Planned **power** and **event-count** requirements	Study feasible
Action feasibility	Nonzero policy disagreement	Action feasible
Independent qualification		
Prediction	Paired-loss **UCB** below the negative margin	Certified
Measurement	Semantic **error bounds** inside the quality region	Qualified
Decision	Policy-gain **LCB** above the practical margin	Qualified

Compositional authorization control. For a registered tuple k, let P_k, M_k, and V_k denote its prediction, measurement, and decision-value claims, respectively. Let $\phi_{jk} = 1$ when the procedure for axis $j \in \{P, M, V\}$ passes tuple k. An invalid tuple has at least one false component claim. Authorization requires all three procedures to pass.

Theorem 2 (Compositional false-authorization control)**.** *For a prespecified finite family, fix error budgets α_{jk} before qualification. Suppose each axis procedure has false-passage probability at most α_{jk} under every law for which its component claim is false. If $\sum_k \max_{j \in \{P,M,V\}} \alpha_{jk} \leq \alpha$, authorizing tuple k only when $\phi_{Pk}\phi_{Mk}\phi_{Vk} = 1$ guarantees*

$$\mathbb{P}(\exists k : \ \textsc{Authorized}_k \text{ and } \neg(P_k \cap M_k \cap V_k)) \leq \alpha. \tag{6}$$

The result permits dependence among axes and tuples. Within a single tuple, each component may use level α.

The theorem controls erroneous *joint authorization*. The experiments use an additive split to control false component certificates simultaneously (Corollary 2).

The population prediction certificate below has a finite-frame counterpart. Paired audits supply simultaneous measurement intervals, and fixed-policy loss comparisons qualify decision value (Appendices D.3 and D.5). The axes can share the probability-sampled qualification cohort; measurement may use an equal-probability subsample within each registered stratum. Tuples in a prespecified family can share an audit when their endpoints and strata coincide, with multiplicity covered by the family budget.

Sharing observations does not make the claims interchangeable. For example, risks $(.01, .49)$ carry Brier information but induce the same accept action at threshold .5, even with perfect measurement. Appendix D.7 gives the component-wise counterexamples and proves the authorization theorem.

4.2 Attainability and certification

Index independent prompt-level units by $i = 1, \ldots, n$. For fixed baseline and augmented predicted probabilities $(p_{0,i}, p_{1,i})$ and their registered binary observed endpoint y, define

$$d_i(y) = (p_{1,i} - y)^2 - (p_{0,i} - y)^2, \qquad a_i = \min_{y \in \{0,1\}} d_i(y), \quad b_i = \max_{y \in \{0,1\}} d_i(y). \tag{7}$$

Let nonnegative weights w_i sum to one. Every quantity in Equation (7) is available before any outcome is observed.

Proposition 2 (Exact outcome-blind falsification)**.** *The exact set of attainable expected paired losses over all binary outcome laws is*

$$\mathcal{S}_w = \left[\sum_i w_i a_i,\ \sum_i w_i b_i \right]. \tag{8}$$

An improvement of at least m is attainable exactly when $\inf \mathcal{S}_w \le -m$.

If the lower endpoint exceeds $-m$, the requested gain is impossible under every binary outcome law for this predictor pair. PGC stops before collecting outcomes. Otherwise, it proceeds to the planning-power and action-feasibility screens.

Theorem 3 (Population simultaneous certificate)**.** *For a prespecified family indexed by $k = 1, \dots, K$, let (W_i, D_i), with $W_i = (X_i, E_i)$, be independent draws from the declared prompt–outcome population and let $d_{ik} = d_k(W_i, D_i)$ be the prompt-level paired loss. The predictors are fixed before these draws, or are trained on data independent of them. Suppose pre-qualification constants A_k, B_k, possibly determined by the independent training data, satisfy $A_k \le B_k$ and $A_k \le d_k(w, y) \le B_k$ for every (w, y) in the declared support. With $\bar{d}_k = n^{-1} \sum_{i=1}^{n} d_{ik}$, define*

$$U_k^{\mathrm{pop}} = \bar{d}_k + (B_k - A_k) \sqrt{\frac{\log(K/\alpha)}{2n}}. \tag{9}$$

With probability at least $1 - \alpha$, every population mean $\mu_k = \mathbb{E}[d_k(W, D)]$ is at most U_k^{pop} simultaneously. Certifying diagnostic k only when $U_k^{\mathrm{pop}} < -m_k$ controls family-wise false certification at level α.

These concentration bounds require a bounded loss contrast; paired Brier loss always lies in $[-1, 1]$. Log loss needs a prespecified restriction $p \in [\epsilon, 1 - \epsilon]$ with $0 < \epsilon < 1/2$, or a different valid concentration argument. A narrower envelope must cover the declared population before sampling. Individualized ranges in Equation (7) instead certify a fixed cohort's conditional mean, without covering prompt-sampling variation. Simple random sampling without replacement permits a finite-population correction for a frame-level claim. Proofs, variants, and power planning are in Appendices C.2–C.5.

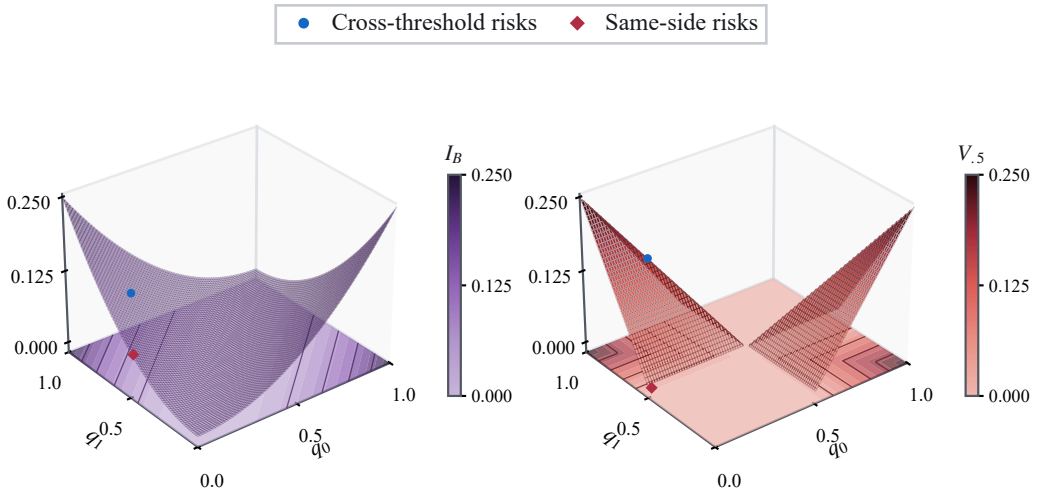
Prediction qualification alone leaves two questions: whether the outcome instrument measures semantic loss, and whether the predictions change decisions. Under nondifferential binary measurement error with $\kappa = \text{sensitivity} + \text{specificity} - 1$, observed diagnostic information is attenuated by κ^2. Differential error can create or reverse the apparent information. For a declared accept/reject policy, predictive information has decision value only where it changes the Bayes action at the cost-induced threshold. Derivations are in Appendices D.1–D.4. PGC returns AUTHORIZED only when prediction is CERTIFIED and both measurement and decision are QUALIFIED. All other combinations return WITHHOLD and identify the first unqualified layer. When a prelaunch criterion fails, PGC stops outcome collection. Policy use remains WITHHOLD, and unobserved qualification axes receive no formal status.

5 VALIDATION

Validation asks whether the three tests detect distinct failures and whether a diagnostic meets all three criteria in a text-to-image study. Controlled component deletions address the first question; probability-sampled qualification addresses the second. Prescribed-coupling controls and stress tests examine the mechanisms behind these results. Source and external studies use prompts as independent clusters; finite-frame qualification samples distinct prompt–transition cases. Future pools and annotations are nested observations (Appendix A.3).

5.1 ISOLATING THE ROLE OF EACH QUALIFICATION AXIS

We construct three worlds that each violate one claim while satisfying the other two, together with a valid-coupling control. At $n = 3,840$, full PGC rejects every failure-world trial and retains every



(a) Predictive information (b) Decision value

Figure 2: Predictive information need not yield decision value. Under balanced strata with conditional risks $q_e = \mathbb{P}(D = 1 \mid E = e)$, (a) $I_B = (q_1 - q_0)^2/4$ measures Brier information; (b) Bayes policy value at $\tau = .5$ uses $c_{\mathrm{FA}} = c_{\mathrm{FR}} = 1/2$. Blue risks $(.20, .80)$ yield $(I_B, V) = (.0900, .150)$; red risks $(.01, .49)$ yield $(.0576, 0)$. Floor projections show contours, and (b) raises only positive value. Equation (44) defines both exact surfaces.

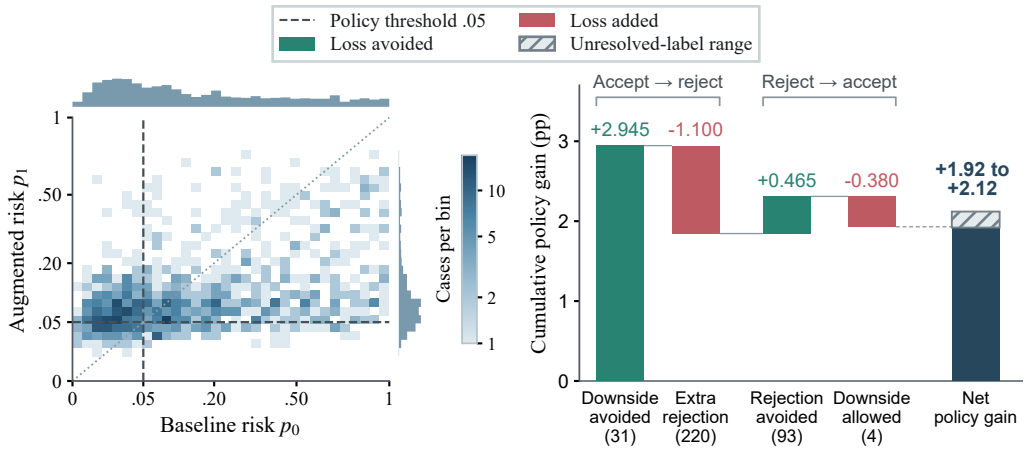
Table 2: Each axis protects a distinct part of the joint authorization claim. Green ✓ denotes 100% authorization and red ✗ denotes 0%, across 4,000 trials per world at $n = 3,840$. Full PGC and each single-axis deletion share fixed predictors and qualification samples. Symbols encode authorization, not correctness. Smaller-sample results appear in Table 12.

Controlled world	Full PGC	Without P	Without M	Without D
Valid coupling	✓	✓	✓	✓
Miscalibrated prediction, useful actions	✗	✓	✗	✗
Invalid measurement	✗	✗	✓	✗
Informative prediction, unchanged actions	✗	✗	✗	✓

valid-world trial. Deleting the corresponding axis authorizes every trial in its failure world (Table 2). The predictors remain fixed across these comparisons, isolating the contribution of each qualification test.

In these component deletions, all outcome-bearing gates are estimated from the independent qualification sample. A separate six-world benchmark fits predictors on development data and includes transport reversal, a semantic null, and an unattainable margin (Appendix E.2). At $n = 960$, its stable-coupling condition is authorized in 3,974/4,000 trials, while every failure world is rejected. Authorization for that condition rises from 3.18% at $n = 240$ to 59.38% at $n = 480$, exposing the measurement audit as the main sample bottleneck. Power, null, adaptive-search, and subgroup grids are in Appendices E.3–E.10.

Supporting analyses map continuous failure boundaries and test prescribed couplings on generated images. The mechanism grid isolates transport, measurement, threshold, and support effects (Appendix E.8); two diffusion positive controls recover their prescribed couplings (Appendices G.3 and G.3).



(a) Paired risk distribution

(b) Policy-gain decomposition

Figure 3: Risk changes and policy consequences in 1,000 qualification cases. (a) Baseline and augmented risks (p_0, p_1) with marginal histograms; axes use square-root spacing and color shows log-scaled bin counts. Dashed lines mark the .05 threshold; the dotted line marks equal risks. (b) The waterfall accumulates policy-gain contributions at the 19:1 cost ratio; parentheses denote case counts. Two of five unresolved labels determine the hatched final range of gains across label completions. Unchanged actions contribute zero.

5.2 Natural diagnostic qualification

Probability-sampled qualification. The natural study tests the complete protocol on a target-attribute diagnostic developed on disjoint data. It passes the feasibility screens and enters formal qualification. Its baseline predictor uses ordinary search-state features. The augmented predictor adds the target-attribute log-odds drop and a monotone risk basis. The diagnostic is produced by the CLIP–FlanT5-XL proxy, whereas a distinct Qwen2.5-VL evaluator defines the automatic endpoint. Because the baseline already contains the current-image attribute probability, the comparison tests the incremental value of candidate-relative deterioration. The predictors, endpoints, margins, and sampling plan were fixed before drawing qualification outcomes. The qualification family contains one tuple ($K = 1$). We drew 1,000 cases without replacement from a 1,227-case frame. Each case received three blinded paired ratings from a pool of 10 annotators. With error budgets $(\alpha_P, \alpha_M, \alpha_V) = (.020, .015, .015)$, all three axes clear the declared margins (Table 3). PGC then returns AUTHORIZED for this diagnostic and policy on the declared finite frame, using the human-semantic endpoint and 19:1 loss ratio.

Table 3: Natural three-axis qualification on a simple random sample of 1,000 cases from the 1,227-case frame. The three error budgets sum to .05.

Axis	Result	Prespecified criterion	Status
Prediction	ΔBrier -0.04947 UCB -0.01134	UCB $< -.001$	Pass
Measurement	min Se LCB 0.912 min Sp LCB 0.957	Se $\geq .80$; Sp $\geq .95$; $\Delta_{\text{err}} \leq .10$	Pass
Decision	sample gain $[0.0192, 0.0212]$ LCB 0.01108	LCB $> .005$	Pass

Five unresolved human majorities are propagated by worst-case bounds. The largest simultaneous between-stratum error-gap bound is 0.088. Together the three passes return AUTHORIZED for the declared frame and policy.

Figure 3 connects the changed predictions to their policy consequences. Avoided downside contributes more than the loss from additional rejection; restored acceptance supplies a smaller net gain. The augmented policy accepts 20.7% of candidate replacements while reducing sample semantic loss by 1.92–2.12 percentage points relative to $\mathcal{M}_0$. This range propagates the five unresolved labels; the formal lower bound is reported in Table 3. Table 14 reports absolute losses and error counts, including an always-retain reference. The paired sample gain over retention is .00035–.00135; the .005-margin certificate concerns the registered $\mathcal{M}_0$-to-$\mathcal{M}_1$ comparison. Human-endpoint prediction, source-wise comparisons, influence analyses and cost sensitivity are reported in Appendix F.

Repeatability without authority. In contrast, Ornstein–Uhlenbeck (OU) energy measures proxy-reward variation under local latent perturbations. Across 57 prompts it is repeatable but provides no evidence of improved grouped out-of-fold prediction (Figure 8). Its fixed predictor pair has insufficient attainable improvement and induces no action change. Appendix G.4 gives the attainable-support and action-feasibility analysis.

6 Discussion and Conclusion

PGC connects a diagnostic's predictive contribution to the semantic value of the policy it informs. Joint tests of prediction, measurement, and decision value establish this connection with false-authorization control.

The experiments show why the distinction matters. Component deletions expose failures missed by the remaining tests, while the target-attribute diagnostic meets all three criteria for the registered finite-frame $\mathcal{M}_0$-to-$\mathcal{M}_1$ comparison. Exact feasibility checks also rule out unattainable gains and unchanged actions before labels are collected. PGC thus identifies which comparisons merit outcome collection and which warrant a change in search policy.

AI Use Statement

The authors led the method development, experimental design, and interpretation of results. OpenAI Codex assisted with experimental design and planning, repository maintenance, debugging, code refactoring, selected implementation tasks, and manuscript drafting and language editing. The authors reviewed AI-assisted contributions and retain full responsibility for the methods, experiments, analyses, and submitted text.

Ethics Statement

Ten volunteers participated without pay or supervisory relationships. They consented to scientific and publication use, remained blind to the hypothesis, and were recorded only by pseudonymous identifiers. The activity followed the applicable institutional policy for minimal-risk, non-identifying research, under which formal review was not required. Appendix I.2 gives details.

References

Anastasios N. Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal risk control. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/f3549ef9b5ff520a7e41ff3cc306ab2b-Abstract-Conference.html.

Anastasios N. Angelopoulos, Stephen Bates, Emmanuel J. Candès, Michael I. Jordan, and Lihua Lei. Learn then test: Calibrating predictive algorithms to achieve risk control. *The Annals of Applied Statistics*, 19(2):1641–1662, 2025. doi: 10.1214/24-AOAS1998.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, et al. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025a. URL https://arxiv.org/abs/2511.21631.

Shuai Bai, Keqin Chen, Xuejing Liu, et al. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025b. URL https://arxiv.org/abs/2502.13923.

Arpit Bansal, Hong-Min Chu, Avi Schwarzschild, Soumyadip Sengupta, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Universal guidance for diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 843–852, 2023. URL https://openaccess.thecvf.com/content/CVPR2023W/GCV/html/Bansal_Universal_Guidance_for_Diffusion_Models_CVPRW_2023_paper.html.

Rémi Bardenet and Odalric-Ambrym Maillard. Concentration inequalities for sampling without replacement. *Bernoulli*, 21(3):1361–1385, 2015. doi: 10.3150/14-BEJ605.

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

Andrew M. Bean, Ryan Othniel Kearns, Angelika Romanou, Franziska Sofia Hafner, Harry Mayne, et al. Measuring what matters: Construct validity in large language model benchmarks. In *Advances in Neural Information Processing Systems*, volume 38, pp. 19868–19949, 2025. doi: 10.52202/085713-0590.

Roger L. Berger and Jason C. Hsu. Bioequivalence trials, intersection–union tests and equivalence confidence sets. *Statistical Science*, 11(4):283–302, 1996. doi: 10.1214/ss/1032280304.

Su Lin Blodgett, Gilsinia Lopez, Alexandra Olteanu, Robert Sim, and Hanna Wallach. Stereotyping Norwegian salmon: An inventory of pitfalls in fairness benchmark datasets. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pp. 1004–1015, 2021. doi: 10.18653/v1/2021.acl-long.81.

Avrim Blum and Moritz Hardt. The ladder: A reliable leaderboard for machine learning competitions. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 1006–1014, 2015. URL https://proceedings.mlr.press/v37/blum15.html.

Glenn W. Brier. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1–3, 1950. doi: 10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2.

William Caban. Measurement without validity: The compounding reliability problem in agentic AI evaluation. *arXiv preprint arXiv:2608.00794*, 2026. doi: 10.48550/arXiv.2608.00794.

C. J. Clopper and E. S. Pearson. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, 26(4):404–413, 1934. doi: 10.1093/biomet/26.4.404.

Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Aaron Roth. The reusable holdout: Preserving validity in adaptive data analysis. *Science*, 349(6248):636–638, 2015. doi: 10.1126/science.aaa9375.

Joseph L. Fleiss. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382, 1971. doi: 10.1037/h0031619.

Leo Gao, John Schulman, and Jacob Hilton. Scaling laws for reward model overoptimization. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 10835–10866, 2023. URL https://proceedings.mlr.press/v202/gao23h.html.

Yonatan Geifman and Ran El-Yaniv. SelectiveNet: A deep neural network with an integrated reject option. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2151–2159, 2019. URL https://proceedings.mlr.press/v97/geifman19a.html.

Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. GenEval: An object-focused framework for evaluating text-to-image alignment. In *Advances in Neural Information Processing Systems*, volume 36, pp. 52132–52152, 2023. doi: 10.52202/075280-2270.

Tilmann Gneiting and Adrian E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007. doi: 10.1198/016214506000001437.

Rogerio Guimaraes and Pietro Perona. Inference-time scaling of diffusion models via progressive seed pruning. *arXiv preprint arXiv:2607.21591*, 2026. doi: 10.48550/arXiv.2607.21591.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963. doi: 10.1080/01621459.1963.10500830.

Xiwei Hu, Rui Wang, Yixiao Fang, Bin Fu, Pei Cheng, and Gang Yu. ELLA: Equip diffusion models with LLM for enhanced semantic alignment. *arXiv preprint arXiv:2403.05135*, 2024. URL https://arxiv.org/abs/2403.05135.

Yushi Hu, Benlin Liu, Jungo Kasai, Yizhong Wang, Mari Ostendorf, Ranjay Krishna, and Noah A. Smith. TIFA: Accurate and interpretable text-to-image faithfulness evaluation with question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 20349–20360, 2023. doi: 10.1109/ICCV51070.2023.01866.

Audrey Huang, Adam Block, Qinghua Liu, Nan Jiang, Akshay Krishnamurthy, and Dylan J. Foster. Is best-of-N the best of them? coverage, scaling, and optimality in inference-time alignment. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 25075–25126, 2025. URL https://proceedings.mlr.press/v267/huang25c.html.

Amita Kamath, Kai-Wei Chang, Ranjay Krishna, Luke Zettlemoyer, Yushi Hu, and Marjan Ghazvininejad. GenEval 2: Addressing benchmark drift in text-to-image evaluation. *arXiv preprint arXiv:2512.16853*, 2025. doi: 10.48550/arXiv.2512.16853.

Hadi Khalaf, Claudio Mayrink Verdun, Alex Oesterling, Himabindu Lakkaraju, and Flavio du Pin Calmon. Inference-time reward hacking in large language models. In *Advances in Neural Information Processing Systems*, volume 38, pp. 68646–68686, 2025. doi: 10.52202/085713-2064.

Jeongsol Kim, Hongeun Kim, Jian Wang, and Jong Chul Ye. Gradient-free noise optimization for reward alignment in generative models. *arXiv preprint arXiv:2605.11347*, 2026. doi: 10.48550/arXiv.2605.11347.

Kwanyoung Kim and Byeongsu Sim. Reward sharpness-aware fine-tuning for diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 36051–36061, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Kim_Reward_Sharpness-Aware_Fine-Tuning_for_Diffusion_Models_CVPR_2026_paper.html.

Sunghwan Kim, Dongjin Kang, Taeyoon Kwon, Hyungjoo Chae, Dongha Lee, and Jinyoung Yeo. Rethinking reward model evaluation through the lens of reward overoptimization. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13252–13280, 2025. doi: 10.18653/v1/2025.acl-long.649.

Klaus Krippendorff. Computing Krippendorff's alpha-reliability. Technical report, Annenberg School for Communication, University of Pennsylvania, 2011. URL https://www.asc.upenn.edu/sites/default/files/2021-03/Computing%20Krippendorff%27s%20Alpha-Reliability.pdf.

Shuangqi Li, Hieu Le, Jingyi Xu, and Mathieu Salzmann. All seeds are not equal: Enhancing compositional text-to-image generation with reliable random seeds. *arXiv preprint arXiv:2411.18810*, 2024. doi: 10.48550/arXiv.2411.18810.

Zhiqiu Lin, Deepak Pathak, Baiqi Li, Jiayao Li, Xide Xia, Graham Neubig, Pengchuan Zhang, and Deva Ramanan. Evaluating text-to-visual generation with image-to-text generation. In *European Conference on Computer Vision*, 2024. URL https://www.ecva.net/papers/eccv_2024/papers_ECCV/html/1435_ECCV_2024_paper.php.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-Eval: NLG evaluation using GPT-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 2511–2522, 2023. doi: 10.18653/v1/2023.emnlp-main.153.

Andreas Maurer and Massimiliano Pontil. Empirical Bernstein bounds and sample variance penalization. In *Proceedings of the 22nd Annual Conference on Learning Theory*, pp. 115–124, 2009. URL https://www.cs.mcgill.ca/~colt2009/papers/012.pdf.

Alexander Pan, Kush Bhatia, and Jacob Steinhardt. The effects of reward misspecification: Mapping and mitigating misaligned models. In *International Conference on Learning Representations*, 2022. URL https://iclr.cc/virtual/2022/poster/6579.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: Improving latent diffusion models for high-resolution image synthesis. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=di52zR8xgf.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763, 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Candace Ross, Melissa Hall, Adriana Romero-Soriano, and Adina Williams. What makes a good metric? evaluating automatic metrics for text-to-image consistency. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=LFfktMPAci.

Frank L. Schmidt, Chockalingam Viswesvaran, and Deniz S. Ones. Reliability is not validity and validity is not reliability. *Personnel Psychology*, 53(4):901–912, 2000. doi: 10.1111/j.1744-6570.2000.tb02422.x.

R. J. Serfling. Probability inequalities for the sum in sampling without replacement. *The Annals of Statistics*, 2(1):39–48, 1974. doi: 10.1214/aos/1176342611.

Joar Skalse, Nikolaus H. R. Howe, Dmitrii Krasheninnikov, and David Krueger. Defining and characterizing reward gaming. In *Advances in Neural Information Processing Systems*, volume 35, pp. 9460–9471, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/3d719fee332caa23d5038b8a90e81796-Abstract-Conference.html.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling LLM test-time compute optimally can be more effective than scaling parameters for reasoning. In *International Conference on Learning Representations*, pp. 10131–10165, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/1b623663fd9b874366f3ce019fdfdd44-Abstract-Conference.html.

Bowen Song, Zecheng Zhang, Zhaoxu Luo, Jason Hu, Wei Yuan, Jing Jia, Zhengxu Tang, Guanyang Wang, and Liyue Shen. CCS: Controllable and constrained sampling with diffusion models via initial noise perturbation. In *Advances in Neural Information Processing Systems*, volume 38, pp. 78545–78581, 2025. doi: 10.52202/085713-2371.

Stability AI. Stable diffusion 3.5 medium model card, 2024. URL https://huggingface.co/stabilityai/stable-diffusion-3.5-medium.

Adarsh Subbaswamy, Peter Schulam, and Suchi Saria. Preventing failures due to dataset shift: Learning predictive models that transport. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pp. 3118–3127, 2019. URL https://proceedings.mlr.press/v89/subbaswamy19a.html.

Elie Tamer. Partial identification in econometrics. *Annual Review of Economics*, 2:167–195, 2010. doi: 10.1146/annurev.economics.050708.143401.

Zhiwei Tang, Jiangweizhi Peng, Jiasheng Tang, Mingyi Hong, Fan Wang, and Tsung-Hui Chang. Inference-time alignment of diffusion models with direct noise optimization. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 58905–58930, 2025. URL https://proceedings.mlr.press/v267/tang25h.html.

Bram Wallace, Akash Gokul, Stefano Ermon, and Nikhil Naik. End-to-end diffusion latent optimization improves classifier guidance. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7280–7290, 2023. URL https://openaccess.thecvf.com/content/ICCV2023/html/Wallace_End-to-End_Diffusion_Latent_Optimization_Improves_Classifier_Guidance_ICCV_2023_paper.html.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, and Zhifang Sui. Large language models are not fair evaluators. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pp. 9440–9450, 2024. doi: 10.18653/v1/2024.acl-long.511.

Bryan Wilder, Bistra Dilkina, and Milind Tambe. Melding the data-decisions pipeline: Decision-focused learning for combinatorial optimization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 1658–1665, 2019. doi: 10.1609/aaai.v33i01.33011658.

Zhuohao Yu, Zhiwei Steven Wu, and Adam Block. From curiosity to caution: Mitigating reward hacking for best-of-N with pessimism. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/0d11b4035d54e449320fc36b0f64ebb7-Abstract-Conference.html.

Kevin Zhai, Utsav Singh, Anirudh Thatipelli, Souradip Chakraborty, Anit Kumar Sahu, Furong Huang, Amrit Singh Bedi, and Mubarak Shah. MIRA: Towards mitigating reward hacking in inference-time alignment of T2I diffusion models. *arXiv preprint arXiv:2510.01549*, 2025. URL https://arxiv.org/abs/2510.01549.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 586–595, 2018. doi: 10.1109/CVPR.2018.00068.

Xiangcheng Zhang, Haowei Lin, Haotian Ye, James Zou, Jianzhu Ma, Yitao Liang, and Yilun Du. Inference-time scaling of diffusion models through classical search. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/41c9aa6c04924c29c6469572c5f6d01b-Abstract-Conference.html.

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 *Advances in Neural Information Processing Systems*, volume 36, pp. 46595–46623, 2023. doi: 10.52202/075280-2020.

Yuchen Zhu, Daniel Augusto De Souza, Zhengyan Shi, Mengyue Yang, Pasquale Minervini, Matt Kusner, and Alexander D'Amour. When can proxies improve the sample complexity of preference learning? In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 79790–79814, 2025. URL https://proceedings.mlr.press/v267/zhu25f.html.

A NOTATION, ESTIMANDS, AND STATUS SEMANTICS

We specify the probability space, prediction unit, estimands, and PGC status semantics used in Sections 3 and 4.

A.1 SEARCH-STATE PROBABILITY SPACE

Let the random prompt be $\mathsf{P} \sim \Pi$. Conditional on a realized prompt c, a generator receives an initial latent Z, sampler randomness Ξ, and sampler configuration Γ. Uppercase variables denote random quantities; the lowercase notation in Section 3 denotes their realizations. A generated image is

$$\mathsf{I} = G_\theta(c, Z; \Xi, \Gamma). \tag{10}$$

The observed candidate set A and all proxy rewards are measurable before the future-search randomness Ω is drawn. The fixed current winner is $Z_A^\star$, with ties resolved by an independent priority generated from the registered tie-breaking seed. Define the future winner $Z'^\star = \mathcal{S}(c, A, Z_A^\star; \Omega)$, where $\mathcal{S}$ is measurable in the realized state and future randomness. Let $\mathsf{I}_A^\star$ and $\mathsf{I}'^\star$ denote the decoded images corresponding to $Z_A^\star$ and $Z'^\star$ under the registered sampler, with realizations $x_A^\star$ and $x'^\star$. The replacement event is

$$C = \mathbf{1}\{Z'^\star \neq Z_A^\star\}. \tag{11}$$

Let O denote the recorded prompt, candidate identities, decoded current and future images, current and future proxy scores, selected winners, margins, locality features, diagnostic responses, and replacement event. It excludes $D^\star$ and $\widetilde{D}$. The observable model specifies the law of O but does not, by itself, specify the unknown semantic correctness rule. Within any fixed semantic world the rule may be deterministic in the prompt and image. Non-identification refers to distinct semantic worlds that induce the same law of O. The evaluator label $\widetilde{D}$ is generated after the selected images are fixed. This order distinguishes a predictor computed from proxy-side data from an outcome instrument used to ground it.

For Ornstein–Uhlenbeck (OU) diagnostics, O_J contains the implemented finite-J antithetic estimates and O_∞ contains the complete OU response around every selected point. The identification result in Appendix B.3 permits $O = O_\infty$ and does not rely on estimator noise.

A.2 OUTCOMES AND ESTIMANDS

For a binary semantic attribute a, define latent correctness $h_a^\star(c, x) \in \{0, 1\}$ and evaluator probability $\widetilde{h}_a(x) \in [0, 1]$. The latent and observed transition outcomes are

$$D^\star = C\mathbf{1}\{h_a^\star(c, \mathsf{I}_A^\star) = 1\}\mathbf{1}\{h_a^\star(c, \mathsf{I}'^\star) = 0\}, \tag{12}$$

$$\widetilde{D} = C\mathbf{1}\{\widetilde{h}_a(\mathsf{I}_A^\star) \geq .5\}\mathbf{1}\{\widetilde{h}_a(\mathsf{I}'^\star) < .5\}. \tag{13}$$

The source and external natural experiments predict prompt- and attribute-specific averages over registered future pools. Let X denote the ordinary prespecified covariates and E the candidate diagnostic. The primary estimand is the expected prompt-equal paired Brier difference

$$\Delta_{\text{Brier}} = \mathbb{E}\Big[(p_1(X, E) - \widetilde{D})^2 - (p_0(X) - \widetilde{D})^2\Big], \tag{14}$$

where negative values favor the augmented predictor. AUROC, AUPRC, calibration, risk strata, and replicate direction are prespecified secondary or stress metrics. They do not replace the paired-score estimand.

For empirical qualification, the intended construct $D^\star$ is operationalized before sampling as a human-semantic endpoint $D^H \in \{0, 1\}$. It defines the semantic endpoint of the joint authorization claim. The prediction axis evaluates the fixed predictor pair against the registered automatic endpoint $\widetilde{D}$; the measurement axis qualifies $\widetilde{D}$ relative to D^H; and the decision axis evaluates policy loss directly against D^H. Extending this operational claim to latent truth $D^\star$ requires an additional measurement model.

Table 4: Inferential units by study. Nested pools and annotations refine the corresponding unit but do not create additional independent units.

Study or object	Inferential unit	Nested observations / aggregation
Source and external studies	Prompt	State–attribute units and future pools; **equal prompt weight**
Matched human audit	Matched case	42 cases; 10 judgments **per case**
Finite-frame qualification	Registered opportunity	1,227 prompt transitions; 3 ratings per **sampled case**
$D^\star, D^H, \widetilde{D}$	Distinct outcomes	Intended construct, **human-semantic** endpoint, and **automatic** endpoint

A.3 Statistical units

Table 4 lists each study's inferential unit and nested observations. In the source and external studies, prompt-equal weighting first averages units within prompt and then averages prompts. For a unit with future event fraction $\bar{y}$, mean Bernoulli Brier loss is $\bar{y}(1-p)^2 + (1-\bar{y})p^2 = (p-\bar{y})^2 + \bar{y}(1-\bar{y})$. The final term cancels in the paired model contrast, but remains in each model's absolute loss. Prompt-cluster bootstrap resamples prompts and carries every nested unit with the sampled cluster. In the finite-frame qualification, every case has a distinct prompt identifier and defines one fixed current-to-candidate transition. The simple random sample is therefore taken over cases, and rater responses are nested measurements rather than additional frame units. Replicate-half analyses evaluate the same fixed prediction pair against two disjoint two-pool targets in place of the four-pool target.

A.4 Status semantics

The prespecified certification rule assigns the predictive axis: CERTIFIED when the simultaneous upper bound clears the declared margin, ADVERSE when a registered harm rule is supported, and INCONCLUSIVE otherwise. If an attainability or prospective-power criterion stops the independent evaluation before outcome collection, the formal predictive status is UNASSIGNED.

The measurement axis is QUALIFIED when the registered outcome audit supports the declared human-semantic endpoint under its sampling model and NOT IDENTIFIED otherwise. The decision axis is QUALIFIED, NOT QUALIFIED, or NOT EVALUATED for a prespecified cost, comparator, budget, and policy class. Policy use is AUTHORIZED only for CERTIFIED$\times$QUALIFIED$\times$QUALIFIED. Every other combination returns WITHHOLD. These labels identify which evidentiary statement is supported.

B Identification and Grounding Theory

The proofs for Section 3 distinguish proper-score information, search exposure, semantic susceptibility, and cross-setting predictive value.

B.1 Incremental Brier information

For an information set $\mathcal{I}$, define the Bayes Brier risk

$$\mathcal{R}^\star(\mathcal{I}) = \inf_{f\ \mathcal{I}\text{-measurable}} \mathbb{E}[(D-f)^2]. \tag{15}$$

Lemma 1 (Incremental proper-score information)**.** *For binary D and nested information sets $X \subseteq (X,E)$,*

$$\mathcal{R}^\star(X) - \mathcal{R}^\star(X,E) = \mathbb{E}[(\eta_{X,E} - \eta_X)^2]. \tag{16}$$

The diagnostic has positive Bayes Brier value exactly when $\mathbb{P}(\eta_{X,E} \neq \eta_X) > 0$.

Proof. Conditional expectation is the L^2 projection, so the Bayes predictors are η_X and $\eta_{X,E}$. Decompose

$$D - \eta_X = (D - \eta_{X,E}) + (\eta_{X,E} - \eta_X). \tag{17}$$

The cross term has expectation zero because

$$\mathbb{E}[(D - \eta_{X,E})(\eta_{X,E} - \eta_X)] = \mathbb{E}[(\eta_{X,E} - \eta_X)\mathbb{E}[D - \eta_{X,E} \mid X, E]] = 0. \tag{18}$$

Squaring Equation (17), taking expectations, and using Equation (18) gives Equation (16). Its right-hand side is zero exactly when the conditional risks agree almost surely. □

B.2 Exposure–susceptibility factorization

Proposition 3 (Search-risk factorization)**.** *Let C indicate replacement, let T indicate semantic loss conditional on replacement, and set $D = CT$. Define*

$$\pi_{X,E} = \mathbb{P}(C = 1 \mid X, E), \qquad q_{X,E} = \mathbb{P}(T = 1 \mid C = 1, X, E), \tag{19}$$

and define π_X, q_X analogously, with $q = 0$ wherever the corresponding replacement probability is zero. Then

$$\eta_{X,E} = \pi_{X,E} q_{X,E}, \qquad \eta_X = \pi_X q_X, \tag{20}$$

and

$$\eta_{X,E} - \eta_X = q_X(\pi_{X,E} - \pi_X) + \pi_{X,E}(q_{X,E} - q_X). \tag{21}$$

If $\pi_{X,E} = \pi_X$ almost surely, the diagnostic's Brier value reduces to $\mathbb{E}[\pi_X^2(q_{X,E} - q_X)^2]$.

Proof. Iterated conditioning yields

$$\mathbb{P}(D = 1 \mid X, E) = \mathbb{P}(C = 1 \mid X, E)\mathbb{P}(T = 1 \mid C = 1, X, E), \tag{22}$$

Equation (22) gives the first equality in Equation (20). Conditioning only on X gives the second. Adding and subtracting $\pi_{X,E} q_X$ gives Equation (21). Substitution into Equation (16) proves the Brier expression and its restricted form. □

B.3 Proof of selection-conditioned non-identifiability

Proof of Theorem 1. It is enough to exhibit one observable law and two members of the stated model class. Let the law of O assign positive probability to a replacement event $\mathcal{E} = \{C = 1\}$. Construct an admissible observable submodel in which every generated candidate has a unique identity and every selected prompt–image pair is distinct. For example, use a generator that is injective on the finite set of prompt, seed, and candidate-index tuples with positive probability. No selected image is reused as a current or future winner in another transition. This restriction leaves the complete proxy trajectory, images, scores, margins, and diagnostic values in O.

Define two deterministic semantic correctness rules on the same prompt–image domain. In world 0, set the current and future winner correct on $\mathcal{E}$. This gives $D_0^\star = 0$. In world 1, set the current winner correct and the future winner incorrect on $\mathcal{E}$, which gives $D_1^\star = 1$. Assign identical labels in the two worlds outside $\mathcal{E}$, and extend each rule arbitrarily to prompt–image pairs that are never selected. Distinct selected prompt–image pairs ensure that each pair receives one label within each world, so both $h_{a,0}^\star$ and $h_{a,1}^\star$ are coherent functions rather than transition-specific labels.

The semantic rule is not an input to the proxy generator or selection mechanism, so replacing $h_{a,0}^\star$ by $h_{a,1}^\star$ leaves the marginal law of O unchanged. On $\mathcal{E}$, however, the two induced downside laws are opposite. Therefore the mapping from the observable law to the downside law is not unique over the assumption-minimal model class. □

The construction also shows why perfect estimator reliability cannot identify semantic validity. Take a nondegenerate, perfectly observed E and choose $D \perp E \mid X$. Then $\eta_{X,E} = \eta_X$ and Lemma 1 with Equation (16) gives zero incremental value.

B.4 STABLE POSITIVE COUPLING

Proof of Proposition 1. Conditioning Equation (5) on X and using the zero-mean condition gives $\eta_X = m(X)$. Lemma 1 then yields

$$\mathcal{R}^\star(X) - \mathcal{R}^\star(X, E) = \mathbb{E}[(\beta g(X, E))^2] = \beta^2 \mathbb{E}[g(X, E)^2] > 0. \tag{23}$$

Under a new prompt law $\mathbb{Q}$, the same calculation applies only when the conditional relation and the centering condition $\mathbb{E}_{\mathbb{Q}}[g(X, E) \mid X] = 0$ both hold under $\mathbb{Q}$. The condition $\mathbb{E}_{\mathbb{Q}}[g(X, E)^2] > 0$ then makes the value strict. Without residual variation beyond X, marginal variation of g alone does not imply incremental value. The external experiment evaluates the fixed predictor without refitting and does not invoke this Bayes transport conclusion. □

B.5 NO UNIFORM PROXY-ONLY HEDGE

Let $\psi(O) \in \{0, 1\}$ decide whether to accept a replacement using only observable proxy-side information, and let ordinary search be $\psi_0(O) = 1$ on $\mathcal{E} = \{C = 1\}$. For $\lambda > 0$, use the loss and replacement-conditioned risk in Equation (4).

Proof of Corollary 1. Choose the unique-identity observable submodel in Appendix B.3 and an observable subset $B \subseteq \mathcal{E}$ with positive conditional probability. Define $\psi = 0$ on B and $\psi = \psi_0 = 1$ elsewhere on $\mathcal{E}$. In world 0, label both current and future winners correct on B, so $D = 0$. In world 1, label the current winner correct and future winner incorrect on B, so $D = 1$. Assign the same semantic labels outside B. Both assignments are coherent correctness functions and leave the observable law unchanged. The policies agree on $\mathcal{E} \setminus B$. Equation (4) therefore gives

$$\mathfrak{R}_{\lambda,\mathbb{P}_0}(\psi) - \mathfrak{R}_{\lambda,\mathbb{P}_0}(\psi_0) = \lambda \mathbb{P}(B \mid C = 1) > 0, \qquad \mathfrak{R}_{\lambda,\mathbb{P}_1}(\psi) - \mathfrak{R}_{\lambda,\mathbb{P}_1}(\psi_0) = -\mathbb{P}(B \mid C = 1) < 0. \tag{24}$$

Thus the same proxy-only hedge is worse in one compatible world and better in the other. □

B.6 EXACT EXOGENOUS-NUISANCE POSITIVE WORLD

Let the fixed current candidate have semantic score and nuisance (h_0, u_0). Draw one independent future candidate (H, U) with $U \perp H$, select by $R_\lambda = H + \lambda U$, and define downside

$$D = \mathbf{1}\{H + \lambda U > h_0 + \lambda u_0,\ H < h_0\}. \tag{25}$$

Theorem 4 (Exogenous nuisance identifies single-future downside)**.** *If $\lambda > 0$ and F_U is continuous and strictly increasing, then conditional on (h_0, u_0),*

$$q_\lambda(h_0, u_0) = \int_{h < h_0} \left[1 - F_U\left(u_0 + \frac{h_0 - h}{\lambda}\right)\right] dF_H(h). \tag{26}$$

When $F_H((-\infty, h_0)) > 0$, this probability is strictly decreasing in u_0 and increasing in diagnostic $E = -u_0$.

Proof. Conditional on $H = h < h_0$, replacement is equivalent to $U > u_0 + (h_0 - h)/\lambda$. Independence supplies the tail probability inside Equation (26). If $u_0' > u_0$, strict monotonicity of F_U makes every integrand at u_0' smaller. Positive mass below h_0 makes the integral inequality strict. □

If X excludes u_0 and the conditional variance of $q_\lambda(h_0, u_0)$ is positive, Lemma 1 gives strictly positive diagnostic Brier value. The real-image bridge uses the theorem's single-future-candidate case as its primary construct-controlled target.

C PREDICTIVE QUALIFICATION AND FINITE-SAMPLE GUARANTEES

We formalize the claim fixed before outcome access, prove the attainability and certificate results from Section 4, and give the prospective power calculation and qualification sequence.

C.1 PRESPECIFIED CLAIM SCHEMA

Before held-out outcome access, a result-bearing PGC study fixes:

1. prompt population, eligibility rule, and independent sampling unit;
2. baseline predictor p_0, augmented predictor p_1, diagnostic family, preprocessing, folds, and tie rules;
3. outcome definition, aggregation, proper score, practical margin, and multiplicity allocation;
4. attainability-compatible planning alternative, prospective sample-size range, and power gate;
5. measurement instrument, sampling design, and qualification rule;
6. decision costs, comparator, action space, budget, and authorization threshold.

Together these elements define the qualification target and sampling law; evaluation then uses outcomes from the declared qualification sample.

C.2 EXACT ATTAINABLE INTERVAL

Proof of Proposition 2. For every unit and binary outcome, $a_i \le d_i(Y_i) \le b_i$ pointwise, so every expected weighted loss lies in $[\sum_i w_i a_i, \sum_i w_i b_i]$. Conversely, at unit i randomize between the labels attaining a_i and b_i. Varying the mixture probability spans the complete interval $[a_i, b_i]$. The weighted Minkowski sum of these intervals is $[\sum_i w_i a_i, \sum_i w_i b_i]$, which is Equation (8), so every point is attainable. Comparing $-m$ with the lower endpoint proves the claim. □

For Brier loss,

$$d_i(0) = p_{1,i}^2 - p_{0,i}^2, \qquad d_i(1) = (1 - p_{1,i})^2 - (1 - p_{0,i})^2, \tag{27}$$

and $b_i - a_i = 2|p_{1,i} - p_{0,i}|$. When a prompt contains several registered units, each prompt endpoint is the same weighted average of unit endpoints used by the primary estimand.

C.3 POPULATION AND FIXED-COHORT CERTIFICATES

Proof of Theorem 3. Condition first on any independent training data used to construct the fixed predictors. For fixed k, the qualification losses are independent draws from the declared prompt population and lie in the deterministic interval $[A_k, B_k]$. If $A_k = B_k$, the claim is immediate. Otherwise Hoeffding's inequality (Hoeffding, 1963) gives

$$\mathbb{P}\{\mu_k - \bar{d}_k \ge t\} \le \exp\left(-\frac{2nt^2}{(B_k - A_k)^2}\right). \tag{28}$$

Setting the right-hand side of Equation (28) to α/K yields Equation (9). A union bound gives simultaneous coverage. On this event, $U_k^{\mathrm{pop}} < -m_k$ implies $\mu_k < -m_k$, so a diagnostic whose population improvement does not exceed the margin cannot be certified. Integrating over the independent training data preserves the bound. □

The individualized endpoint widths in Equation (7) support a different, conditional statement. Let $\mu_{ik} = \mathbb{E}[d_{ik} \mid \mathcal{T}, W_{1:n}]$, where $\mathcal{T}$ denotes the training data and $W_{1:n}$ the realized qualification units.

Proposition 4 (Fixed-cohort range-aware certificate)**.** *Conditional on $(\mathcal{T}, W_{1:n})$, suppose the outcomes are independent and $d_{ik} \in [a_{ik}, b_{ik}]$. Define*

$$U_k^{\mathrm{fix}} = \bar{d}_k + \sqrt{\frac{\log(K/\alpha)}{2n^2} \sum_{i=1}^{n} (b_{ik} - a_{ik})^2}. \tag{29}$$

With conditional probability at least $1 - \alpha$, every realized-cohort mean $n^{-1} \sum_i \mu_{ik}$ is at most U_k^{fix} simultaneously.

Proof. Conditional on $(\mathcal{T}, W_{1:n})$, the intervals $[a_{ik}, b_{ik}]$ are fixed. Hoeffding's inequality for independent, non-identically bounded variables gives

$$\mathbb{P}\left\{ n^{-1} \sum_i \mu_{ik} - \bar{d}_k \geq t \mid \mathcal{T}, W_{1:n} \right\} \leq \exp\left(-\frac{2n^2 t^2}{\sum_i (b_{ik} - a_{ik})^2} \right). \tag{30}$$

If all widths vanish, the claim is immediate. Otherwise, substitution of α/K followed by a union bound proves it. □

Proposition 4 conditions on which prompts were sampled. Its individualized widths quantify outcome noise within that cohort. They do not control variation from sampling prompts from Π.

Proposition 5 (Finite-frame prediction certificate)**.** *Let a fixed frame contain N registered frame-unit losses $d_{jk} \in [A_k, B_k]$ for each registered tuple k. Draw $1 \leq n \leq N$ frame units by simple random sampling without replacement and write $\bar{d}_{k,n}$ for the sample mean and $\mu_{k,N} = N^{-1} \sum_{j=1}^{N} d_{jk}$ for the frame mean. Define*

$$U_k^{\mathrm{ff}} = \bar{d}_{k,n} + (B_k - A_k) \sqrt{\left(1 - \frac{n-1}{N}\right) \frac{\log(K/\alpha)}{2n}}. \tag{31}$$

With probability at least $1 - \alpha$, every $\mu_{k,N}$ is at most U_k^{ff} simultaneously. Requiring $U_k^{\mathrm{ff}} < -m_k$ controls false certification of the declared frame mean at level α.

Proof. For $n < N$ and $A_k < B_k$, apply the without-replacement Hoeffding–Serfling inequality to $-d_{jk}$ (Serfling, 1974); the second inequality of Theorem 2.4 in Bardenet & Maillard (2015) gives this same bound. It bounds $\mathbb{P}\{\mu_{k,N} - \bar{d}_{k,n} \geq t\}$ by $\exp\{-2nt^2/[(1 - (n-1)/N)(B_k - A_k)^2]\}$. Substitution of α/K and a union bound give Equation (31) and the stated false-certification control. When $n = N$ or $A_k = B_k$, the sample mean equals the frame mean, so coverage is deterministic. □

C.4 Certificate variants

Table 5 records the assumptions and information used by each predictive bound. Measurement and decision guarantees concern different objects and remain in Appendix D.

Table 5: Predictive bounds and their sampling scope. All certify the declared mean proper-loss contrast under the fixed family and error allocation.

Bound	Loss information	Sampling scope
Population Hoeffding (Hoeffding, 1963)	Uniform $[A_k, B_k]$ over the declared support	Independent population draws
Finite-frame Serfling (Serfling, 1974)	Fixed frame; uniform $[A_k, B_k]$	Random sampling **without replacement**
Fixed-cohort Hoeffding (Hoeffding, 1963)	Realized unit-wise $[a_{ik}, b_{ik}]$	**Conditional** on units and outputs
Empirical Bernstein (Maurer & Pontil, 2009)	$[-1, 1]$ support and sample variance	I.i.d. population draws; $n > 1$

For paired Brier loss, the generic population certificate uses $d_{ik} \in [-1, 1]$ and radius $\sqrt{2\log(K/\alpha)/n}$. A narrower predictor-specific range is valid only if it is a deterministic envelope over the declared population conditional on independent training data. A variance-adaptive population alternative uses sample variance

$$\widehat{V}_k = (n-1)^{-1} \sum_i (d_{ik} - \bar{d}_k)^2 \tag{32}$$

and

$$U_k^{\mathrm{EB}} = \bar{d}_k + \sqrt{\frac{2\widehat{V}_k \log(2K/\alpha)}{n}} + \frac{14\log(2K/\alpha)}{3(n-1)}. \tag{33}$$

Theorem 5 (Variance-adaptive simultaneous certificate)**.** *If $n > 1$ and, for each k, the prompt-level losses are i.i.d. in $[-1, 1]$ and $\mu_k = \mathbb{E}[\bar{d}_k]$, all μ_k are at most the corresponding U_k^{EB} in Equation* (33) *simultaneously with probability at least $1 - \alpha$.*

Proof. Transform $Y_{ik} = (d_{ik} + 1)/2 \in [0, 1]$. The empirical Bernstein bound of Maurer & Pontil (2009, Theorem 4) at level α/K gives

$$\mathbb{E}Y_{1k} \le \bar{Y}_k + \sqrt{\frac{2\widehat{V}(Y_k)\log(2K/\alpha)}{n}} + \frac{7\log(2K/\alpha)}{3(n-1)}. \tag{34}$$

Since $\widehat{V}(Y_k) = \widehat{V}_k/4$, multiplying Equation (34) by two and subtracting one gives Equation (33). A union bound proves simultaneous coverage. □

C.5 Prospective power procedure

Power is projected only after the attainability check. The calculation is a plug-in planning analysis conditional on a fully specified empirical prompt-level distribution. Its scope is this planning law rather than all binary outcome laws with the same mean contrast. The planned statistic must match the declared estimand: a population qualification uses the support-uniform radius in Equation (9), whereas a fixed-cohort qualification may use Equation (29). The source development calculation below is the latter and did not assign a population certification status. For each diagnostic, let ℓ_j be the development prompt's prompt-equal paired Brier contrast and let $r_j = b_j - a_j$ be its outcome-blind range width. The prespecified planning distribution is uniform over the empirical pairs $\{(\ell_j - \bar{\ell}, r_j)\}_{j=1}^{57}$. For registered alternative $-\delta$, replicate $b = 1, \ldots, B$ at size n samples prompt indices $J_1, \ldots, J_n$ independently with replacement and sets

$$\bar{d}^{(b)} = -\delta + \frac{1}{n}\sum_{i=1}^{n}(\ell_{J_i} - \bar{\ell}), \qquad R^{(b)} = \sqrt{\frac{\log(K/\alpha)}{2n^2}\sum_{i=1}^{n} r_{J_i}^2}. \tag{35}$$

The replicate certifies exactly when the two quantities in Equation (35) satisfy $\bar{d}^{(b)} + R^{(b)} < -m$. The procedure then:

1. draws 10,000 prompt-resampling replicates at each planned n;
2. shifts only the mean contrast to the registered alternative, retaining the centered empirical prompt heterogeneity and sampled range widths;
3. applies the prespecified family-wise fixed-cohort rule;
4. records the fraction of simulations clearing the practical margin;
5. separately projects event-count feasibility by resampling the empirical number of units per prompt and drawing binomial events at the prespecified development prevalence.

"Power" below denotes the clearance frequency of this resampled statistic, not power under a prompt-specific Bernoulli outcome model. Item 5 separately models event-count feasibility. The source natural-family gate fixes four diagnostics, $\alpha = .05$, a 0.5% relative Brier margin, a 1% relative planning effect, prompt counts $80, 100, 120$, and 10,000 simulations. An independently seeded broader-budget sensitivity evaluates $n \in \{60, 120, 240, 480, 960\}$ without altering the prespecified launch decision. A separate 17-point visualization grid spans 60–280 prompts, uses 4,000 trials per cell, and locates the first grid crossings above 80% at 180 and 220 prompts.

C.6 Qualification procedure

The inputs are the prespecified prompt law, prediction unit, predictor pair, diagnostic family, practical margins, family-wise error allocation, planning alternative, power threshold, outcome instrument, and policy specification. PGC first computes the exact attainable interval and action geometry from the fixed predictions. Attainable contrasts then enter the registered power and event calculations. A family that clears these pre-outcome checks is evaluated once on independent grouped outcomes, without changing its predictors, features, folds, margins, or error allocation. Paired losses, the outcome audit, and the policy comparison assign the three statuses under the rules in Appendix A.4.

D MEASUREMENT AND DECISION QUALIFICATION

We derive the measurement and policy results summarized in Section 4.

D.1 NONDIFFERENTIAL MEASUREMENT ATTENUATION

Let $D^\star$ be the intended semantic event and $\widetilde{D}$ the evaluator label. Assume

$$\mathbb{P}(\widetilde{D} = 1 \mid D^\star = 1, X, E) = s, \qquad \mathbb{P}(\widetilde{D} = 0 \mid D^\star = 0, X, E) = t, \tag{36}$$

where sensitivity s and specificity t are constant across (X, E), and set $\kappa = s + t - 1$.

Theorem 6 (Measurement attenuation)**.** *For either $\mathcal{I} \in \{X, (X, E)\}$,*

$$\widetilde{\eta}_{\mathcal{I}} = (1 - t) + \kappa \eta^\star_{\mathcal{I}}. \tag{37}$$

Consequently, the Brier information carried by E about $\widetilde{D}$ is κ^2 times its Brier information about $D^\star$.

Proof. The law of total probability gives $\mathbb{P}(\widetilde{D} = 1 \mid \mathcal{I}) = s\eta^\star_{\mathcal{I}} + (1 - t)(1 - \eta^\star_{\mathcal{I}})$. Collecting terms yields Equation (37). Subtracting the two information sets multiplies the semantic risk contrast by κ. Equation (16) squares this contrast, giving the κ^2 information ratio. □

In empirical qualification, D^H replaces $D^\star$ as the registered semantic endpoint. Equations (36)–(37) then describe evaluator error relative to that operational endpoint.

D.2 DIFFERENTIAL-ERROR REVERSAL

Proposition 6 (Differential error can create or reverse grounding)**.** *If sensitivity or specificity depends on E after conditioning on $D^\star$, the observed diagnostic contrast need not have the semantic sign.*

Proof. Take semantic risks 0.2 and 0.4 in strata $E = 0$ and $E = 1$. Let sensitivity be 0.9 in both strata and specificity be 0.55 and 0.98, respectively. The observed positive rates are $0.9(0.2) + 0.45(0.8) = 0.54$ and $0.9(0.4) + 0.02(0.6) = 0.372$. The semantic contrast is $+0.2$, while the observed contrast is -0.168. □

D.3 FINITE-SAMPLE MEASUREMENT CERTIFICATE

Let $k = 1, \ldots, K$ index a fixed family of diagnostic–outcome–policy tuples, and let $G \in \{1, \ldots, G_0\}$ index policy-relevant strata. Within each registered stratum of an independent population-sampled qualification cohort, audited cases are selected independently with equal inclusion probability and receive paired evaluator and human-semantic labels $(\widetilde{D}, D^H)$. Different strata may use different audit fractions. Prediction and decision statistics use the population-sampled cohort rather than treating this stratified audit subsample as an unweighted population sample. Within tuple k and stratum g, write $s_{kg} = \mathbb{P}(\widetilde{D} = 1 \mid D^H = 1, G = g)$ and $t_{kg} = \mathbb{P}(\widetilde{D} = 0 \mid D^H = 0, G = g)$. Conditional on the observed numbers of human-positive and human-negative cases, form two-sided Clopper–Pearson (Clopper & Pearson, 1934) intervals $[s_{kg,L}, s_{kg,U}]$ and $[t_{kg,L}, t_{kg,U}]$, each with error $\alpha_M/(2G_0K)$. Thus each interval has the binomial sampling model required for exact Clopper–Pearson coverage. An empty human-label class receives the interval $[0, 1]$ and cannot establish its quality requirement. Sensitivity and specificity targets require positive human-positive and human-negative mass in each registered stratum.

For fixed s_0, t_0 and differential-error tolerances δ_s, δ_t, define the validity region for tuple k

$$\begin{aligned}\mathcal{M}_k = \{(s_k, t_k) : \min_g s_{kg} \geq s_0, \quad \min_g t_{kg} \geq t_0, \\ \max_g s_{kg} - \min_g s_{kg} \leq \delta_s, \quad \max_g t_{kg} - \min_g t_{kg} \leq \delta_t\}.\end{aligned} \tag{38}$$

Proposition 7 (Stratified measurement certificate)**.** *Qualify tuple k only when its confidence rectangle is contained in $\mathcal{M}_k$. The probability that any tuple outside its validity region qualifies is at most α_M.*

Proof. There are $2G_0K$ binomial parameters. Their conditional intervals cover simultaneously with probability at least $1-\alpha_M$ by a union bound. On this event, a confidence rectangle contained in $\mathcal{M}_k$ implies that the true parameter vector lies in $\mathcal{M}_k$. A false passage therefore requires a simultaneous-coverage failure. □

For a fixed finite frame, let N_g be the number of units in stratum g and let the four evaluator–human cell totals be $(K_{kg}^{\mathrm{TP}}, K_{kg}^{\mathrm{FN}}, K_{kg}^{\mathrm{TN}}, K_{kg}^{\mathrm{FP}})$. Under a simple random sample without replacement from the full frame, condition on the sampled stratum size n_g. Each observed cell count is then hypergeometric with population size N_g, draw size n_g, and the corresponding unknown cell total.

Proposition 8 (Finite-frame measurement certificate)**.** *Invert equal-tailed hypergeometric tests to obtain a two-sided interval for each of the $4G_0K$ cell totals at error $\alpha_M/(4G_0K)$. Convert each four-cell rectangle into sensitivity and specificity bounds by using the smallest possible numerator with the largest possible opposing error count, and vice versa. If the resulting confidence rectangle is contained in $\mathcal{M}_k$, the probability that any tuple outside its finite-frame validity region qualifies is at most α_M.*

Proof. Conditional on n_g, sampling within each stratum is simple random sampling without replacement, so inversion gives exact coverage for every cell total. A union bound over $4G_0K$ intervals gives simultaneous coverage at least $1-\alpha_M$. The sensitivity and specificity bounds are monotone functions of the four cell extrema. Rectangle containment then implies membership in $\mathcal{M}_k$; false qualification requires a coverage failure. □

If a cell rectangle permits a zero sensitivity or specificity denominator, use $[0,1]$ for that rate. A stratum with no sampled cases likewise retains its full feasible cell range.

Operationally, containment requires the smallest sensitivity and specificity lower bounds to clear s_0, t_0, while the corresponding upper-minus-lower ranges remain below δ_s, δ_t. Known across-stratum sampling fractions may weight a separate population aggregate, but do not enter these stratum-specific intervals. Clopper–Pearson applies to the i.i.d. design; hypergeometric inversion applies to finite-frame sampling without replacement. Unequal inclusion probabilities within a stratum require a prespecified design-valid simultaneous interval. A matched audit without a probability design cannot supply this certificate.

Remark 1 (Measurement qualification is partition-relative)**.** *The certificate controls only the registered partition G. Passing source-level bounds does not imply the same error limits on an unregistered refinement by diagnostic magnitude, predicted risk, or action change. A claim requiring such conditional validity must register the finer partition before outcome access or qualify prediction directly against D^H. Aggregate accuracy cannot substitute for either route because Proposition 6 applies within every coarse stratum.*

D.4 From prediction to decision value

Throughout this subsection, probabilities and expectations are taken over the fixed decision-opportunity population. For replacement policies, this is the eligible stratum $C = 1$.

Let $a = 1$ accept a replacement and $a = 0$ reject it, with

$$L(a, D) = c_{\mathrm{FA}} aD + c_{\mathrm{FR}}(1-a)(1-D), \qquad \tau = \frac{c_{\mathrm{FR}}}{c_{\mathrm{FA}} + c_{\mathrm{FR}}}. \tag{39}$$

The Bayes action under information $\mathcal{I}$ is $a_{\mathcal{I}} = \mathbf{1}\{\eta_{\mathcal{I}} < \tau\}$ with a fixed reject tie rule. For any action rule a, write $\mathcal{R}(a) = \mathbb{E}[L(a, D)]$.

Theorem 7 (Exact grounding-to-decision value)**.** *For $X \subseteq (X, E)$,*

$$\mathcal{R}(a_X) - \mathcal{R}(a_{X,E}) = (c_{\mathrm{FA}} + c_{\mathrm{FR}})\mathbb{E}[|\eta_{X,E} - \tau| \mathbf{1}\{a_X \neq a_{X,E}\}] \geq 0. \tag{40}$$

The value is strict exactly when the actions disagree away from the threshold on a set of positive probability.

Proof. Conditional on (X, E) with risk η, the losses of accepting and rejecting are $c_{\mathrm{FA}}\eta$ and $c_{\mathrm{FR}}(1-\eta)$. Their absolute difference is $(c_{\mathrm{FA}} + c_{\mathrm{FR}})|\eta - \tau|$ by Equation (39). The refined action chooses the smaller loss. The X-only action incurs the same loss when actions agree and exactly this difference when they disagree. Taking expectations proves Equation (40) and strictness statement. □

D.5 FINITE-SAMPLE DECISION CERTIFICATE

For tuple k, let (W_i, D_i^H) be independent draws from the declared decision-opportunity population. Let the baseline and diagnostic policies be fixed before these draws, or trained on independent data, and define their paired semantic-loss improvement

$$Y_{ik} = L(a_{0,k}(W_i), D_i^H) - L(a_{1,k}(W_i), D_i^H). \tag{41}$$

Suppose pre-qualification constants $A_k^V \le B_k^V$, possibly determined by independent training data, bound this contrast for every opportunity and binary endpoint in the declared support. Such bounds follow from the fixed action space and costs before qualification sampling.

Proposition 9 (Population decision certificate)**.** *For a fixed family of K tuples, define*

$$L_k^V = \bar{Y}_k - (B_k^V - A_k^V)\sqrt{\frac{\log(K/\alpha_V)}{2n}}. \tag{42}$$

With probability at least $1 - \alpha_V$, every mean decision improvement $\mu_k^V = \mathbb{E}[Y_k(W, D^H)]$ is at least L_k^V. Qualifying tuple k only when $L_k^V > m_k^V$ controls family-wise passage of false claims $\mu_k^V \le m_k^V$ at level α_V.

Proof. Conditioning on any independent training data, Hoeffding's lower-tail inequality on $[A_k^V, B_k^V]$ gives the stated bound when the interval has positive width. The zero-width case is immediate. Thus $\mathbb{P}(\mu_k^V < L_k^V) \le \alpha_V/K$ for each tuple. A union bound gives the simultaneous statement, and integrating over the training data preserves it. On this event, $L_k^V > m_k^V$ implies $\mu_k^V > m_k^V$. □

For a fixed finite frame, the two policies partition action-changing units into directions $\mathcal{H}_+$ (baseline accepts, augmented rejects) and $\mathcal{H}_-$ (baseline rejects, augmented accepts). Let N_h be a direction's frame size and K_h its unknown number of human-positive outcomes. Conditional on the sampled direction size n_h, the observed positive count is hypergeometric. Let $[K_{h,L}, K_{h,U}]$ be simultaneous intervals obtained by inverting these distributions. The lower frame-level policy gain is

$$L_{\mathrm{ff}}^V = \frac{1}{N}\left[\sum_{h\in\mathcal{H}_+}\{(c_{\mathrm{FA}} + c_{\mathrm{FR}})K_{h,L} - c_{\mathrm{FR}}N_h\} + \sum_{h\in\mathcal{H}_-}\{c_{\mathrm{FR}}N_h - (c_{\mathrm{FA}} + c_{\mathrm{FR}})K_{h,U}\}\right]. \tag{43}$$

Proposition 10 (Finite-frame decision certificate)**.** *Under simple random sampling without replacement, allocate α_V across the nonempty action-change directions and construct the exact intervals above. Then the finite-frame mean policy improvement is at least L_{ff}^V with probability at least $1-\alpha_V$. Requiring $L_{\mathrm{ff}}^V > m^V$ controls false decision qualification at level α_V.*

For a family of K tuples, allocate α_V/K to each tuple before splitting its budget across directions. Then the same result holds simultaneously across tuples at family-wise level α_V.

Proof. Conditional on n_h, the sampled outcomes in direction h are a simple random sample from its N_h frame units. Hypergeometric inversion therefore covers each K_h exactly. A union bound covers all nonempty directions. On this event, the gain is increasing in K_h for $h \in \mathcal{H}_+$ and decreasing for $h \in \mathcal{H}_-$, so Equation (43) is a valid lower bound. The qualification claim follows immediately. For multiple tuples, a further union bound over their allocated budgets gives the family-wise statement. An unsampled direction uses $[0, N_h]$; a direction with $N_h = 0$ contributes zero. □

Unresolved human labels are handled by taking the union of the confidence sets over all admissible label completions. This union contains the set computed from the true completion, so its coverage is at least that of the complete-label procedure without any additional multiplicity penalty. Measurement must qualify throughout this enlarged set; decision uses its smallest policy-gain lower bound. This construction also applies to the finite-frame measurement certificate.

As in prediction, replacing $[A_k^V, B_k^V]$ by ranges computed from the realized qualification opportunities would control only their conditional mean. The population decision claim uses the deterministic envelope above.

Corollary 2 (Finite-sample PGC authorization)**.** *For a fixed family of K tuples, use either the population procedures in Theorem 3 and Propositions 7 and 9, or the common finite-frame design in Propositions 5, 8, and 10. Allocate error budgets α_P, α_M, and α_V. If $\alpha_P + \alpha_M + \alpha_V \leq \alpha$, requiring all three certificates for authorization controls family-wise false authorization at level α. The three statistics may use overlapping observations from the same qualification cohort.*

Proof. The three cited results control their respective family-wise false-passage events. False authorization implies at least one of these events, so their union has probability at most $\alpha_P + \alpha_M + \alpha_V$. This allocation therefore also protects the component certificates simultaneously. The union bound permits dependence from shared observations. Each axis must target the same declared population or finite frame with a design-valid procedure. □

Figure 2 specializes Theorem 7 to no ordinary covariates, balanced binary strata $\mathbb{P}(E = 0) = \mathbb{P}(E = 1) = 1/2$, conditional risks $q_e = \mathbb{P}(D = 1 \mid E = e)$, and normalized symmetric costs $c_{\mathrm{FA}} = c_{\mathrm{FR}} = 1/2$. Writing $\bar{q} = (q_0 + q_1)/2$, $a_\varnothing = \mathbf{1}\{\bar{q} < .5\}$, and $a_e = \mathbf{1}\{q_e < .5\}$ with reject ties gives

$$I_B = \frac{(q_1 - q_0)^2}{4}, \qquad V_{.5} = \frac{1}{2}\sum_{e=0}^{1} |q_e - .5|\mathbf{1}\{a_\varnothing \neq a_e\}. \tag{44}$$

Thus $(q_0, q_1) = (.20, .80)$ yields $(I_B, V) = (.0900, .150)$, whereas $(.01, .49)$ yields $(.0576, 0)$ because neither stratum changes the baseline action. The cost normalization fixes the numerical scale of V.

Corollary 3 (Root-Brier decision-regret bound)**.** *For an (X, E)-measurable estimate $\widehat{\eta}$ and action $\widehat{a} = \mathbf{1}\{\widehat{\eta} < \tau\}$,*

$$\mathcal{R}(\widehat{a}) - \mathcal{R}(a_{X,E}) \leq (c_{\mathrm{FA}} + c_{\mathrm{FR}})\sqrt{\mathbb{E}[(\widehat{\eta} - \eta_{X,E})^2]}. \tag{45}$$

Proof. On action disagreement, $\widehat{\eta}$ and $\eta_{X,E}$ lie on opposite sides of τ, so $|\eta_{X,E} - \tau| \leq |\widehat{\eta} - \eta_{X,E}|$. Apply the preceding wrong-action loss identity in Equation (40), then Cauchy–Schwarz to obtain Equation (45). □

D.6 Joint qualification criterion

Prediction, measurement, and decision statuses are assigned separately using the rules in Appendix A.4. Stress metrics such as AUROC, top-risk event rate, calibration, and replicate direction can diagnose failure but do not modify the prespecified family-wise proper-score criterion. Development analyses inform whether to initiate formal qualification; formal status is reserved for an independent certification sample.

D.7 Compositional authorization guarantee

Proof of Theorem 2. Fix any data-generating law. Let $\mathcal{I}$ index its invalid tuples, and choose a false component $j(k)$ for each $k \in \mathcal{I}$. This choice depends on the law, not on observed test outcomes. Authorization of k requires $\phi_{j(k)k} = 1$. Consequently,

$$\mathbb{P}(\exists k \in \mathcal{I} : \ \textsc{Authorized}_k) \leq \sum_{k \in \mathcal{I}} \mathbb{P}(\phi_{j(k)k} = 1) \leq \sum_k \max_j \alpha_{jk} \leq \alpha, \tag{46}$$

which proves Equation (6). The argument uses no independence assumption. This is an intersection–union test within each tuple (Berger & Hsu, 1996) and a union bound across tuples. It controls erroneous joint authorization, not simultaneous correctness of every separately reported axis.

Corollary 2 gives the additive allocation that also protects all component certificates simultaneously. □

Proposition 11 (Non-substitutability of qualification evidence)**.** *With prediction and decision margins .001, minimum sensitivity and specificity .75, and maximum between-stratum error gap .25, for each of the prediction, measurement, and decision axes there is a balanced binary diagnostic world in which the other two claims hold while that claim fails. Thus no rule guaranteeing the joint claim $P_k \cap M_k \cap V_k$ can omit an axis uniformly over these worlds.*

Proof of Proposition 11. Let E be balanced and use $c_{\mathrm{FA}} = c_{\mathrm{FR}} = .5$, hence threshold $\tau = .5$.

For the prediction witness, let the observed and semantic risks be $(.49, .51)$ and use perfect measurement. Set the baseline forecast to .5 and the fixed diagnostic forecasts to $(0, 1)$. The diagnostic forecast has expected Brier loss .49, compared with .25 for the baseline, so the prediction claim fails. With reject ties, however, the baseline rejects both strata whereas the diagnostic policy accepts the lower-risk stratum and rejects the higher-risk stratum. Equation (40) gives decision value .005, so the measurement and decision claims hold.

For the measurement witness, retain semantic risks $(.20, .80)$ but let the observed risks be $(.10, .90)$. These arise from a differential channel with $(s_0, t_0) = (.5, 1)$ and $(s_1, t_1) = (1, .5)$. A forecast equal to the observed risks has Brier information .160 beyond their marginal mean. The intended semantic risks still have decision value .150, but sensitivity or specificity falls below .75 in each stratum and the between-stratum error gap exceeds .25. The measurement claim therefore fails.

For the decision witness, let the observed and semantic risks be $(.01, .49)$, with forecasts equal to those risks and baseline forecast .25. Prediction improves Brier risk by .0576 and measurement is valid. Both conditional risks remain below .5, so the baseline and refined Bayes actions always accept and decision value is zero. Each construction leaves the two retained claims true while the omitted claim fails. □

Proposition 12 (Policy value does not identify diagnostic authority)**.** *Two worlds can induce the same joint law of the human endpoint and both policy actions, hence the same policy value, while the prediction and measurement claims have opposite truth values under the same thresholds as Proposition 11.*

Proof of Proposition 12. Let E be balanced, set $D^H = E$, and use equal false-accept and false-reject costs. Take $p_0 = .5$ and $p_1 = .25$ when $E = 0$ and .75 when $E = 1$, with reject ties. The baseline rejects both strata, whereas the augmented policy accepts $E = 0$ and rejects $E = 1$. Its human-endpoint decision value is therefore .25.

In the first world, set $\widetilde{D} = D^H$. The augmented Brier risk is .0625, compared with .25 for the baseline, and measurement is perfect. In the second world, set $\widetilde{D} = 1 - D^H$ while leaving E, D^H, both predictions, and both actions unchanged. The augmented Brier risk becomes .5625, and the automatic endpoint reverses the human endpoint. The direct human-endpoint policy comparison has the same law and value in both worlds, but prediction and measurement change from true to false. Hence policy value alone does not identify diagnostic authority. □

E Controlled Validation

This section gives the controlled data-generating processes, grids, ablations, and stress tests underlying Section 5.1.

E.1 Controlled data-generating processes

The controlled prediction study has a balanced binary core feature X and a prespecified family $E_1, \ldots, E_K$. Only E_1 is coupled to the semantic outcome:

$$\mathbb{P}(D^\star = 1 \mid X, E_1) = \pi + \lambda(2E_1 - 1), \tag{47}$$

with λ represented as a fraction of the largest value preserving probabilities in $[0, 1]$. Symmetric nondifferential measurement with Youden index κ yields

$$\mathbb{P}(\widetilde{D} = 1 \mid X, E_1) = \frac{1 - \kappa}{2} + \kappa\mathbb{P}(D^\star = 1 \mid X, E_1). \tag{48}$$

Equation (48) is Equation (37) specialized to symmetric error with $s = t = (1+\kappa)/2$. Development prompts fit cell-mean predictors. Predictors and family are then fixed before an independent certification sample. The design uses 300 development prompts, certification $n \in \{40, 80, 120\}$, family sizes $1, 5, 10$, three α levels, 12,000 null trials, and 4,000 positive trials per cell.

Table 6: Six oracle-assisted benchmark worlds. Ordered pairs are risks for the two balanced diagnostic strata; the last column gives the evaluator risks in this benchmark.

World	Development truth	Certification truth	Evaluator risk
Semantic null	$(.30, .30)$	$(.30, .30)$	$(.30, .30)$
Stable positive	$(.20, .80)$	$(.20, .80)$	$(.20, .80)$
Transport reversal	$(.20, .80)$	$(.80, .20)$	$(.80, .20)$
Differential measurement	$(.30, .30)$	$(.30, .30)$	$(.20, .80)$
Attainability infeasible	$(.45, .55)$	$(.45, .55)$	$(.45, .55)$
Decision irrelevant	$(.01, .49)$	$(.01, .49)$	$(.01, .49)$

Table 7: Comparison rules for the oracle-assisted benchmark. Reliability and development estimates are reference rules. Measurement, attainability, and decision checks augment the independent prediction rule.

Pathway	Evidence used for its authorization rule
Reliability only	Repeatability screen; no **outcome-grounding** requirement
Development estimate	Favorable **development** paired-score estimate
Predictive certificate	**Independent** fixed-cohort predictive certificate
+ Measurement	Predictive certificate and exact **measurement-grounding** flag
Analytic PGC pathway	Predictive certificate, measurement qualification, **attainable margin**, and positive **decision value**

The authorization benchmark uses a balanced binary diagnostic with the exact risks in Table 6 and comparison rules in Table 7. Each trial fits on 300 prompts, evaluates a development point rule on 120 prompts, and reserves 120 independent certification prompts.

The attainability-infeasible world shrinks the candidate contrast by 0.02 and requires practical margin 0.005. All other worlds use unit shrinkage and zero margin. The decision threshold is 0.5. Measurement and decision checks use exact DGP information. This isolates the qualification logic from sampling variation in a human audit or deployment trial. The development estimate and independent certificate are alternative rules. The power curve evaluates the fixed test directly at each registered sample size.

E.2 Finite-sample authorization with sampled gates

The finite-sample benchmark applies the population prediction rule in Theorem 3 and the sample procedures in Propositions 7 and 9. Prediction and decision use deterministic ranges that cover every diagnostic stratum and binary endpoint. Each trial fits forecasts on 300 development prompts and evaluates them on an independent qualification sample of 240, 480, or 960 units. Prediction, measurement, and decision receive error budgets 0.020, 0.015, and 0.015. The measurement region uses minimum sensitivity and specificity 0.75 and maximum cross-stratum range 0.25. Units are sampled independently from balanced diagnostic strata, satisfying the within-stratum binomial model used by the measurement certificate.

The stable and non-measurement failure worlds use sensitivity and specificity 0.95 in both diagnostic strata, replacing the oracle-assisted benchmark's observation channel. Thus semantic risk η gives evaluator risk $q = .05 + .90\eta$: the stable pair becomes $(.23, .77)$ and the null pair becomes $(.32, .32)$. In the differential world, stratum-specific channels induce evaluator risks 0.20 and 0.80 from equal semantic risks 0.30. Authorization uses sampled paired labels, forecasts, actions, and costs. DGP parameters label ground-truth validity afterward.

At $n = 960$, the complete rule authorizes 3,974/4,000 stable trials and none in the five failure worlds. Stable authorization is 127/4,000 at $n = 240$ and 2,375/4,000 at $n = 480$, identifying measurement qualification as the main sample cost. Every failure world has zero PGC authorization at all three sizes; the one-sided 95% binomial upper bound is 7.49×10^{-4} per cell.

Table 9: Null-grid summary. Each row spans $n \in \{40, 80, 120\}$, $K \in \{1, 5, 10\}$, three prevalences and three measurement qualities. Each cell has 12,000 trials.

Bound	α	Cells	Max. FWER	Min. coverage	Cellwise 95% UCB
Empirical Bernstein	0.01	81	0.0000	1.0000	2.50×10^{-4}
Empirical Bernstein	0.05	81	**0.0000**	1.0000	2.50×10^{-4}
Empirical Bernstein	0.10	81	0.0000	1.0000	2.50×10^{-4}
Hoeffding	0.01	81	0.0000	1.0000	2.50×10^{-4}
Hoeffding	0.05	81	**0.0000**	1.0000	2.50×10^{-4}
Hoeffding	0.10	81	0.0000	1.0000	2.50×10^{-4}

The UCB is pointwise for each zero-count cell. This grid uses the global $[-1, 1]$ loss range and evaluates the population certificate in Theorem 3.

Table 8: Finite-sample operating characteristics of the complete authorization rule. All three outcome-bearing gates are estimated from each independent qualification sample; DGP quantities are used only to determine whether authorization is warranted.

World	n	Authorized / trials	Rate (%)	False / trials	False UCB (%)
Stable coupling	240	127/4,000	3.175	–	–
	480	2,375/4,000	59.375	–	–
	960	3,974/4,000	**99.350**	–	–
Semantic null	240	0/4,000	0.000	0/4,000	0.0749
	480	0/4,000	0.000	0/4,000	0.0749
	960	0/4,000	0.000	0/4,000	0.0749
Transport reversal	240	0/4,000	0.000	0/4,000	0.0749
	480	0/4,000	0.000	0/4,000	0.0749
	960	0/4,000	0.000	0/4,000	0.0749
Differential measurement	240	0/4,000	0.000	0/4,000	0.0749
	480	0/4,000	0.000	0/4,000	0.0749
	960	0/4,000	0.000	0/4,000	0.0749
Unattainable margin	240	0/4,000	0.000	0/4,000	0.0749
	480	0/4,000	0.000	0/4,000	0.0749
	960	0/4,000	0.000	0/4,000	0.0749
Decision-irrelevant prediction	240	0/4,000	0.000	0/4,000	0.0749
	480	0/4,000	0.000	0/4,000	0.0749
	960	0/4,000	0.000	0/4,000	0.0749

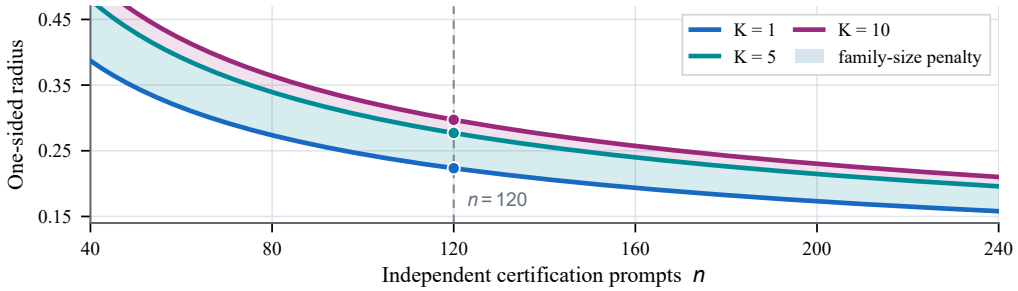
False UCB is the one-sided 95% exact binomial upper bound. Prediction and decision use ranges uniform over the DGP support; measurement uses independent equal-probability draws within each stratum. The stable world contains warranted authorizations; all other worlds are failure mechanisms. Alpha is allocated as 0.020/0.015/0.015 across prediction, measurement and decision.

E.3 Null calibration

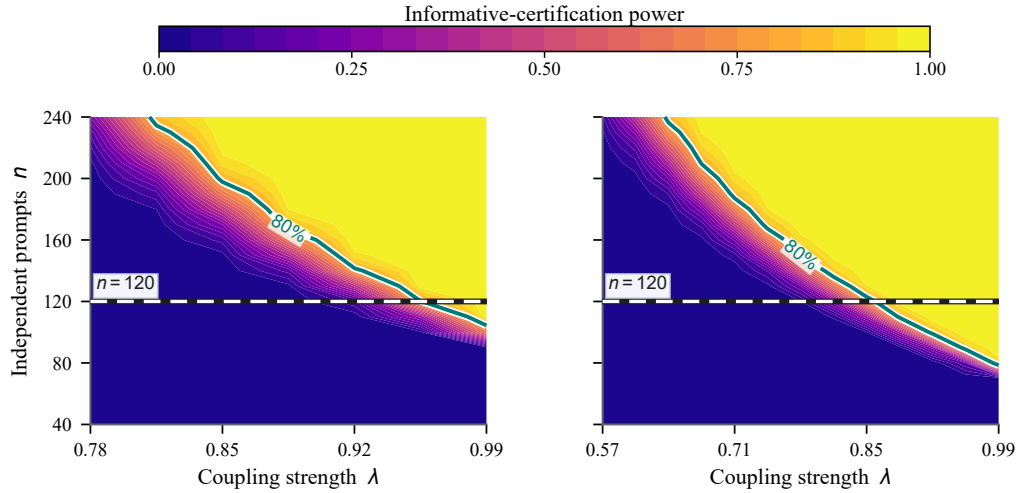
Across 486 null grid cells, zero false certifications are observed in 12,000 trials per cell. For a zero count, the one-sided 95% exact binomial upper bound is 2.50×10^{-4} per cell. The largest Monte Carlo standard error among the nonzero reported grid rates is 0.00457. Empirical simultaneous coverage never falls below nominal in the prespecified grid. These are finite Monte Carlo observations, separate from the analytical guarantee in Theorem 3.

E.4 Power surfaces

At $n = 120$ and $\lambda = .85$, the empirical-Bernstein certificate has 73.3% power. At $\lambda = .92$ it has 100% power, while the range-agnostic Hoeffding comparison has 11.4% at $\lambda = .92$. The dense sensitivity contains 315 cells per bound, 630 cells total, with 4,000 trials per DGP setting.



(a) Family-wise radius contracts with sample size



(b) Global-range Hoeffding

(c) Empirical Bernstein

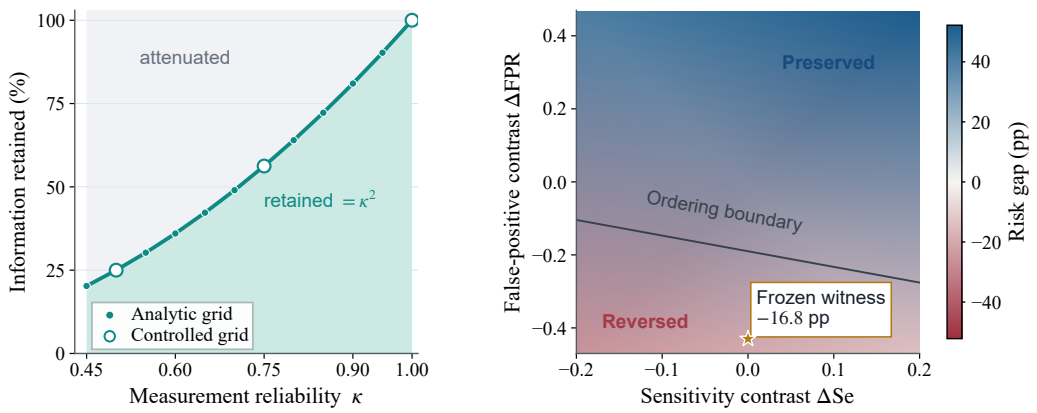
Figure 4: Certificate behavior across sample size, family size and coupling. (a) One-sided global-range Hoeffding radius for $K \in \{1, 5, 10\}$ at $\alpha = .05$; shading is the family-size penalty and the reference marks $n = 120$. (b,c) Controlled power for global-range Hoeffding and empirical-Bernstein certificates, using 4,000 trials per cell. Color gives power, contours mark 80%, and horizontal references mark $n = 120$. Null calibration is in Table 9.

E.5 Measurement grid

For every controlled coupling and κ in the prespecified grid, the observed Brier information divided by semantic information equals κ^2. The maximum absolute numerical discrepancy is 1.11×10^{-16}. The exact differential witness in Appendix D.2 reverses the ranking. The information ratios are computed analytically from the DGP.

E.6 Analytic-gate pathway comparison

The oracle-assisted benchmark compares reliability and development estimates with an independent fixed-cohort prediction certificate. It then adds measurement, decision, and attainability checks to



(a) Information attenuation (b) Ordering phase map

Figure 5: Measurement error determines whether proxy information preserves ordering. (a) Nondifferential error follows exact κ^2 attenuation; 12 filled analytic points and three open κ locations cover 27 controlled DGP cells. (b) Differential error changes ordering for true risks 0.20 and 0.40. Color is the observed-risk gap, the black contour is zero ordering, and the star is the prespecified witness ($q_{\text{high}} - q_{\text{low}} = -0.168$). Appendices D.1 and D.2 derive the attenuation and reversal, respectively.

that certificate. Measurement and decision use DGP quantities to isolate their contributions. Table 10 reports all 30 cells. The sample-based experiment in Table 8 is the operating-characteristic result used in the main text.

Table 10: Oracle-gate mechanism benchmark retained for comparison. Prediction is sampled, while measurement and decision use DGP quantities. Intervals are pointwise exact binomial Monte Carlo intervals, not PGC certificates.

World	Pathway	Count / trials	Rate (%)	95% CI (%)
Stable coupling	Reliability only	4,000/4,000	100.000	[99.908, 100.000]
	Development estimate	4,000/4,000	100.000	[99.908, 100.000]
	Predictive certificate	3,412/4,000	85.300	[84.164, 86.384]
	+ Measurement	3,412/4,000	85.300	[84.164, 86.384]
	PGC authorization	3,412/4,000	**85.300**	[84.164, 86.384]
Semantic null	Reliability only	4,000/4,000	100.000	[99.908, 100.000]
	Development estimate	1,609/4,000	40.225	[38.701, 41.764]
	Predictive certificate	7/4,000	0.175	[0.070, 0.360]
	+ Measurement	7/4,000	0.175	[0.070, 0.360]
	PGC authorization	0/4,000	0.000	[0.000, 0.092]
Transport reversal	Reliability only	4,000/4,000	100.000	[99.908, 100.000]
	Development estimate	4,000/4,000	100.000	[99.908, 100.000]
	Predictive certificate	0/4,000	0.000	[0.000, 0.092]
	+ **Measurement**	0/4,000	0.000	[0.000, 0.092]
	PGC authorization	0/4,000	0.000	[0.000, 0.092]
Differential measurement	Reliability only	4,000/4,000	100.000	[99.908, 100.000]
	Development estimate	4,000/4,000	100.000	[99.908, 100.000]
	Predictive certificate	3,398/4,000	84.950	[83.804, 86.045]
	+ Measurement	0/4,000	0.000	[0.000, 0.092]
	PGC authorization	0/4,000	0.000	[0.000, 0.092]

Table 10 (continued)

World	Pathway	Count / trials	Rate (%)	95% CI (%)
Unattainable margin	Reliability only	4,000/4,000	100.000	[99.908, 100.000]
	Development estimate	0/4,000	0.000	[0.000, 0.092]
	Predictive certificate	0/4,000	0.000	[0.000, 0.092]
	+ Measurement	0/4,000	0.000	[0.000, 0.092]
	PGC authorization	0/4,000	0.000	[0.000, 0.092]
Decision irrelevant	Reliability only	4,000/4,000	100.000	[99.908, 100.000]
	Development estimate	4,000/4,000	100.000	[99.908, 100.000]
	Predictive certificate	2,430/4,000	60.750	[59.217, 62.267]
	+ Measurement	2,430/4,000	60.750	[59.217, 62.267]
	PGC authorization	0/4,000	0.000	[0.000, 0.092]

For every zero count, the one-sided 95% upper probability bound is 7.49×10^{-4} (0.0749%). The two-sided intervals above use a different tail allocation.

E.7 Leave-one-requirement ablation

We remove one requirement at a time from the complete sampled-gate rule. Each comparison reuses the primary experiment's development fits and qualification draws, with $n = 960$, 4,000 trials per world, and error allocation $(.020, .015, .015)$. All three outcome-bearing procedures use sampled data.

Table 11: Paired component deletion under the primary sampled-gate protocol. Authorization rates (%) use the same $n = 960$, 4,000 trials, development fits, qualification samples, and error allocation as Table 8. Columns remove one requirement from Full: prediction (P), measurement (M), decision (D; claim V_k), or attainability (A).

World	Full	$-P$	$-M$	$-D$	$-A$
Stable coupling	99.350	99.350	100.000	99.350	99.350
Semantic null	0.000	0.000	0.000	0.175	0.000
Transport reversal	0.000	0.000	0.000	0.000	0.000
Differential error	0.000	0.000	**52.825**	0.000	0.000
Unattainable margin	0.000	0.000	0.000	0.000	0.000
Same-action risks	0.000	0.000	0.000	0.000	0.000

Measurement blocks the differential-error world, while Decision removes the remaining semantic-null passages (Table 11). The other deletions leave final authorization unchanged in this grid because the retained checks also reject those worlds.

Fixed-predictor axis isolation. Table 2 separates the three claims using a balanced binary diagnostic and predictors fixed before qualification sampling. In the valid world, semantic risks are $(.2, .8)$, baseline prediction is .5, augmented predictions are $(.23, .77)$, and the action threshold is .5. The prediction-failure world uses semantic risks $(.35, .65)$, baseline .49, augmented predictions $(.001, .501)$, and threshold .5. Its population Brier difference is $+.056901$, although its semantic policy gain is .075. The measurement-failure world uses semantic risks $(.3, .3)$, baseline .5, augmented predictions $(.2, .8)$, and threshold .5. Its evaluator sensitivities are $(.5, 1)$ and specificities $(13/14, 2/7)$, yielding evaluator risks $(.2, .8)$ and semantic policy gain .1. The decision-failure world uses semantic risks $(.2, .6)$, baseline .41, augmented predictions $(.23, .59)$, and threshold .75. Its Brier difference is $-.0324$, but both policies accept every case, so policy gain is zero. All other worlds use sensitivity and specificity .95 in both strata. The valid world's Brier difference and policy gain are $-.0729$ and .15, respectively.

Each world uses 4,000 trials at $n = 3,840$, with the same error allocation and qualification thresholds as the preceding experiment. Baseline and augmented predictors are specified directly rather

Table 12: Sample-size sensitivity of fixed-predictor axis isolation. Entries are authorization rates (%) from 4,000 trials per world and sample size. Every deletion is evaluated on the same samples as Full. All four prespecified smaller sample sizes are shown.

n	World	Full PGC	Without P	Without M	Without D
240	Valid coupling	3.2	4.325	93.6	3.225
240	Miscalibrated prediction	0	7.625	0	0
240	Invalid measurement	0	0	63.925	0
240	Unchanged actions	0	0	0	1.7
480	Valid coupling	59.075	59.075	100	59.075
480	Miscalibrated prediction	0	70.175	0	0
480	Invalid measurement	0	0	98.275	0
480	Unchanged actions	0	0	0	54.975
960	Valid coupling	99.225	99.225	100	99.225
960	Miscalibrated prediction	0	99.5	0	0
960	Invalid measurement	0	0	100	0
960	Unchanged actions	0	0	0	98.725
1920	Valid coupling	100	100	100	100
1920	Miscalibrated prediction	0	100	0	0
1920	Invalid measurement	0	0	100	0
1920	Unchanged actions	0	0	0	100

than estimated from development data. Full and deleted rules share qualification samples; only the indicated requirement is removed. The sample-size extension fixes $n \in \{240, 480, 960, 1920\}$ before simulation and retains all four worlds (Table 12). Smaller samples reduce valid-world authorization, while the complete rule authorizes none of the three failure worlds at any tested size.

E.8 Continuous qualification boundaries

A mechanism study varies four conditions across 86 grid points, using 2,000 trials per point. Its prediction test targets the conditional mean on the sampled qualification units and uses their binary-outcome ranges. Measurement uses the known κ threshold, and decision uses exact Bayes value; their conjunction illustrates mechanism boundaries rather than sampled population qualification. The latter is evaluated in Appendix E.2.

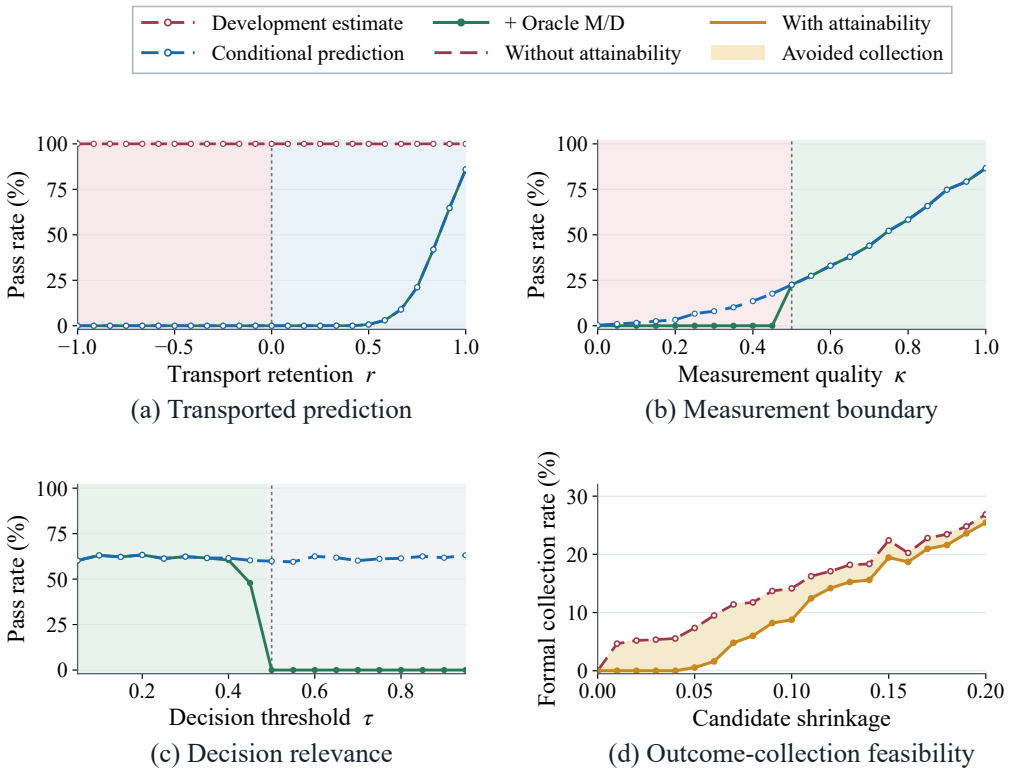
Under complete transport reversal, the development rule always passes while the conditional prediction test never passes. Stable transport gives an 85.95% joint pass rate. At $\kappa = .45$, prediction passes in 17.65% of trials, but the $\kappa \geq .5$ requirement blocks the joint rule. At $\tau = .5$, same-side risks give a 59.85% prediction pass rate and zero Bayes value. At shrinkage .02, attainability prevents collection in every trial; retaining action feasibility alone permits collection in 5.20%.

E.9 Adaptive-search and subgroup stress

In each of 4,000 trials, 20 null diagnostics are evaluated on the same 120 outcomes. Selecting the most favorable point estimate produces an apparent improvement in every trial. Freezing the selected predictor and evaluating it on an independent sample with the global-range population certificate produces false-certification rate zero. In the subgroup world, the stable 80% majority yields a favorable aggregate ΔBrier of $-.018$, while the 20% reversal stratum has ΔBrier $+.270$ (Figure 7a).

E.10 Rare events and decision stress

Differential labels generate 0.09 apparent Brier information from a semantic null. At $n = 120$, rare-event certificate power remains below 1% through 10% prevalence even when AUROC is approximately 0.70. Decision curves evaluate two exact diagnostic-risk worlds over 19 thresholds; the



(a) Transported prediction (b) Measurement boundary

(c) Decision relevance (d) Outcome-collection feasibility

Figure 6: Oracle-assisted mechanism boundaries. Blue tests prediction conditional on sampled units; green also requires known measurement quality and positive Bayes value. (a) Transport retention. (b) Measurement threshold $\kappa \geq .5$. (c) Decision value for risks $(.01, .49)$. (d) Gold marks collections avoided by attainability. There are 2,000 trials per point across 25, 21, 19, and 21 grid points, respectively.

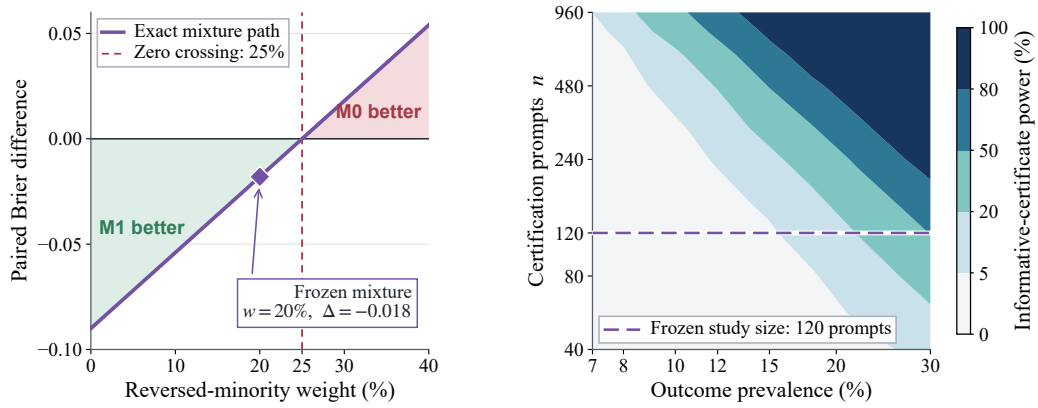
$(.01, .49)$ world is shown in Figure 6c. Value is positive precisely where the additional risk separates actions.

F Natural Three-Axis Qualification

This study evaluates a naturally arising target-attribute diagnostic. Development and qualification use disjoint prompts; qualification draws a probability sample from a new finite frame.

F.1 Target, predictors, and qualification design

For each search transition, the diagnostic is the decrease in clipped target-attribute log odds from the current winner to the proposed candidate. The CLIP–FlanT5-XL proxy (Lin et al., 2024) scores the target-attribute question with the prompt "{question} Please answer yes or no." Its current-image probability enters $\mathcal{M}_0$; the candidate-relative log-odds decrease enters only $\mathcal{M}_1$. The qualification endpoint is instead produced by the distinct Qwen2.5-VL-7B evaluator (Bai et al., 2025b) with its fixed binary-answer adapter and .5 threshold. Neither its score nor its binary endpoint is a predictor feature. The baseline predictor $\mathcal{M}_0$ uses current global proxy, candidate-minus-current proxy, current attribute probability, and prompt-source family. The augmented predictor $\mathcal{M}_1$ adds the diagnostic and a fixed monotone risk basis. Both predictors, their preprocessing and coefficients, the automatic and human endpoints, the margins, the sampling plan, and the reject tie rule were fixed before qualification outcomes were collected.



(a) Mixture sensitivity (b) Qualification frontier

Figure 7: Subgroup composition and event prevalence determine qualification. (a) Mixing effects of -0.090 and $+0.270$ crosses zero at 25% reversed-minority weight; the prespecified 20% mixture gives $\Delta\text{Brier} = -0.018$. (b) The displayed 99 controlled cells span 7–30% prevalence and 40–960 prompts on log-spaced axes (4,000 trials per cell). Colored regions encode the 0–5%, 5–20%, 20–50%, 50–80%, and 80–100% power ranges; the dashed purple reference marks the prespecified study size of 120 prompts. A supplementary grid contains 33 cells at 4–6% prevalence, all at most 1.2% power; AUROC is 0.70–0.77.

The qualification frame contains 1,227 cases from two prompt-source families: 456 GenEval 2-derived cases (Kamath et al., 2025) and 771 cases derived from the Dense Prompt Graph Benchmark (DPG-Bench) (Hu et al., 2024). DPG-Bench metadata came from the official ELLA repository at revision `3c228f1dc6c4d3cad0a47493816151a419f14db3`. Eligibility required an answerable target-attribute question; normalized duplicates and development prompts were excluded. One attribute, generation seed, and current-to-candidate transition were specified per remaining prompt. All cases use Stable Diffusion 3.5 Medium (Stability AI, 2024) revision `b940f670f0eda2d07fbb75229e779da1ad11eb80`, a current pool A of 16 images (A_{16}), and a new candidate pool B of eight images (B_8). The resulting 1,227 cases have distinct prompt identifiers and form the finite population for qualification. Every recorded transition has $C = 1$, so this analysis concerns semantic susceptibility among replacements. Before sampling, the study fixed a Brier-improvement margin of .001, a policy-improvement margin of .005, minimum sensitivity and specificity of .80 and .95, and a maximum between-source error-rate difference of .10. False-accept and false-reject costs are .95 and .05, giving the fixed acceptance threshold $\tau = .05$. The prediction and decision margins define the minimum improvements treated as meaningful. The measurement thresholds prioritize downside sensitivity while requiring stricter false-alarm control, and the 19:1 cost ratio defines the downside-averse study policy. Across the frame, the attainable paired-Brier interval is $[-.15980, .18018]$, so the prediction margin is attainable. The fixed policies differ on 426 frame cases, establishing that the decision contrast is nondegenerate. A design-stage power analysis then fixed the qualification sample size at 1,000.

F.2 Probability sample and semantic measurement

The qualification sample contains 1,000 cases drawn without replacement from the frame, with common inclusion probability $1000/1227 = .814996$. The certificate targets the mean over the full 1,227-case frame. Prediction uses the automatic evaluator endpoint. Decision uses the operational human-semantic endpoint D^H: the current image satisfies the requested attribute and its selected replacement does not, as determined by the prespecified two-of-three case-vote majority. Measurement links these two endpoints.

Each case received separately elicited judgments from three of 10 blinded annotators. Left–right presentation was randomized, and annotators were not shown model identities, proxy scores, eval-

uator outputs, or predicted risks. Each image received a yes, no, or unsure attribute judgment. The 3,000 blinded paired case ratings comprise 6,000 image-level judgments; 109 were unsure and 10 were missing. Five cases lacked a two-of-three majority and remain unresolved; among the other 995 cases, 153 are human-positive and 842 human-negative. The evaluator agrees with 983 resolved majorities (TP=153, FN=0, TN=830, FP=12). Table 13 propagates unresolved cases through simultaneous finite-frame intervals.

Image-level labels have pooled pairwise agreement .875, Fleiss $\kappa = .722$ (Fleiss, 1971), and nominal Krippendorff $\alpha = .722$ (Krippendorff, 2011). Case-level downside votes have pairwise agreement .875, Fleiss $\kappa = .587$, and Krippendorff $\alpha = .587$. Of the 1,000 cases, 814 receive three identical definite votes, 181 receive a two-of-three definite majority, and five remain unresolved. Across individual annotators, downside-vote rates range from 13.8% to 19.6% and indeterminate-vote rates from 0.7% to 3.0%.

Table 13: Finite-frame measurement qualification. TP, FN, TN, and FP entries are simultaneous intervals for frame totals, obtained by inverting hypergeometric tests conditional on the sampled stratum sizes.

Stratum	n_g/N_g	TP / FN	Sensitivity	TN / FP	Specificity
GenEval 2	367/456	$[62, 88]$ / $[0, 6]$	$[\mathbf{0.912}, 1.000]$	$[360, 386]$ / $[5, 16]$	$[\mathbf{0.957}, 0.987]$
DPG-Bench	633/771	$[103, 132]$ / $[0, 7]$	$[0.936, 1.000]$	$[628, 660]$ / $[7, 15]$	$[0.977, 0.990]$

Eight two-sided cell intervals share $\alpha_M = .015$; unresolved majorities enlarge every compatible cell. The confidence rectangle remains inside the registered region: sensitivity at least .80, specificity at least .95, and each between-stratum error-rate difference at most .10.

This design samples 81.5% of the frame. The controlled sample-size analysis in Appendix E.2 examines the role of measurement precision at smaller sample sizes.

F.3 Prediction, decision, and authorization

Prediction uses Proposition 5 with the uniform Brier-loss range $[-1, 1]$. The paired estimate is -0.049469. Its one-sided .02 Serfling UCB is -0.011340, below the threshold $-.001$. The prediction axis therefore passes.

At $\tau = .05$, the baseline and augmented policies disagree on 350 sampled cases. On the full frame, 307 cases change from accept to reject and 119 from reject to accept. The five unresolved majorities yield a sample policy-gain identification interval $[.0192, .0212]$. Exact conditional hypergeometric intervals for the two action directions give a one-sided finite-frame lower bound of .01108, above the .005 decision margin. The simultaneous measurement rectangle in Table 13 also lies inside the declared sensitivity, specificity, and differential-error region.

For this single tuple ($K = 1$), the three error allocations sum to .05. Corollary 2 therefore returns AUTHORIZED for the fixed diagnostic, predictors, automatic and human endpoints, finite frame, and cost ratio.

F.4 Robustness

Absolute policy losses and a retention reference. The augmented policy accepts 207 candidate replacements and reduces accepted downside cases from 36–39 to 9–10 relative to $\mathcal{M}_0$. Table 14 reports the accompanying rejection costs and includes always retaining the current image as a reference. This secondary comparison evaluates the fixed policies on the same 1,000 cases at .95/.05 costs. We enumerate all $2^5 = 32$ completions of unresolved human labels and compute each paired gain within the same completion. The retention-to-$\mathcal{M}_1$ sample gain is $[.00035, .00135]$, compared with $[.0192, .0212]$ for the registered $\mathcal{M}_0$-to-$\mathcal{M}_1$ comparison.

Endpoint, source, and label sensitivity. The remaining analyses test whether the result depends on the automatic endpoint, one prompt source, a label convention, or a small number of cases. Replacing the automatic endpoint by D^H without refitting either predictor gives a sample paired-

Table 14: Absolute policy outcomes on the qualification sample at the fixed 19:1 cost ratio. FA and FR count accepted downside cases and rejected safe replacements. Ranges give each column's minimum and maximum across unresolved-label completions. Retention is a secondary reference policy.

Policy	Accept (%)	FA	FR	Mean loss
$\mathcal{M}_0$	36.3	36–39	518–520	.06010–.06305
$\mathcal{M}_1$	20.7	9–10	645–649	.04080–.04195
Always retain current	0.0	0	842–847	.04210–.04235

Brier interval of $[-.05378, -.05282]$ after propagating the five unresolved cases. The corresponding one-sided finite-frame upper bound is $-.01469$.

The two prompt sources support the same direction. GenEval 2 and DPG-Bench have automatic-endpoint ΔBrier values of $-.03798$ and $-.05613$; their descriptive bootstrap intervals remain below zero. Their sample policy-gain intervals are $[.01567, .02112]$ and $[.02125, .02125]$. Recoding every unresolved vote in either direction yields human-endpoint ΔBrier between $-.05347$ and $-.05313$ and policy gain between .01920 and .02120. On the 814 cases with unanimous definite votes, these quantities remain $-.05766$ and .01026. Removing the 50 largest absolute paired prediction contributions leaves ΔBrier at $-.03794$. Table 15 reports these checks with the agreement statistics.

Table 15: Human agreement and robustness of the natural qualification. Source-specific Brier intervals are descriptive case bootstraps; policy-gain intervals reflect unresolved labels. Agreement uses Fleiss' κ (Fleiss, 1971) and nominal Krippendorff's α (Krippendorff, 2011). Label and influence analyses use the fixed qualification records and leave the formal certificate unchanged.

A. Agreement			
Unit	Pair agreement	Fleiss κ	Krippendorff α
Image-level attribute	.875	.722	.722
Case-level downside	.875	.587	.587
B. Prompt-source consistency			
Source	Cases	ΔBrier [95% CI]	Policy-gain interval
GenEval 2	367	$-.\mathbf{03798}$ $[-.06257, -.01288]$	$[.01567, .02112]$
DPG-Bench	633	$-.\mathbf{05613}$ $[-.07392, -.03824]$	$[.02125, .02125]$
C. Human-label convention			
Rule	Cases (positive)	Human ΔBrier	Policy gain
Unresolved $\rightarrow$ non-downside	1,000 (153)	$-.05347$	.01920
Unresolved $\rightarrow$ downside	1,000 (158)	$-.05313$	.02120
Unanimous definite only	814 (93)	$-.05766$	$\mathbf{.01026}$
D. Prediction influence			
Removed by $\lvert d_i \rvert$	Cases retained	ΔBrier	Removed $\lvert d_i \rvert$ share
None	1,000	$-.04947$	0.0%
Largest 10	990	$-.04279$	7.1%
Largest 50	950	$-.\mathbf{03794}$	30.7%

Cost sensitivity. A cost-threshold sweep recomputes the fixed policy comparison for $\tau = .01, .02, \ldots, .50$, with additional checks at .025 and .075. Costs remain normalized as $c_{\mathrm{FA}} = 1 - \tau$ and $c_{\mathrm{FR}} = \tau$. On the 50-point regular grid, the decision lower bound exceeds the .005 margin at 47 of 50 thresholds, including $\tau = .50$; it does not at .09–.11. The formal authorization uses the

prespecified $\tau = .05$. Changing the threshold changes which cases switch actions and in which direction, producing a nonmonotone lower bound. Table 16 reports representative thresholds.

Table 16: Policy-threshold sensitivity. Selected points summarize the 50-threshold cost sweep; † marks the formal operating point and the remaining rows characterize robustness.

τ	Cost ratio	Frame changes	Decision LCB	Clears .005
$.05^{\dagger}$	19.0:1	426	.0111	Yes
.09	10.1:1	465	.0030	No
.10	9.0:1	452	$-.0002$	No
.11	8.1:1	431	.0045	No
.20	4.0:1	370	.0134	Yes
.50	1.0:1	201	.0338	Yes

G Motivating Counterexample and Positive Controls

The preceding study qualifies a target-attribute diagnostic. Here, the OU study examines a contrasting case: repeatability without local-future outcome grounding. Two prescribed-coupling controls then test whether known relations can be recovered on generated images.

G.1 Motivating repeatability counterexample

The source analysis in Section 5.2 evaluates OU energy against local-future outcomes on 57 prompts.

Experimental design. The source study uses 57 GenEval 2 attribute-bearing development prompts and the settings below. The analysis computes OU energy from pair IDs 0/1 and evaluates it against independent local-future pools.

Table 17: Implementation settings for the motivating repeatability counterexample.

Component	Setting
Generator	SDXL base 1.0 (Podell et al., 2024); 30 steps; CFG 7.0; 1024^2; float16
Proxy	CLIP–FlanT5-XL (Lin et al., 2024)
Evaluator	Qwen3-VL-8B (Bai et al., 2025a); Soft-TIFA attribute at threshold .5 (Kamath et al., 2025)
Prompts	GenEval 2 (Kamath et al., 2025); 57 prompts
Search design	**Current pool** of 16 images (A_{16}); two diagnostic pairs at each of two scales; four local-future pools of eight calls
Predictor fitting	Five **grouped outer folds** with grouped inner selection of C

For diagnostic scales $\rho \in \{.995, .98\}$ and two antithetic pairs,

$$z^{\pm}_{\rho,j} = \rho z^{\star}_A + \sqrt{1-\rho^2}(\pm\epsilon^{\mathrm{diag}}_j), \tag{49}$$

with diagnostic noise disjoint from every future-pool noise. Letting $r_0 = r(z^{\star}_A)$,

$$G_\rho = \frac{1}{2J(1-\rho^2)} \sum_{j=1}^{J} \sum_{q\in\{+,-\}} \big(r(z^q_{\rho,j}) - r_0\big)^2, \qquad E = \log(10^{-8} + G_{.995} + G_{.98}). \tag{50}$$

Equations (49)–(50) fix the implemented diagnostic conditional on prompt, sampler, and step randomness. The empirical question is whether E adds information about independent local-future semantic outcomes.

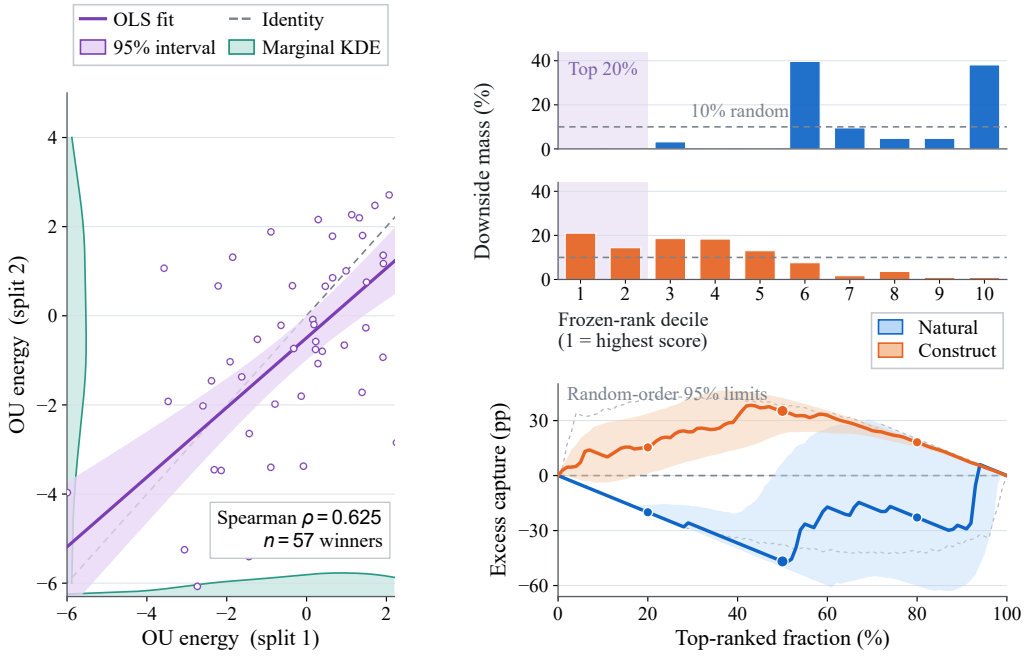
Each prompt uses four local future pools at $\rho = .98$, with four antithetic pairs per pool and the current winner included in selection. Diagnostic and future epsilon namespaces are disjoint.

Table 18: Prompt-equal source predictions on 57 prompts, 98 units and 392 future outcomes. Brier and log loss are proper losses for the binary future outcome, averaged over the observed future pools.

Predictor	Brier ↓	Log loss ↓	AUROC	AUPRC	ECE (5 bins)
Intercept	0.04467643	0.195490	0.2676	0.0348	0.0416
Ordinary $\mathcal{M}_0$	**0.04467602**	0.195479	0.2602	0.0347	0.0421
OU $\mathcal{M}_1$	**0.04467904**	0.195498	0.2699	0.0351	0.0462

For unit event fraction y, Brier loss is $y(1-p)^2 + (1-y)p^2$, not $(p-y)^2$. Lower Brier favors the model. AUROC/AUPRC and five-bin calibration are descriptive; a fold-specific intercept need not have AUROC 0.5 after pooling folds.

Predictors and evaluation. Each of the 57 prompts contributes one current-winner state and one unit for each registered attribute question. These 98 state–attribute units each average four future-pool outcomes, giving $98 \times 4 = 392$ nested outcomes. Five outer folds split prompts. Continuous variables are standardized on the training fold. Categorical atom type is encoded on the training fold. Ridge penalty $C \in \{.01, .1, 1, 10\}$ is selected by grouped inner validation. The baseline contains point reward, winner margin, latent typicality, atom type, neighbor exceedance, CLIP locality (Radford et al., 2021), and LPIPS locality (Zhang et al., 2018). The augmented model adds only log total normalized OU energy. Prompt-equal Brier loss is primary. Log loss, AUROC, AUPRC, calibration, top-quintile rate, and replicate direction are stress metrics.



(a) Diagnostic repeatability

(b) Outcome grounding

Figure 8: Repeatability does not establish outcome grounding. (a) Split-half OU energy is repeatable across 57 winners (Spearman $\rho = .625$). (b) Its ranking concentrates downside mass only under prescribed outcome coupling. Shading in (a) is an approximate 95% interval for the fitted regression mean; random-order limits in (b) are pointwise permutation quantiles. Arms use their respective outcomes; Appendix G.3 defines the construct-controlled arm.

Results. There are 14 evaluator-defined downsides across nine prompts. The primary scores are reported below. Split-half OU energy has Spearman $\rho = .625$ with 95% prompt-bootstrap interval $[.388, .798]$. In Figure 8, the top two deciles contain no natural downside, compared with 35.4% of the positive control's known risk. The paired $\mathcal{M}_1 - \mathcal{M}_0$ Brier difference is

$+3.02 \times 10^{-6}$ with prompt-cluster 95% interval $[-2.91 \times 10^{-5}, 3.77 \times 10^{-5}]$. The $\mathcal{M}_1$ AUPRC interval is $[.0134, .0794]$. Its highest-risk 10% and 20% of prompt-equal weight contain no downside events. The latter's 2,000-draw bootstrap interval is $[0, .0073]$. Five-bin ECE is $.0462$ with interval $[.0289, .0968]$. These secondary stress metrics use cumulative prompt-equal mass. In contrast, Figure 8 partitions 98 units into ten fixed-rank deciles and averages prompts within each bin. Its 10,000-draw bootstrap rebuilds the bins, giving the top-decile interval $[0, .0192]$. The point estimate is zero under both definitions.

Replicate halves give paired Brier differences of $+7.61 \times 10^{-6}$ and -1.58×10^{-6}. Together with the primary interval, these results show repeatability without predictive qualification in the source study.

G.2 Matched human audit

All 14 evaluator-positive transitions and 28 attribute-matched retained controls were selected before annotation. Each annotator sees the prompt, pinned GenEval 2 attribute question, and two neutrally named images. Proxy scores, evaluator scores, case type, and current/replacement identity are hidden. Ten annotators complete all 42 cases, yielding 420 paired ratings. Fleiss' κ is 0.6763 and mean pairwise image agreement is 87.9%. Strict majorities identify one downside among 14 evaluator-positive transitions and 22 retained cases among 28 controls. Because this matched sample is not a probability sample of the full outcome population, it is descriptive rather than a measurement certificate.

G.3 Diffusion positive controls

These studies test whether the certificate recovers a known coupling on real candidate images and under an independent generator.

Real-candidate construct bridge. The real-candidate bridge uses fixed candidate images and the exact exogenous-nuisance mechanism in Appendix B.6. The primary setting assigns a fixed current nuisance and independent standard normal future nuisance, uses loading $\lambda = 10$, and computes exact conditional downside probabilities for the single-pair theorem. The nominal unit is prompt. Because this bridge evaluates a fixed 57-prompt cohort, its certificate uses the conditional prompt-averaged bound in Proposition 4.

The primary oracle-oriented arm has ΔBrier -0.182943 and fixed-cohort UCB -0.029407. All 100 registered nuisance-seed sensitivities certify. The natural-replicate-1 sensitivity has UCB $+0.02423$ and certification rate zero. The prescribed coupling passes the fixed-cohort prediction criterion, whereas assigning natural roles to the same images does not.

Independent SD3.5 positive control. The independent arm uses 64 score-blind prompts and Stable Diffusion 3.5 Medium (Stability AI, 2024) at 1024px, with Qwen2.5-VL-7B (Bai et al., 2025b) as the held-out evaluator. It defines H as VQAScore (Lin et al., 2024) on a shared generated image, adds independent nuisance with loading 10, and integrates conditional risk with 20,000 nuisance draws per future pool. Diagnostic–risk Spearman correlation is $.510$, and the paired Brier difference is $-.00500$ with a prompt-bootstrap 95% interval $[-.00869, -.00193]$. The independent arm therefore recovers the prescribed coupling under a second generator and evaluator.

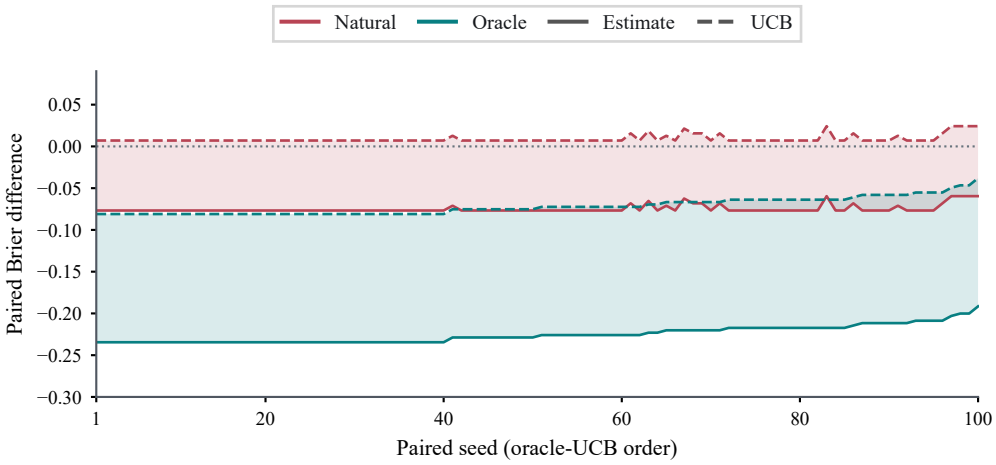


Figure 9: Construct certification across 100 paired registered nuisance repetitions. Seeds share the oracle-UCB ordering. Each ribbon spans paired ΔBrier to its fixed-cohort UCB, so its thickness is the registered range penalty. Natural and oracle-oriented roles certify in 0/100 and 100/100 repetitions, respectively.

G.4 Exact geometry of the motivating predictor pair

The fixed OU predictor pair illustrates how support and action geometry can rule out authority without an outcome-based qualification test.

Attainability. The prelaunch screen applies the source-trained OU predictors to 114 state–attribute units from a separate 64-prompt SD3.5 cohort. Units within a prompt share equal prompt weight. For every unit, the two possible Brier differences are those in Equation (27). Applying Proposition 2, equivalently Equation (8), gives

$$\mathcal{S}_w = [-8.1326575 \times 10^{-6},\ 8.8271505 \times 10^{-6}], \tag{51}$$

so the maximum attainable improvement is 8.1326575×10^{-6}. The prespecified design sets the practical margin at 0.5% and the planning alternative at 1% of the prior baseline Brier 0.0304922945:

$$m_{\text{pred}} = 1.5246147 \times 10^{-4}, \qquad \delta_{\text{plan}} = 3.0492295 \times 10^{-4}. \tag{52}$$

The fixed predictor contrast covers only

$$\frac{8.1326575 \times 10^{-6}}{1.5246147 \times 10^{-4}} = 0.0533424 \tag{53}$$

of the declared practical target. Both the certification target and planning alternative in Equation (52) lie outside the exact interval in Equation (51). The same support rules out every target above 0.026671% of baseline Brier.

Action geometry. The false-accept and false-reject costs are 19 and 1, representing a policy that weights accepting a downside nineteen times as heavily as rejecting a safe replacement. They give threshold

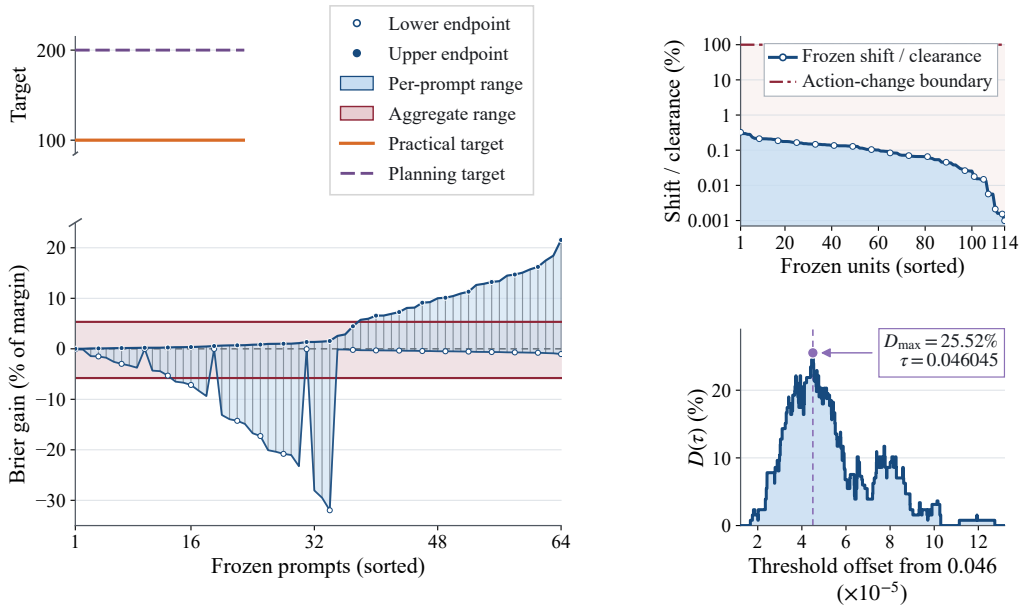
$$\tau = \frac{1}{19+1} = 0.05. \tag{54}$$

Across the fixed units, p_0 ranges from 0.0460167 to 0.0461194 and p_1 from 0.0460176 to 0.0461275. The maximum absolute probability contrast $\max_i |p_{1i} - p_{0i}|$ is 2.5549243×10^{-5}. Both predictors choose the same action everywhere:

$$\mathbb{P}(a_{\mathcal{M}_0} \neq a_{\mathcal{M}_1}) = 0. \tag{55}$$

The two prediction-induced policies are identical, so their pairwise policy-value difference is exactly zero under any realized outcome law by Equation (55). This calculation does not identify the Bayes value of observing E, which depends on the unknown true conditional risks in Appendix D.4.

Because the attainable interval and induced actions are properties of the predictor pair, increasing the cohort size cannot change either conclusion.



(a) Attainable gain (b) Action feasibility

Figure 10: Outcome-blind screen for the motivating predictor pair. (a) Its attainable Brier-gain interval misses the practical and planning targets. (b) Half the prediction shift is at most 0.323% of the midpoint-to-threshold clearance, producing no action change at $\tau = .05$. The lower panel scans action disagreement over thresholds near .046; the registered threshold .05 lies outside this range.

H Extended Related Work

Table 19 extends Section 2 by comparing the objects and guarantees of related methods.

PGC uses proper losses and valid risk bounds within its qualification tests. Its contribution is to connect these tests to the diagnostic's role in search: the fixed predictor pair must first be able to attain the gain margin and change an action, after which independent outcomes test prediction, measurement and decision value. A direct comparison on probability-sampled human outcomes can establish the value of a fixed policy pair (Proposition 10); PGC additionally qualifies the predictive signal and its automatic endpoint. With the same paired audit, the decision contrast measures policy gain, while the measurement intervals identify missed downsides, false alarms, and between-stratum errors in the automatic instrument. Prediction tests the incremental signal in that instrument. The shared cohort thus evaluates the signal, evaluator, and policy together.

I Authorization Target and Human Annotation

I.1 Authorization scope

PGC indexes every authorization by a diagnostic, prediction pair, automatic and human endpoints, population, and policy. The natural study instantiates this object on the 1,227-case target-attribute frame. Cross-system positive controls and the matched audit address complementary implementa-

Table 19: Related methods: evaluation targets and principal guarantees.

Approach	Evaluation target	Principal guarantee or objective
Proper scoring rules (Gneiting & Raftery, 2007)	Probabilistic forecasts	Truthful reporting optimizes the **expected score**
SelectiveNet (Geifman & El-Yaniv, 2019)	Prediction with abstention	Joint optimization of the **risk–coverage** trade-off
Learn-then-Test (Angelopoulos et al., 2025)	Predictors or policies under specified losses	**Joint risk control** for multiple requirements using calibration outcomes
Conformal risk control (Angelopoulos et al., 2024)	Predictor family with monotone loss	**Expected-risk control** under its calibration rule
Adaptive holdout methods (Dwork et al., 2015; Blum & Hardt, 2015)	Repeated queries or submissions	Protected evaluation under **adaptive feedback**
PGC	Diagnostic-guided search decisions	Joint tests of predictive gain, measurement validity and decision value, with **false-authorization control**

tion and measurement questions. A randomized policy comparison targets the distinct causal effect of deployment.

I.2 Ethics and annotator privacy

Ten volunteers participated without compensation or supervisory relationships with the authors. All consented to the scientific and publication use of their judgments. They were blinded to the hypothesis, diagnostic, predicted direction, and qualification results. The task asked only whether generated images satisfied a stated attribute. We retained pseudonymous judgments and collected no demographic or directly identifying information. The manuscript reports aggregate results; the numerical supplement retains pseudonymous case-level ratings for recomputing the analyses. The annotation activity followed the applicable institutional policy for minimal-risk, non-identifying research, under which formal review was not required.